Statistické modelování

Jan Koláček

Obsah

Úvod

Tento text je určen zejména pro studenty předmětu „M5VM05 Statistické modelování“. Jde o nadstavbovou část základního kurzu pravděpodobnosti a matematické statistiky „Spojité modely a statistika B“, který je výchozím pro další teoretické i aplikačně zaměřené stochastické předměty.

Kurz nejprve obsahuje základy popisné statistiky a průzkumovou analýzu dat. Zabývá se vlastnostmi bodových odhadů, zejména nestranností a konzistencí. Zmínka je též o intervalových odhadech parametrů, především normálního rozdělení a také o odhadech založených na centrální limitní větě. V návaznosti na tuto problematiku kurz pokračuje testováním hypotéz. V další části jsou probrány obecné základy regresní a korelační analýzy. Na ty pak kurz navazuje klasickým lineárním regresním modelem, kde je kladen důraz zejména na odhad neznámých parametrů a testování hypotéz o těchto parametrech. Jako speciální příklad lineárních regresních modelů je dále podrobněji studována analýza rozptylu. V další části jsou úvahy rozšířeny na zobecněné lineární modely (tzv. „GLM modely“). Zejména se jedná o popis základních komponent modelů a aplikací GLM modelů v konkrétních případech. Závěr kurzu se ještě věnuje modelování závislosti mezi kvalitativními proměnnými.

Podstatná část textu byla převzata ze zdrojů [**3, 6, 7**] a také z výukových materiálů dr. Forbelské k předmětům „Lineární statistické modely I, II“ a „Zobecněné lineární modely“. Většina tvrzení je uvedena bez důkazů, ty lze nalézt ve výše uvedených zdrojích. Zkoumaná problematika je demonstrována na příkladech se snahou o lepší srozumitelnost textu. Pro více příkladů odkazujeme studenty na cvičení k tomuto kurzu.

Závěrem bych rád poděkoval RNDr. Marii Budíkové, Dr. a RNDr. Marii Forbelské, Ph.D. za poskytnuté studijní materiály a za cenné rady a připomínky k tomuto textu.

Jan Koláček

Literatura

[1] J. Anděl. *Matematická statistika*. SNTL, Praha, 1985.
[2] J. Anděl. *Statistické metody*. Matfyzpress, Praha, 1993.
[3] M. Budíková, T. Lerch, and Š. Mikoláš. *Základní statistické metody*. Masarykova univerzita, Brno, 2005.
[4] V. Dupač and M. Hušková. *Pravděpodobnost a matematická statistika*. Karolinum, Praha, 1999.
[5] M. Forbelská. *M722 Zobecněné lineární modely – pracovní text*. 2013.
[6] M. Forbelská and J. Koláček. *Pravděpodobnost a statistika I [online]*. Elportál. Masarykova univerzita, 1. vyd. edition, 2013.
[7] M. Forbelská and J. Koláček. *Pravděpodobnost a statistika II [online]*. Elportál. Masarykova univerzita, 1. vyd. edition, 2013.
[8] P. Hebák and J. Hustopecký. *Vícerozměrné statistické metody s aplikacemi*. SNTL, Praha, 1987.
[9] J. Michálek. *Úvod do teorie pravděpodobnosti a matematické statistiky*. Státní pedagogické nakladatelství, Praha, 1984.
[10] R. C. Rao. *Lineární metody statistické indukce a jejich aplikace*. Academia Praha, 1978.
[11] K. Zvára and J. Štěpán. *Pravděpodobnost a matematická statistika*. Matfyzpress, Praha, 2001.

KAPITOLA 1

Průzkumová analýza jednorozměrných dat

Základní informace

(1) V následující kapitole se budeme zabývat průzkumovou analýzou jednorozměrných dat. Popíšeme funkcionální a číselné charakteristiky datového souboru a uvedeme základní diagnostické grafy.
(2) Předpokládá se znalost pouze nejzákladnějších pojmů z teorie pravděpodobnosti, např. náhodná veličina.

Výstupy z výukové jednotky

Studenti

- umí spočítat a graficky znázornit bodové rozložení četností
- umí spočítat a graficky znázornit intervalové rozložení četností
- umí spočítat a interpretovat zadaný kvantil
- vypočítají a interpretují průměr a rozptyl
- umí popsat a vysvětlit krabicový diagram
- umí vykreslit a interpretovat N–P, Q–Q a P–P plot

1. Motivace

Průzkumová analýza dat je odvětví statistiky, které pomocí různých postupů odhaluje zvláštnosti v datech. Při zpracování dat se často používají metody, které jsou založeny na předpokladu, že data pocházejí z nějakého konkrétního rozložení, nejčastěji normálního. Tento předpoklad nemusí být vždy splněn, protože data

- mohou pocházet z jiného rozložení
- mohou být zatížena hrubými chybami
- mohou pocházet ze směsi několika rozložení.

Proto je důležité provést průzkumovou analýzu dat, abychom se vyvarovali neadekvátního použití statistických metod.

Poznámka 1.1. Většina textu v této kapitole byla převzata z [**3**]. Pro podrobnější studium tohoto tématu proto odkazujeme na tento zdroj.

2. Funkcionální charakteristiky datového souboru

2.1. Označení. Na množině objektů $\{\varepsilon_1, \ldots, \varepsilon_n\}$ zjišťujeme hodnoty znaku X. Hodnotu znaku X na objektu ε_i označíme $x_i, i = 1, \ldots, n$. V teorii pravděpodobnosti se jim také říká **realizace** náhodné veličiny X. Tyto hodnoty zaznamenáme do jednorozměrného datového souboru:

$$\mathbf{x} = (x_1, \ldots, x_n)'.$$

Uspořádané hodnoty $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ tvoří uspořádaný datový soubor:

$$\mathbf{x}_{(\cdot)} = (x_{(1)}, \ldots, x_{(n)})'.$$

Vektor

$$\mathbf{x}_{[\cdot]} = (x_{[1]}, \ldots, x_{[r]})',$$

kde $x_{[1]} < \ldots < x_{[r]},\ r \leq n$, jsou navzájem různé hodnoty znaku X, se nazývá vektor variant.

2.2. Bodové rozložení četností. Je-li počet variant malý, přiřazujeme četnosti jednotlivým variantám a hovoříme o bodovém rozložení četností.
Zaveďme tzv. **indikátor množiny** předpisem:

$$I_B(x) = \begin{cases} 1 & x \in B, \\ 0 & x \notin B. \end{cases}$$

DEFINICE 2.1. Pro datový soubor $\mathbf{x} = (x_1, \ldots, x_n)'$ definujeme následující pojmy

- **absolutní četnost varianty** $x_{[j]}$:
$$n_j = \sum_{i=1}^{n} I_{\{x_{[j]}\}}(x_i)$$
- **relativní četnost varianty** $x_{[j]}$:
$$p_j = \frac{n_j}{n}$$
- **absolutní kumulativní četnost prvních** j **variant**:
$$N_j = n_1 + \ldots + n_j$$
- **relativní kumulativní četnost prvních** j **variant**:
$$F_j = \frac{N_j}{n} = p_1 + \ldots + p_j$$
- **četnostní funkce**:
$$p(x) = \begin{cases} p_j & \text{pro } x = x_{[j]},\ j = 1, \ldots, r \\ 0 & \text{jinak} \end{cases}$$
- **empirická distribuční funkce**:
$$F(x) = \tfrac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x>}(x_i)$$

Poznámka 2.2. Absolutní či relativní četnosti znázorňujeme graficky např. pomocí sloupkového diagramu či polygonu četností.

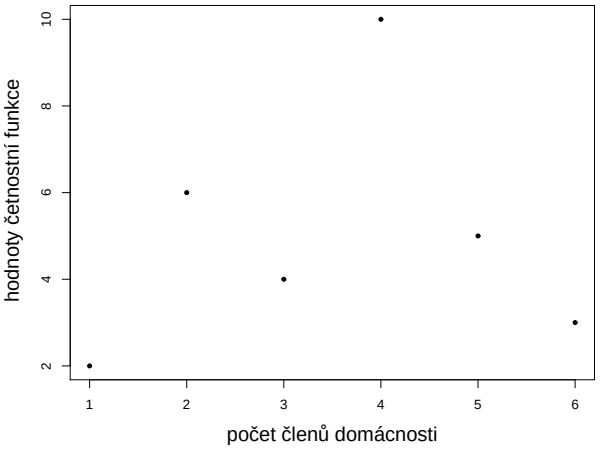
Příklad 2.3. U 30 domácností byl zjišťován počet členů.

Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

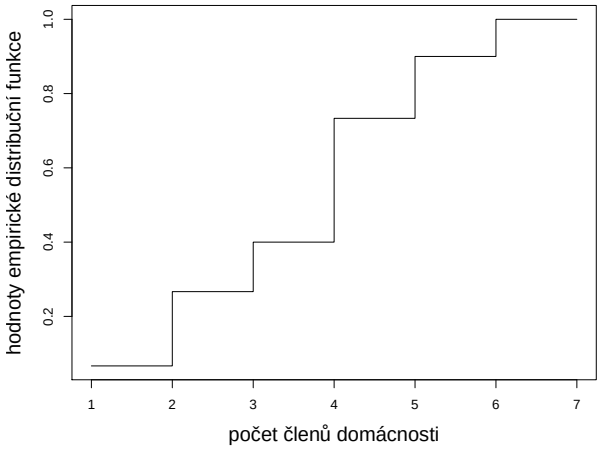
Vytvořte tabulku rozložení četností. Nakreslete grafy četnostní funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností počtu členů domácnosti.

Řešení. Tabulka rozložení četností:

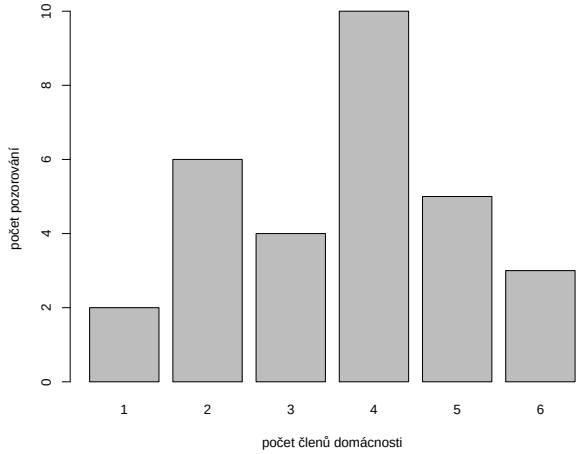
$x_{[j]}$	n_j	p_j	N_j	F_j
1	2	2/30	2	2/30
2	6	6/30	8	8/30
3	4	4/30	12	12/30
4	10	10/30	22	22/30
5	5	5/30	27	27/30
6	3	3/30	30	1



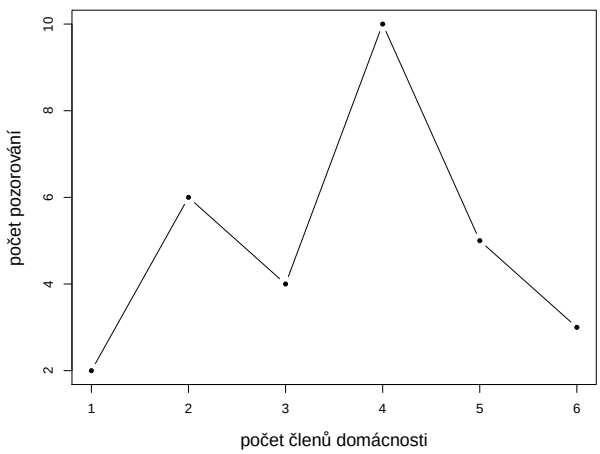
Obr. 1: Graf četnostní funkce



Obr. 2: Graf empirické distribuční funkce



Obr. 3: Sloupkový diagram



Obr. 4: Polygon četností

Příslušné grafy jsou na Obr. 1 až Obr.4.

2.3. Intervalové rozložení četností. Je-li počet variant velký, přiřazujeme četnosti nikoli jednotlivým variantám, ale třídicím intervalům $(u_1, u_2\rangle, \ldots, (u_r, u_{r+1}\rangle$ a hovoříme o intervalovém rozložení četností. Názvy četností jsou podobné jako v bodě 2.2. Stanovení počtu třídicích intervalů je dosti subjektivní záležitost. Často se doporučuje volit r blízké $\sqrt{n}$.

DEFINICE 2.4. Četnostní hustota j-tého třídicího intervalu je definována vztahem

$$f_j = \frac{p_j}{d_j}$$

kde p_j jsou relativní četnosti v jednotlivých intervalech a $d_j = u_{j+1} - u_j$. Soustava obdélníků sestrojených nad třídicími intervaly, jejichž plochy jsou rovny relativním četnostem (a výšky tedy četnostním hustotám), se nazývá **histogram**.

DEFINICE 2.5.

- **hustota četnosti**:

$$f(x) = \begin{cases} f_j & \text{pro } u_j < x \leq u_{j+1},\ j = 1, \ldots, r \\ 0 & \text{jinak} \end{cases}$$

  (grafem hustoty četnosti je schodovitá čára shora omezující histogram)
- **Intervalová empirická distribuční funkce**:

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Příklad 2.6. U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	$(35, 65\rangle$	$(65, 95\rangle$	$(95, 125\rangle$	$(125, 155\rangle$	$(155, 185\rangle$	$(185, 215\rangle$
Počet domácností	7	16	27	14	4	2

Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

Řešení. Tabulka rozložení četností

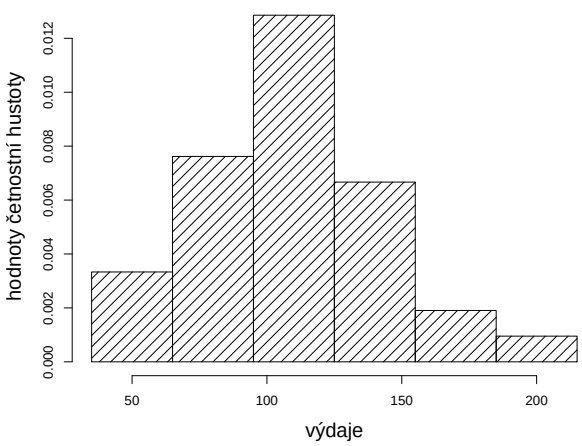
$(u_j, u_{j+1}\rangle$	n_j	p_j	f_j	N_j	F_j
$(35, 65\rangle$	7	7/70	7/2100	7	7/70
$(65, 95\rangle$	16	16/70	16/2100	23	23/70
$(95, 125\rangle$	27	27/70	27/2100	50	50/70
$(125, 155\rangle$	14	14/70	14/2100	64	64/70
$(155, 185\rangle$	4	4/70	4/2100	68	68/70
$(185, 215\rangle$	2	2/70	2/2100	70	1

3. Číselné charakteristiky datového souboru

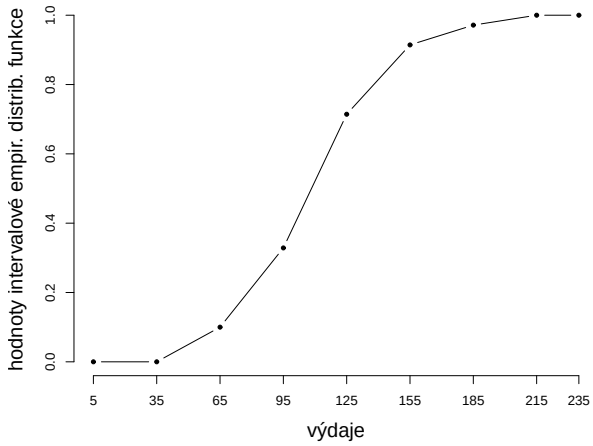
3.1. Znaky nominálního typu. Nominální škála klasifikuje objekty do určitých předem vymezených tříd či kategorií. Hodnoty v nominální škále se dají vyjádřit slovně a mezi různými hodnotami není definováno žádné uspořádání. Pokud jsou hodnoty nominální škály někdy označovány číselně, mějme na paměti, že toto číslo je pouze jakousi zkratkou (kódem) slovní hodnoty. O znacích měřených v nominální škále hovoříme jako o znacích nominálního typu.

Příklad 3.1. Příklady znaků nominálního typu mohou být např.:

- pohlaví (s možnými hodnotami mužské, ženské)
- barva očí (modrá, hnědá, černá)
- výsledek léčby (uzdraven, zemřel)



Obr. 5: Histogram



Obr. 6: Graf intervalové empirické distribuční funkce

- národnost (česká, slovenská, polská, německá, ...)

DEFINICE 3.2. Charakteristikou polohy je **modus** – nejčetnější varianta či střed nejčetnějšího intervalu. (Modus je jediná charakteristika polohy vhodná pro nominální veličiny).

3.2. Znaky ordinálního typu. Znaky ordinálního typu lze podle sledované vlastnosti nejen rozlišovat, ale také uspořádat ve smyslu vztahů „je větší“, „je menší“ nebo „předchází“, „následuje“, aniž bychom však byli schopni vyjádřit číselně vzdálenost mezi větším a menším či mezi předcházejícím a následujícím.

Příklad 3.3. Znaky ordinálního typu mohou být např.:

- dosažené vzdělání (základní, střední, vysokoškolské)
- prospěch ve školním předmětu (výborně, velmi dobře, dobře, nevyhověl)
- důstojnická hodnost (podporučík, poručík, nadporučík, kapitán, ...)
- stav pacienta (vyléčen, remise, recidiva)
- hodnocení funkce technických zařízení (stupně závažnosti poruchy jaderné elektrárny)
- ohrožení povodní (stupně povodňové aktivity)
- hodnocení postojů v sociologických průzkumech (škála má hodnoty např. souhlasím, spíše souhlasím, spíše nesouhlasím, nesouhlasím)
- četnost výskytu (často, občas, zřídka, nikdy)

Vhodnou charakteristikou polohy je α-kvantil.

DEFINICE 3.4. Je-li $\alpha \in (0;1)$, pak α-kvantil x_a je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl $1-\alpha$ všech dat.

NÁVOD 3.5. Pro výpočet α-kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c & \Rightarrow x_\alpha = \frac{x_{(c)}+x_{(c+1)}}{2} \\ \text{necelé číslo} & \Rightarrow \text{ zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

Označení. Pro speciálně zvolená α užíváme názvů:

- $x_{0,50}$ – medián
- $x_{0,25}$ – dolní kvartil
- $x_{0,75}$ – horní kvartil
- $x_{0,1}, \ldots, x_{0,9}$ – decily
- $x_{0,01}, \ldots, x_{0,99}$ – percentily.

Jako charakteristika variability slouží kvartilová odchylka: $q = x_{0,75} - x_{0,25}$.

Příklad 3.6. Během semestru se studenti podrobili písemnému testu z matematiky, v němž bylo možno získat 0 až 10 bodů. Výsledky jsou uvedeny v tabulce:

Počet bodů	0	1	2	3	4	5	6	7	8	9	10
Počet studentů	1	4	6	7	11	15	19	17	12	6	3

Zjistěte modus, medián, 1. decil, 9. decil a kvartilovou odchylku počtu bodů.

Řešení. Modus je nejčetnější varianta znaku, v tomto případě tedy 6. Pro výpočet kvantilů musíme znát rozsah datového souboru: $n = 1 + 4 + \cdots + 3 = 101$.
Výpočty uspořádáme do tabulky.

α	$n\alpha$	c	$x_\alpha = x_{(c)}$
0,50	50,5	51	6
0,10	10,1	11	2
0,90	90,9	91	8
0,25	25,25	26	4
0,75	75,75	76	7

Kvartilová odchylka: $q = 7 - 4 = 3$.

3.3. Znaky intervalového a poměrového typu. U znaků intervalového typu lze stanovit vzdálenost mezi hodnotami měřené veličiny. Je zde definována jednotka měření, avšak nula je definována pouze relativně. To nám dovoluje proto počítat s rozdíly naměřených hodnot, nikoliv s jejich podíly. Typickým příkladem je teplota, která se dá měřit v různých stupnicích (Celsiova, Fahrenheitova).

U znaků poměrového typu lze určit nejen rozdíly (intervaly) mezi hodnotami, ale i podíly hodnot, neboť tyto znaky mají nulu stanovenu absolutně a jednoznačně.

Definice 3.7. **Aritmetický průměr** $\bar{x}$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

U poměrových znaků, které nabývají pouze kladných hodnot, lze použít **geometrický průměr**:

$$\sqrt[n]{x_1 \cdot \ldots \cdot x_n} \tag{2}$$

Oba dva průměry jsou charakteristikou polohy.

DEFINICE 3.8.

(1) **rozptyl**:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

(2) **směrodatná odchylka**:

$$s = \sqrt{s^2} \tag{4}$$

(3) **koeficient variace** (pro poměrové znaky):

$$\frac{s}{\bar{x}} \tag{5}$$

Všechny uvedené charakteristiky jsou charakteristikami variability datového souboru.

Poznámka 3.9. Rozptyl se zpravidla počítá podle vzorce $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.

DEFINICE 3.10. Známe-li absolutní či relativní četnosti variant $x_{[1]}, \ldots, x_{[r]}$, můžeme spočítat **vážený průměr**:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]} \tag{6}$$

nebo **vážený rozptyl**:

$$s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j (x_{[j]} - \bar{x})^2 \tag{7}$$

Poznámka 3.11. Vážený rozptyl se zpravidla počítá podle vzorce $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^2 - \bar{x}^2$.

Aritmetický průměr a rozptyl jsou speciální případy tzv. momentů. V následující definici obecně zavedeme k-tý počáteční a centrální moment.

DEFINICE 3.12.

- **k-tý počáteční moment**:

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad \text{kde } k = 1, 2, \ldots \tag{8}$$

- **k-tý centrální moment**:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^k, \quad \text{kde } k = 1, 2, \ldots \tag{9}$$

Pomocí 3. a 4. centrálního momentu se definuje šikmost a špičatost:

DEFINICE 3.13.

- **šikmost**:

$$\alpha_3 = \frac{m_3}{s^3} \tag{10}$$

Šikmost měří nesouměrnost rozložení četností kolem průměru.

- **špičatost**:

$$\alpha_4 = \frac{m_4}{s^4} - 3 \tag{11}$$

Špičatost měří koncentraci rozložení četností kolem průměru.

Příklad 3.14. Pro údaje z Příkladu 2.3 vypočtěte průměr a rozptyl počtu členů domácnosti.
Řešení.

$$\bar{x} = \tfrac{1}{30}(1 \cdot 2 + 2 \cdot 6 + 3 \cdot 4 + 4 \cdot 10 + 5 \cdot 5 + 6 \cdot 3) = \tfrac{109}{30} = 3,6\bar{3}$$

$$s^2 = \tfrac{1}{30}\left(1^2 \cdot 2 + 2^2 \cdot 6 + 3^2 \cdot 4 + 4^2 \cdot 10 + 5^2 \cdot 5 + 6^2 \cdot 3\right) - \left(\tfrac{109}{30}\right)^2 = \tfrac{1769}{900} = 1,96\bar{5}$$

Příklad 3.15. Nechť $\bar{x}$ je průměr a s_1^2 rozptyl hodnot $x_1, \ldots, x_n$. Nechť a, b jsou reálné konstanty. Položme $y_i = a + bx_i, i = 1, \ldots, n$. Vypočtěte průměr $\bar{y}$ a rozptyl s_2^2 hodnot $y_1, \ldots, y_n$.
Řešení.

$$\bar{y} = \tfrac{1}{n}\sum_{i=1}^{n} y_i = \tfrac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\tfrac{1}{n}\sum_{i=1}^{n} x_i = a + b\bar{x},$$

$$s_2^2 = \tfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \tfrac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\bar{x})^2 = b^2\tfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 s_1^2.$$

4. Diagnostické grafy

4.1. Krabicový diagram (Box plot). Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot. Můžete se setkat i z názvem **box plot**.

DEFINICE 4.1. Krabicový diagram je specifikován těmito pojmy:
Dolní vnitřní hradba:

$$x_{0,25} - 1,5q$$

Horní vnitřní hradba:

$$x_{0,75} + 1,5q$$

Dolní vnější hradba:

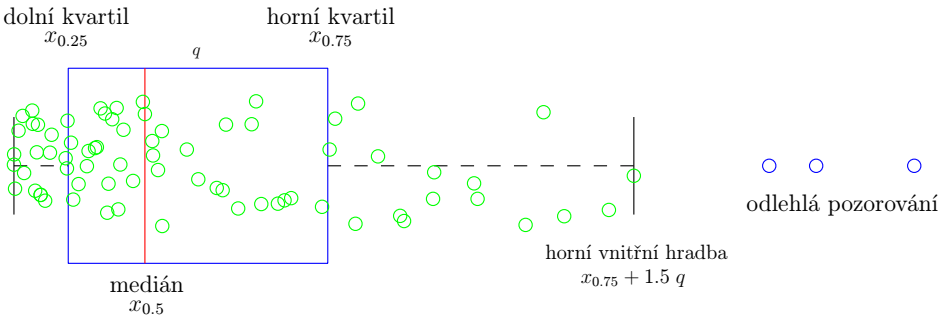
$$x_{0,25} - 3q$$

Horní vnější hradba:

$$x_{0,75} + 3q$$

Odlehlá hodnota je hodnota, která leží mezi vnitřními a vnějšími hradbami. **Extrémní hodnota** je hodnota, která leží za vnějšími hradbami.

ZPŮSOB KONSTRUKCE 4.2. Způsob konstrukce krabicového diagramu je zřejmý z následujícího obrázku.

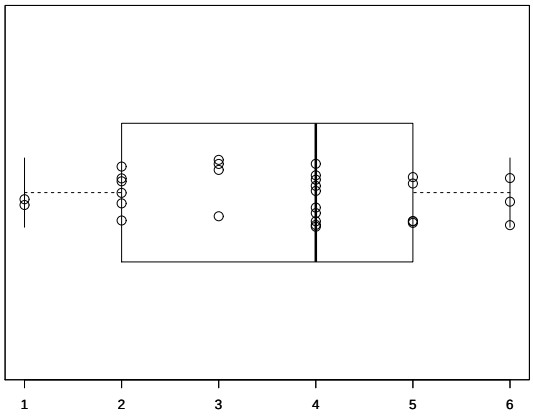


Příklad 4.3. Pro data z Příkladu 2.3 sestrojte krabicový diagram.

Řešení. Rozsah souboru $n = 30$. Výpočty potřebných kvantilů uspořádáme do tabulky.

α	$n\alpha$	c		x_α
0,25	7,5	8	$x_{(c)} = x_{(8)}$	2
0,50	15	15	$\frac{x_{(15)}+x_{(16)}}{2}$	4
0,75	22,5	23	$x_{(c)} = x_{(23)}$	5

$q = 5 - 2 = 3$
Dolní vnitřní hradba: $x_{0,25} - 1,5q = 2 - 1,5.3 = -2,5$
Horní vnitřní hradba: $x_{0,75} + 1,5q = 5 + 1,5.3 = 9,5$
Krabicový graf je znázorněn na Obr. 7.



Obrázek 7: Krabicový diagram

4.2. Normal probability plot (N–P plot). Umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.

ZPŮSOB KONSTRUKCE 4.4. N–P plot konstruujeme tak, že na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na svislou osu kvantily normálního rozdělení u_{α_j}, kde

$$\alpha_j = \frac{3j-1}{3n+1}.$$

Poznámka 4.5. Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.

Poznámka 4.6.

- Pocházejí-li data z normálního rozložení, pak budou všechny dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ ležet na přímce.
- Pro data z rozložení s kladnou šikmostí se budou dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ řadit do konkávní křivky.
- Pro data z rozložení se zápornou šikmostí se budou dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ řadit do konvexní křivky.

Příklady N–P plotu pro jednotlivé případy jsou v části 4.6.

4.3. Quantile – quantile plot (Q–Q plot). Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení.

ZPŮSOB KONSTRUKCE 4.7. Q–Q plot konstruujeme tak, že na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde

$$\alpha_j = \frac{j - r_{adj}}{n + n_{adj}},$$

přičemž r_{adj} a n_{adj} jsou korigující faktory $\leq 0,5$. Implicitně se klade $r_{adj} = 0,375$ a $n_{adj} = 0,25$. Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadují z dat, nebo se volí na základě teoretického modelu. Body $(K_{\alpha_j}(X), x_{(j)})$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím lepší je soulad mezi empirickým a teoretickým rozložením.

Poznámka 4.8. Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.

Příklad 4.9. Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření:
$2 \quad 1,8 \quad 2,1 \quad 2,4 \quad 1,9 \quad 2,1 \quad 2 \quad 1,8 \quad 2,3 \quad 2,2.$
Pomocí N–P plotu a Q–Q plotu ověřte, zda se tato data řídí normálním rozložením.

Řešení.

usp.hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

- **N–P plot:**
  $j = (1,5;\ 3;\ 4,5;\ 6,5;\ 8;\ 9;\ 10)$
  $\alpha_j = \frac{3j-1}{3n+1} = (0,1129;\ 0,2581;\ 0,4032;\ 0,5968;\ 0,7419;\ 0,8387;\ 0,9355)$
  $u_{\alpha_j} = (-1,2112;\ -0,6493;\ -0,245;\ 0,245;\ 0,6493;\ 0,9892;\ 1,5179)$
- **Q–Q plot:**
  $j = (1,5;\ 3;\ 4,5;\ 6,5;\ 8;\ 9;\ 10)$
  $\alpha_j = \frac{j-0,375}{n+0,25} = (0,1098;\ 0,2561;\ 0,4024;\ 0,5976;\ 0,7439;\ 0,8415;\ 0,939)$
  $u_{\alpha_j} = (-1,2278;\ -0,6554;\ -0,247;\ 0,247;\ 0,6554;\ 1,0005;\ 1,566)$

Vzhled obou grafů nasvědčuje tomu, že data pocházejí z normálního rozložení.

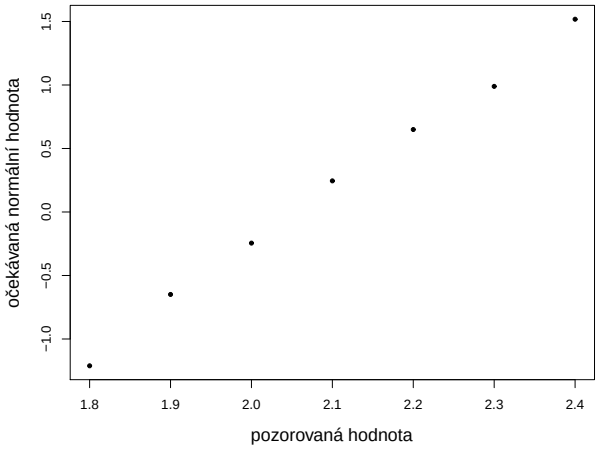
4.4. Probability – probability plot (P–P plot). Používá se ke stejným účelům jako Q–Q plot, ale jinak se konstruuje.

ZPŮSOB KONSTRUKCE 4.10. Spočtou se standardizované hodnoty

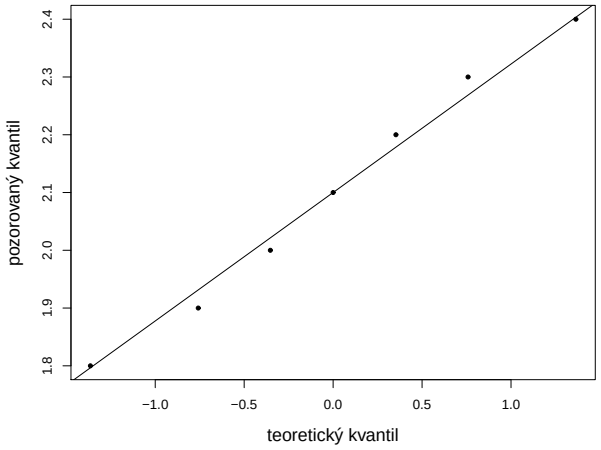
$$z_{(j)} = \frac{x_{(j)} - \bar{x}}{s}, \quad j = 1, \ldots, n.$$

Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce $\langle 0,1\rangle \times \langle 0,1\rangle$, lze usuzovat na dobrou shodu empirického a teoretického rozložení.

Poznámka 4.11. Jsou-li některé hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.



Obr. 8: N–P plot

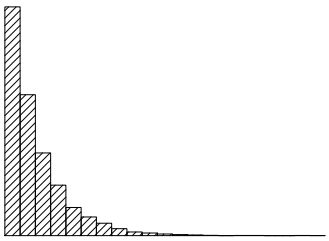


Obr. 9: Q–Q plot

4.5. Histogram. Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. Např. normálního, Pearsonova, Studentova a jiných. Příklady histogramů pro některá rozložení jsou v části 4.6.

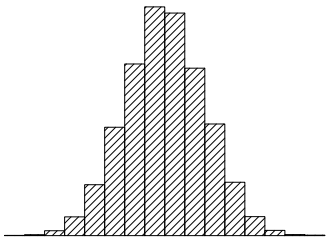
4.6. Vzhled diagnostických grafů pro rozložení s různou šikmostí. Vlastnosti rozložení četností datového souboru se projeví ve vzhledu histogramu, N–P plotu a krabicového diagramu, jak ukazují následující obrázky na straně 14.

Rozložení s kladnou šikmostí



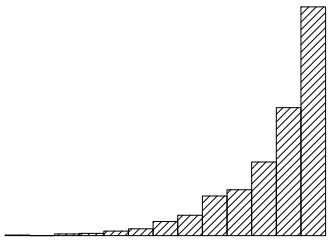
Obr. 10: Histogram

Normální rozložení

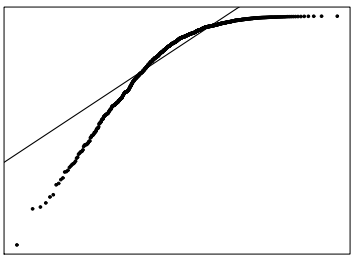


Obr. 11: Histogram

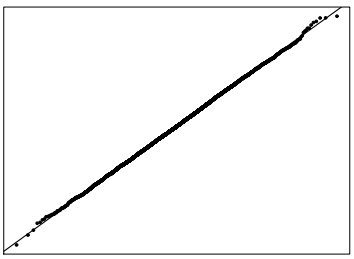
Rozložení se zápornou šikmostí



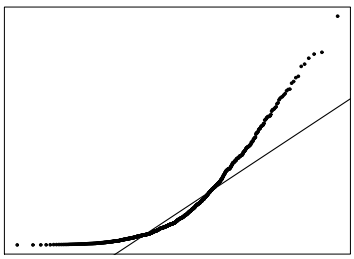
Obr. 12: Histogram



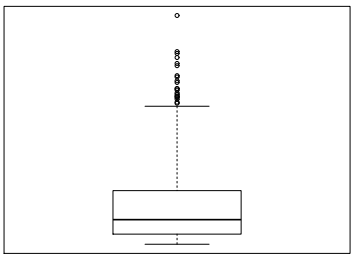
Obr. 13: N–P plot



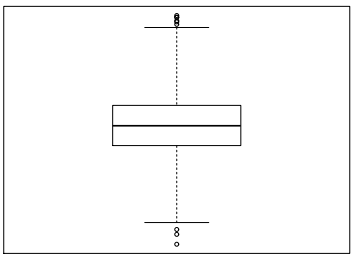
Obr. 14: N–P plot



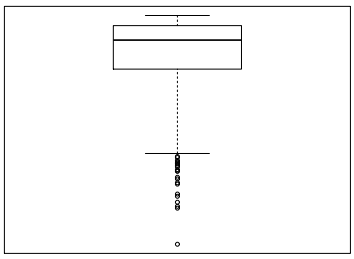
Obr. 15: N–P plot



Obr. 16: Box plot



Obr. 17: Box plot



Obr. 18: Box plot

Úlohy k procvičení

Cvičení 4.1. U 20 studentů 1. ročníku byla zjišťována známka z matematiky na prvním zkušebním termínu.

Známka	1	2	3	4
Počet studentů	7	3	2	8

Vytvořte tabulku rozložení četností. Nakreslete grafy četnostní funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností známek.

Cvičení 4.2. U 60 vzorků oceli byla zjišťována mez plasticity.

Mez plasticity	$\langle 30, 50\rangle$	$\langle 50, 70\rangle$	$\langle 70, 90\rangle$	$\langle 90, 110\rangle$	$\langle 110, 130\rangle$	$\langle 130, 150\rangle$	$\langle 150, 170\rangle$
Počet vzorků	8	4	13	15	9	7	4

Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

Cvičení 4.3. Pro údaje z příkladu 4.2 vypočtěte průměr a rozptyl meze plasticity.
$[\bar{x} = 96,67,\ s^2 = 1148,89]$

Cvičení 4.4. V datovém souboru, z něhož byl vypočten průměr 110 a rozptyl 800, byly zjištěny 2 chyby: místo 85 má být 95 a místo 120 má být 150. Ostatních 18 údajů je správných. Opravte průměr a rozptyl.
$[\bar{x} = 112,\ s^2 = 851]$

Cvičení 4.5. Pro údaje z příkladu 4.1 sestrojte krabicový diagram.
$[x_{0,50} = 2,5,\ x_{0,25} = 1,\ x_{0,75} = 4,\ q = 3$, dolní vnitřní hradba $= -3,5$, horní vnitřní hradba $= 8,5]$

KAPITOLA 2

Základní pojmy matematické statistiky

Základní informace

(1) V následující kapitole se budeme zabývat základními pojmy matematické statistiky. Popíšeme náhodný výběr a uvedeme základní výběrové charakteristiky. Dále se budeme zabývat bodovými a intervalovými odhady parametrů rozdělení, ze kterého náhodný výběr pochází. Zaměříme se na normální rozdělení a také na některá jiná rozdělení. Nakonec se budeme zabývat testováním hypotéz.
(2) Předpokládá se znalost základních pojmů z teorie pravděpodobnosti – náhodná veličina, náhodný vektor, střední hodnota, rozptyl, centrální limitní věta.

Výstupy z výukové jednotky

Studenti

- umí určit nestrannost a konzistenci bodových odhadů
- umí sestrojit a interpretovat interval spolehlivosti pro parametry normálního, alternativniho a Poissonova rozdělení
- umí sestrojit a interpretovat interval spolehlivosti pro rozdíl středních hodnot a podíl rozptylů u dvou výběrů z normálního rozdělení
- umí testovat hypotézy o parametrech normálního, alternativního a Poissonova rozdělení
- umí testovat hypotézy o porovnání dvou parametrů normálního rozdělení

1. Motivace

V teorii pravděpodobnosti se předpokládá, že

- je známý **pravděpodobnostní prostor** $(\Omega, \mathcal{A}, P)$
- a že také známe **rozdělení pravděpodobnosti** náhodných veličin (resp. náhodných vektorů), které na tomto pravděpodobnostním prostoru uvažujeme.

V matematické statistice však

- máme k dispozici výsledky n nezávislých pozorování hodnot sledované náhodné veličiny X, které se ve statistice říká *statistický znak*, tj. máme

$$x_1 = X(\omega_1), \ldots, x_n = X(\omega_n),\ \omega_1, \ldots, \omega_n \in \Omega$$

- a na základě těchto pozorování chceme učinit výpověď o rozdělení zkoumané náhodné veličiny.

Poznámka 1.1. Většina textu v této kapitole byla převzata z [**7**]. Pro podrobnější studium tohoto tématu, zejména důkazy jednotlivých tvrzení, proto odkazujeme na tento zdroj.

2. Náhodný výběr a výběrové charakteristiky

Definujme nejprve základní pojmy matematické statistiky. Základním pojmem matematické statistiky je pojem náhodného výběru.

DEFINICE 2.1. Náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)'$ nazýváme **náhodným výběrem z rozdělení pravděpodobnosti** P, pokud

(i) $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny,
(ii) $X_1, \ldots, X_n$ mají stejné rozdělení pravděpodobnosti P.

Číslo n nazýváme **rozsah náhodného výběru**. Libovolný bod $\mathbf{x} = (x_1, \ldots, x_n)'$, kde x_i je realizace náhodné veličiny X_i $(i = 1, \ldots, n)$, budeme nazývat **realizací náhodného výběru** $\mathbf{X} = (X_1, \ldots, X_n)'$. Množinu všech hodnot, kterých může náhodný výběr nabýt, nazýváme **výběrový prostor** a budeme jej značit $\mathcal{X}$.

Základní dělení matematické statistiky je dané strukturou množiny všech možných rozdělení (označme ji $\mathcal{P}$) náhodného výběru $\mathbf{X}$. Velmi často vybíráme do množiny $\mathcal{P}$ jen rozdělení, která jsou stejného typu a která závisí pouze na nějakém (skalárním či vícerozměrném) parametru. Tento parametr se většinou značí $\boldsymbol{\theta}$ a pravděpodobnostní míry z množiny $\mathcal{P}$ symbolem $P_{\boldsymbol{\theta}}$. Přitom předpokládáme, že parametr $\boldsymbol{\theta}$ nabývá hodnot z nějaké množiny $\boldsymbol{\Theta}$.

DEFINICE 2.2. Množinu $\mathcal{P}$ pravděpodobnostních měr tvaru

$$\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

nazýváme **parametrickou třídou rozdělení**. Vektor $\boldsymbol{\theta}$ nazýváme **parametrem rozdělení pravděpodobnosti** $P_{\boldsymbol{\theta}}$ a množinu $\boldsymbol{\Theta}$ možných hodnot parametru $\boldsymbol{\theta}$ **parametrický prostor**.

Nechť náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ je z rozdělení, které je dáno distribuční funkcí $F(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Zkráceně budeme značit:

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq F(x; \boldsymbol{\theta}).$$

Cílem teorie odhadu je **na základě náhodného výběru** odhadnout

- rozdělení pravděpodobnosti,
- popřípadě některé parametry tohoto rozdělení,
- anebo nalézt odhad nějaké funkce parametrů $\boldsymbol{\theta}$, tj. $\gamma(\boldsymbol{\theta})$.

Funkci $\gamma(\boldsymbol{\theta})$ nazýváme **parametrickou funkcí**. V matematické statistice se pro funkce, pomocí kterých budeme odhady provádět, nazývají statistikou. Tyto funkce jsou navíc měřitelné.

DEFINICE 2.3. Libovolnou náhodnou veličinu T_n, která vznikne jako funkce náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$, budeme nazývat **statistikou**, tj. $T_n = T(X_1, \ldots, X_n)'$.

DEFINICE 2.4. VÝBĚROVÉ CHARAKTERISTIKY. Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr rozsahu n z rozdělení s distribuční funkcí $F(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Potom statistika

$\bar{X}_n = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	se nazývá	**výběrový průměr**
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$		**výběrový rozptyl**
$S = \sqrt{S^2}$		**výběrová směrodatná odchylka**
$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x>}(X_i)$		**výběrová (empirická) distribuční funkce**

3. Bodové odhady

Bodovým odhadem parametrické funkce $\gamma(\boldsymbol{\theta})$ budeme rozumět nějakou statistiku $T_n = T(X_1, \ldots, X_n)'$, která bude pro různé náhodné výběry kolísat kolem $\gamma(\boldsymbol{\theta})$.

Statistika $T_n = T(X_1, \ldots, X_n)'$ závisí na parametru $\boldsymbol{\theta}$ prostřednictvím distribuční funkce rozdělení, z něhož výběr pochází. Také rozdělení této statistiky, tj. náhodné veličiny, závisí na parametru $\boldsymbol{\theta}$. Proto střední hodnotu a rozptyl této statistiky budeme značit $E_{\boldsymbol{\theta}} T_n$ a $D_{\boldsymbol{\theta}} T_n$.

Za lepší odhad se považuje ten, jehož rozdělení je více koncentrované okolo neznámé hodnoty parametru. Tento přirozený požadavek koncentrace rozdělení T_n okolo skutečné hodnoty parametru vyjadřujeme pomocí střední hodnoty a rozptylu.

DEFINICE 3.1. Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení pravděpodobnosti $P_{\boldsymbol{\theta}}$, kde $\boldsymbol{\theta}$ je vektor neznámých parametrů. Nechť $\gamma(\boldsymbol{\theta})$ je daná parametrická funkce.

Řekneme, že statistika $T_n = T(X_1, \ldots, X_n)'$ je

nestranným (nevychýleným)	odhadem parametrické funkce $\gamma(\boldsymbol{\theta})$	pokud pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí $E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$.
kladně vychýleným		$E_{\boldsymbol{\theta}} T_n > \gamma(\boldsymbol{\theta})$.
záporně vychýleným		$E_{\boldsymbol{\theta}} T_n < \gamma(\boldsymbol{\theta})$.
asymptoticky nestranným		$\lim_{n\to\infty} E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$.
(slabě) konzistentním		pokud pro $\forall \varepsilon > 0$ platí $\lim_{n\to\infty} P_{\boldsymbol{\theta}}(\lvert T_n - \gamma(\boldsymbol{\theta})\rvert > \varepsilon) = 0$ tj. $T_n \xrightarrow{P_{\boldsymbol{\theta}}} \gamma(\boldsymbol{\theta})$

Poznámka 3.2. Vlastnost **nestrannosti** (tj. nevychýlenosti) ještě neposkytuje záruku dobrého odhadu, pouze vylučuje systematickou chybu.

Poznámka 3.3 (polopatě)**.** Používání **konzistentních** odhadů zaručuje

- malou pravděpodobnost velké chyby v odhadu parametru, pokud rozsah výběru dostatečně roste;
- volbou dostatečně velkého počtu pozorování lze učinit chybu odhadu libovolně malou.

Příklad 3.4. GEOMETRICKÉ ROZDĚLENÍ.
Nechť náhodná veličina X má geometrické rozdělení,

$$f_X(x) = P(X = x) = (1-\theta)^x \theta \qquad 0 < \theta < 1 \;\; x = 0, 1, \ldots$$

Veličina X udává počet neúspěchů při výběru z alternativního rozdělení před výskytem prvního úspěchu. Nalezněte nestranný odhad pro θ.

Řešení. Je-li $T(X)$ takový nestranný odhad, musí pro něj platit

$$E_\theta T(X) = \sum_{x=0}^{\infty} T(x)(1-\theta)^x \theta = \theta \qquad 0 < \theta < 1,$$

Odtud dostáváme

$$\sum_{x=0}^{\infty} T(x)(1-\theta)^x = 1 \qquad 0 < \theta < 1,$$

takže musí platit

$$\begin{aligned} T(0) &= 1 \\ T(x) &= 0 \qquad \text{pro } x \geq 1. \end{aligned}$$

Tento odhad však není pokládán za vhodný, protože jen minimálně přihlíží k počtu neúspěchů před prvním úspěchem. Závisí jen na tom, zda úspěch nastal hned v prvním pokusu či nikoli. Může se také stát, že nestranný odhad neexistuje, viz následující příklad.

Příklad 3.5. Parametrická funkce $\frac{1}{\theta}$ v případě BINOMICKÉHO ROZDĚLENÍ.

Nechť náhodná veličina X má binomické rozdělení, tj. $X \sim Bi(n, \theta)$ a

$$f_X(x) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad n \geq 1, \;\; 0 < \theta < 1 \;\; x = 0, 1, \ldots, n.$$

Ukažte, že neexistuje nestranný odhad pro parametrickou funkci

$$\gamma(\theta) = \tfrac{1}{\theta}.$$

Řešení. Dokážeme sporem. Nechť existuje taková funkce T, že pro každé $\theta \in (0, 1)$ platí

$$E_\theta T(X) = \sum_{x=0}^{n} T(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = \tfrac{1}{\theta} \qquad 0 < \theta < 1.$$

Na levé straně je však polynom proměnné θ nejvýše stupně n, který samozřejmě nemůže být identicky roven $\frac{1}{\theta}$ na intervalu $(0, 1)$.

Nyní vyšetříme případ, kdy odhadovanými parametry jsou **střední hodnota** a **rozptyl** rozdělení, ze kterého náhodný výběr pochází.

VĚTA 3.6. *Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má střední hodnotu $\mu(\boldsymbol{\theta})$ pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Pak **výběrový průměr** je **nestranným odhadem** střední hodnoty, tj.*

$$E_{\boldsymbol{\theta}}\bar{X} = \mu(\boldsymbol{\theta}).$$

VĚTA 3.7. *Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má rozptyl $\sigma^2(\boldsymbol{\theta})$ pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Pak **výběrový rozptyl** je **nestranným odhadem** rozptylu, tj.*

$$E_{\boldsymbol{\theta}}S^2 = \sigma^2(\boldsymbol{\theta}).$$

Následující věta udává postačující podmínku pro konzistentní odhad.

VĚTA 3.8. *Nechť statistika $T_n = T(X_1, \ldots, X_n)'$ je nestranný nebo asymptoticky nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$ a platí*

$$\lim_{n\to\infty} D_{\boldsymbol{\theta}} T_n = 0.$$

Pak je statistika $T_n = T(X_1, \ldots, X_n)$ konzistentním odhadem parametrické funkce $\gamma(\boldsymbol{\theta})$.

Využitím této věty se dají ukázat následující vlastnosti výběrového průměru a výběrového rozptylu.

Důsledek 3.9. *Nechť* $\mathbf{X} = (X_1, \ldots, X_n)'$ *je náhodný výběr z rozdělení, které má pro* $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ *střední hodnotu* $\mu(\boldsymbol{\theta})$ *a rozptyl* $\sigma^2(\boldsymbol{\theta})$*, tj.*

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

Potom je-li $\mu(\boldsymbol{\theta}) < \infty$*, pak* ***výběrový průměr*** $\bar{X}$ *je* ***slabě konzistentním odhadem*** $\mu(\boldsymbol{\theta})$*.*

Důsledek 3.10. *Nechť* $\mathbf{X} = (X_1, \ldots, X_n)'$ *je náhodný výběr z rozdělení, které má pro* $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ *střední hodnotu* $\mu(\boldsymbol{\theta})$ *a rozptyl* $\sigma^2(\boldsymbol{\theta})$*, tj.*

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

Potom je-li $\sigma^2(\boldsymbol{\theta}) < \infty$*, pak* ***výběrový rozptyl*** S^2 *je* ***slabě konzistentním odhadem*** $\sigma^2(\boldsymbol{\theta})$*.*

Poznámka 3.11. Více nestranných odhadů.
Obecně může existovat více nestranných odhadů. Například nejen výběrový průměr $\bar{X}$ je nestranným odhadem střední hodnoty $\mu(\boldsymbol{\theta})$, ale i každé jednotlivé pozorování X_i nebo každá jeho lineární kombinace $\sum_{i=1}^n c_i X_i$, pro kterou platí $\sum_{i=1}^n c_i = 1$.

Pokud tedy existuje více nestranných odhadů je přirozenou otázkou, který z nich je nejlepší. Za nejlepší můžeme považovat ten, který má **nejmenší rozptyl** mezi všemi nestrannými odhady.

Rozdělení každé statistiky však závisí na parametru $\boldsymbol{\theta}$, z čehož vyplývá, že i rozptyl nestranné statistiky T_n závisí na parametru $\boldsymbol{\theta}$. Může se stát, že odhad minimalizující rozptyl při určité hodnotě parametru není vhodný pro jinou hodnotu parametru – existuje jiný nestranný (nevychýlený) odhad, který má při této hodnotě parametru menší rozptyl. Pokud taková situace nenastane, mluvíme o rovnoměrně nejlepším nestranném odhadu.

Definice 3.12. Nechť T_n je nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$ a pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí

$$D_{\boldsymbol{\theta}} T_n \leq D_{\boldsymbol{\theta}} T_n^*,$$

kde T_n^* je libovolný nestranný odhad parametru $\gamma(\boldsymbol{\theta})$. Potom odhad T_n nazveme (**rovnoměrně**) **nejlepším nestranným odhadem** parametrické funkce $\gamma(\boldsymbol{\theta})$.

Příklad 3.13. Nalezněte nejlepší nestranný lineární odhad střední hodnoty $\mu(\boldsymbol{\theta})$.

Řešení. Jak jsme již dříve spočítali, pro náhodný výběr $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$ platí, že střední hodnota výběrového průměru $\bar{X}$ je rovna

$$E_{\boldsymbol{\theta}} \bar{X} = \mu(\boldsymbol{\theta})$$

a rozptyl výběrového průměru $\bar{X}$ je roven

$$D_{\boldsymbol{\theta}} \bar{X} = \frac{\sigma^2(\boldsymbol{\theta})}{n}.$$

Tedy variabilita této statistiky je n krát menší než variabilita jednotlivých pozorování $X_1, \ldots, X_n$ a tedy hodnoty statistiky $\bar{X}$ jsou více koncentrovány kolem odhadované střední hodnoty $\mu(\boldsymbol{\theta})$ než jednotlivá pozorování $X_1, \ldots, X_n$. Navíc je statistika $\bar{X}$ je lineární funkcí náhodných veličin $X_1, \ldots, X_n$.

Uvažujme všechny lineární statistiky tvaru $\sum_{i=1}^n c_i X_i$, kde $c_1, \dots, c_n \in \mathbb{R}$, které jsou nestrannými odhady střední hodnoty $\mu(\boldsymbol{\theta})$, tj. pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ musí platit

$$\mu(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left(\sum_{i=1}^n c_i X_i \right) = \sum_{i=1}^n c_i \underbrace{E_{\boldsymbol{\theta}} X_i}_{=\mu(\boldsymbol{\theta})} = \mu(\boldsymbol{\theta}) \sum_{i=1}^n c_i \quad \Rightarrow \quad \sum_{i=1}^n c_i = 1.$$

Tím jsme dostali první podmínku, která se týká nestrannosti odhadu. Nyní budeme hledat taková $c_1, \dots, c_n \in \mathbb{R}$, která minimalizují rozptyl

$$D_{\boldsymbol{\theta}} \left(\sum_{i=1}^n c_i X_i \right) \overset{nez.}{=} \sum_{i=1}^n c_i^2 D_{\boldsymbol{\theta}} X_i = \sigma^2(\boldsymbol{\theta}) \sum_{i=1}^n c_i^2$$

a pro něž platí $\sum_{i=1}^n c_i = 1$, tedy hledáme vázaný extrém, takže použijeme Lagrangeovu[1] funkci s multiplikátorem λ, tj.

$$L(c_1, \dots, c_n, \lambda) = \sum_{i=1}^n c_i^2 - \lambda \left(\sum_{i=1}^n c_i - 1 \right).$$

Pak pro $j = 1, \dots, n$

$$\frac{\partial L}{\partial c_j} = 2c_j - \lambda = 0 \quad \Rightarrow \quad c_j = \tfrac{1}{2}\lambda$$

$$\frac{\partial L}{\partial \lambda} = -\sum_{i=1}^n c_i + 1 = 0 \quad \Rightarrow \quad \sum_{i=1}^n c_i = 1.$$

Prvních n rovnic implikuje, že

$$c_1 = c_2 = \cdots = c_n.$$

Označme společnou hodnotu symbolem c. Díky poslední rovnici dostaneme

$$1 = \sum_{i=1}^n c_i = nc \quad \Rightarrow \quad c = c_1 = c_2 = \cdots = c_n = \tfrac{1}{n},$$

tedy výběrový průměr $\bar{X}$ je **nejlepším nestranným lineárním odhadem** střední hodnoty $\mu(\boldsymbol{\theta})$.

Zkusme provést důkaz ještě jiným způsobem. Nechť $\sum\limits_{i=1}^n c_i X_i$ je libovolný nestranný lineární odhad pro μ (tj. nutně musí platit $\sum\limits_{i=1}^n c_i = 1$).

Položíme-li	$c_i = \frac{1}{n} + \delta_i$	pro	$i = 1, \dots, n$
je minimalizace výrazu	$\sum\limits_{i=1}^n c_i^2$	za podmínky	$\sum\limits_{i=1}^n c_i = 1$
ekvivalentní s úlohou minimalizovat	$\sum\limits_{i=1}^n \left(\frac{1}{n} + \delta_i\right)^2$	za podmínky	$\sum\limits_{i=1}^n \delta_i = 0.$

Za této podmínky je však

$$\sum_{i=1}^n \left(\tfrac{1}{n} + \delta_i\right)^2 = \underbrace{\sum_{i=1}^n \left(\tfrac{1}{n}\right)^2}_{=n\frac{1}{n^2}} + 2\tfrac{1}{n} \underbrace{\sum_{i=1}^n \delta_i}_{=0} + \sum_{i=1}^n \delta_i^2 = \tfrac{1}{n} + \sum_{i=1}^n \delta_i^2,$$

což je minimální pro

$$\delta_i = 0 \qquad \text{pro} \qquad i = 1, \dots, n.$$

Tedy nejlepším nestranným lineárním odhadem je lineární kombinace X_i s koeficienty $c_i = \frac{1}{n}$.

[1] Joseph Louis Lagrange (1736-1813) – italsko–francouzský matematik a astronom

4. Intervalové odhady

Odhady, jimiž jsme se doposud zabývali, se někdy nazývají **bodové odhady** parametrické funkce $\gamma(\boldsymbol{\theta})$. Je tomu tak proto, že pro danou realizaci náhodného výběru $x_1, \ldots, x_n$ představuje odhad daný statistikou $T_n(x_1, \ldots, x_n)$ **jediné číslo** (**bod**), které je v jistém smyslu přiblížením ke skutečné hodnotě parametrické funkce $\gamma(\boldsymbol{\theta})$.

Úlohu odhadu však lze formulovat i jiným způsobem. Jde o to, sestrojit na základě daného náhodného výběru takový interval, jehož **hranice** jsou **statistiky**, a který se s dostatečně velkou přesností pokryje skutečnou hodnotu parametrické funkce $\gamma(\boldsymbol{\theta})$. V tomto případě mluvíme o **intervalovém odhadu** parametrické funkce $\gamma(\boldsymbol{\theta})$.

Podobná je úloha zkonstruovat na základě náhodného výběru statistiku, o níž lze s dostatečně velkou spolehlivostí prohlásit, že skutečná hodnota parametrické funkce je větší než tato statistika. V tomto případě mluvíme o **dolním odhadu** parametrické funkce $\gamma(\boldsymbol{\theta})$. Analogicky lze zavést pomocí opačné nerovnosti pojem **horního odhadu** $\gamma(\boldsymbol{\theta})$.

DEFINICE 4.1. Nechť $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq F(x; \boldsymbol{\theta})$ je náhodný výběr rozsahu n z rozdělení o distribuční funkci $F(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Dále mějme parametrickou funkci $\gamma(\boldsymbol{\theta})$, $\alpha \in (0, 1)$ a statistiky $D = D(X_1, \ldots, X_n)$ a $H = H(X_1, \ldots, X_n)$.

Potom intervaly $\langle D, H\rangle$ nazveme $100(1-\alpha)$ % **intervalem spolehlivosti** pro parametrickou funkci $\gamma(\boldsymbol{\theta})$ jestliže

$$P_{\boldsymbol{\theta}}(D(X_1, \ldots, X_n) \leq \gamma(\boldsymbol{\theta}) \leq H(X_1, \ldots, X_n)) = 1 - \alpha$$

Jestliže

$$P_{\boldsymbol{\theta}}(D(X_1, \ldots, X_n) \leq \gamma(\boldsymbol{\theta})) = 1 - \alpha,$$

pak statistiku $D = D(X_1, \ldots, X_n)$ nazýváme **dolním odhadem parametrické funkce** $\gamma(\boldsymbol{\theta})$ se spolehlivostí $1 - \alpha$ (nebo s rizikem α).

Jestliže

$$P_{\boldsymbol{\theta}}(\gamma(\boldsymbol{\theta}) \leq H(X_1, \ldots, X_n)) = 1 - \alpha$$

pak statistiku $H = H(X_1, \ldots, X_n)$ nazýváme **horním odhadem parametrické funkce** $\gamma(\boldsymbol{\theta})$ se spolehlivostí $1 - \alpha$ (nebo s rizikem α).

Poznámka 4.2. (polopatě) Vysvětleme si nyní smysl pojmu **spolehlivost intervalových odhadů**.
Konkrétní data $x_1, \ldots, x_n$ (tj. realizace náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$) nejsou náhodnými veličinami, nýbrž jsou to výsledky určitého pokusu ω, tj.

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

Sestrojíme-li tedy na jejich základě intervalový odhad, řekněme (a, b), parametrické funkce $\gamma(\theta)$, pak nemá smysl mluvit o pravděpodobnosti $P(a < \gamma(\theta) < b)$, protože všechny tři symboly jsou reálná čísla (třebaže $\gamma(\theta)$ neznáme) a nerovnost $a < \gamma(\theta) < b$ buď platí nebo neplatí, tj. náš intervalový odhad je buď správný nebo nesprávný.

Budeme-li však sestrojovat intervalové odhady vícekrát po sobě, pak poměrná četnost případů, kdy intervalový odhad bude správný, bude přibližně rovna $1 - \alpha$.

Číslo $\boxed{\alpha}$ se volí poměrně malé, nejčastěji $\begin{matrix} 0.05 \\ 0.01 \end{matrix}$ spolehlivost je pak $\begin{matrix} 0.95 \\ 0.99 \end{matrix}$ tj. $\begin{matrix} 95\% \\ 99\% \end{matrix}$

Kromě dostatečné spolehlivosti bychom chtěli, aby interval $\langle D_n(\mathbf{X}), T_n(\mathbf{X})\rangle$ byl co možná nejkratší.

Tyto požadavky jsou však (při pevném rozsahu výběru n) protichůdné. Žádáme-li větší spolehlivost, musíme se smířit s delším intervalem; žádáme-li naopak kratší interval, musíme se smířit s nižší spolehlivostí.

NÁVOD 4.3 (konstrukce intervalových odhadů). Popíšeme nyní jednu metodu konstrukce intervalových odhadů, která je použitelná ve většině případů.

(1) Najdeme nějakou tzv. PIVOTOVOU STATISTIKU , tj. funkci $\boxed{h}$ náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$ a parametrické funkce $\gamma(\theta)$, tedy náhodnou veličinu

$$\boxed{h(\mathbf{X}, \gamma(\theta))},$$

tak aby její rozdělení již **nezáviselo na parametru** θ.

(2) Nechť $q_{\alpha/2}$ a $q_{1-\alpha/2}$ jsou kvantily rozdělení statistiky

$$h(\mathbf{X}, \gamma(\theta)).$$

Pak pro všechna θ platí

$$P_\theta(q_{\alpha/2} < h(\mathbf{X}, \gamma(\theta)) \leq q_{1-\alpha/2}) = 1 - \alpha$$

(3) Jestliže lze nerovnosti v závorce převést ekvivalentními úpravami na tvar, kde mezi nerovnostmi stojí jen $\gamma(\theta)$, pak jsme sestrojili intervalový odhad

$$D_n(\mathbf{X}) \leq \gamma(\theta) \leq H_n(\mathbf{X})$$

o spolehlivosti $1 - \alpha$.

Tedy, je-li $h(\mathbf{X}, \gamma(\theta))$ **ryze monotónní funkce**, pak existuje inverzní funkce

$$h^{-1}\left(h(\mathbf{X}, \gamma(\theta))\right) = \gamma(\theta).$$

(a) Pokud je $h(\mathbf{X}, \gamma(\theta))$ **rostoucí funkce**, pak platí

$$P_\theta(h^{-1}(q_{\alpha/2}) \leq \gamma(\theta) \leq h^{-1}(q_{1-\alpha/2})) = 1 - \alpha.$$

(b) Pokud je $h(\mathbf{X}, \gamma(\theta))$ **klesající funkce**, pak platí

$$P_\theta(h^{-1}(q_{1-\alpha/2}) \leq \gamma(\theta) \leq h^{-1}(q_{\alpha/2})) = 1 - \alpha.$$

Při konstrukci intervalových odhadů hrají důležitou roli kvantily. Tabulka 1 udává jejich značení pro některá rozdělení. Navíc je dobré si uvědomit následující vlastnost.

Φ	distribuční funkce standardizovaného normálního rozdělení
G_n	distribuční funkce rozdělení χ^2 o n stupních volnosti
H_n	distribuční funkce Studentova rozdělení o n stupních volnosti
$Q_{n,m}$	distribuční funkce Fisherova–Snedecorova rozdělení o n a m stupních volnosti
u_α	kvantily standardizovaného normálního rozdělení
$\chi^2_\alpha(\nu)$	kvantily rozdělení χ^2 o ν stupních volnosti
$t_\alpha(\nu)$	kvantily Studentova rozdělení o ν stupních volnosti
$F_\alpha(\nu_1, \nu_2)$	kvantily Fisherova–Snedecorova rozdělení o ν_1 a ν_2 stupních volnosti

Tabulka 1: Kvantily některých důležitých rozdělení

Je-li distribuční funkce F absolutně spojitá a ryze monotónní a je-li příslušná hustota f **sudá funkce**, pak platí

$$F(x) = 1 - F(-x) \qquad x \in \mathbb{R}$$

a odtud

$$x_\alpha = -x_{1-\alpha} \qquad \alpha \in (0,1),$$

což speciálně platí pro **normální** a **Studentovo rozdělení**.

5. Bodové a intervalové odhady parametrů normálního rozdělení

Nechť $k, n \in \mathbb{N},\ \nu, \nu_1, \nu_2, \ldots, \nu_k \in \mathbb{N},\ b_0, b_1, \ldots, b_n \in \mathbb{R}, \exists\, i \in \{1, \ldots, n\} : b_i \neq 0$ Připomeňme, že platí:

Normální rozdělení s hustotou

$$\boxed{X \sim N(\mu, \sigma^2) \sim f(x) = \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} \qquad x \in \mathbb{R}$$

má střední hodnotu $EX = \mu$ a rozptyl $DX = \sigma^2$. Toto rozdělení má následující vlastnosti:

$$\perp\!\!\!\perp \{X_1, \ldots, X_n\} \ \wedge\ X_i \sim N(\mu_i, \sigma_i^2) \quad \Rightarrow \quad b_0 + \sum_{i=1}^n b_i X_i \sim N\left(b_0 + \sum_{i=1}^n b_i \mu_i, \sum_{i=1}^n b_i^2 \sigma_i^2\right)$$

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad U = \tfrac{X-\mu}{\sigma} \sim N(0,1)$$

χ^2 **rozdělení**:

$$\perp\!\!\!\perp \{U_1, \ldots, U_\nu\} \simeq N(0,1) \quad \Rightarrow \quad K = U_1^2 + \cdots + U_\nu^2 \sim \chi^2(\nu)$$

$$\perp\!\!\!\perp \{K_1 \sim \chi^2(\nu_1), \ldots, K_k \sim \chi^2(\nu_k)\} \quad \Rightarrow \quad K = K_1 + \cdots + K_k \sim \chi^2(\nu_1 + \cdots + \nu_k)$$

Studentovo t-rozdělení:

$$U \sim N(0,1) \perp K \sim \chi^2(\nu) \quad \Rightarrow \quad T = \frac{U}{\sqrt{\frac{K}{\nu}}} \sim t(\nu)$$

Fisherovo–Snedecorovo F-rozdělení:

$$K_1 \sim \chi^2(\nu_1) \perp K_2 \sim \chi^2(\nu_2) \quad \Rightarrow \quad F = \frac{K_1/\nu_1}{K_2/\nu_2} \sim F(\nu_1, \nu_2)$$

VĚTA 5.1. *Mějme* $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq N(\mu, \sigma^2)$ *a výběrový průměr* $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *a výběrový rozptyl* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. *Pak platí*

(1) *Výběrový průměr* $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

(2) *Statistika* $U = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$

(3) *Statistika* $K = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$

(4) *Statistika* $T = \frac{\bar{X}-\mu}{S}\sqrt{n} \sim t(n-1)$

Poznámka 5.2. Statistiky $\boxed{U}$, $\boxed{K}$ a $\boxed{T}$ se nazývají PIVOTOVÉ STATISTIKY, přičemž

$U = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ je pivotovou stastistikou pro neznámý parametr μ při známém σ

$K = \frac{n-1}{\sigma^2} S^2$ - ” - σ^2

$T = \frac{\bar{X}-\mu}{S}\sqrt{n}$ - ” - μ při neznámém σ

DŮSLEDEK 5.3. *Mějme* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$, *kde* μ *je* ***neznámý parametr*** *a* $\sigma^2 \in \mathbb{R}$ *je* ***známé*** *reálné číslo. Pak*

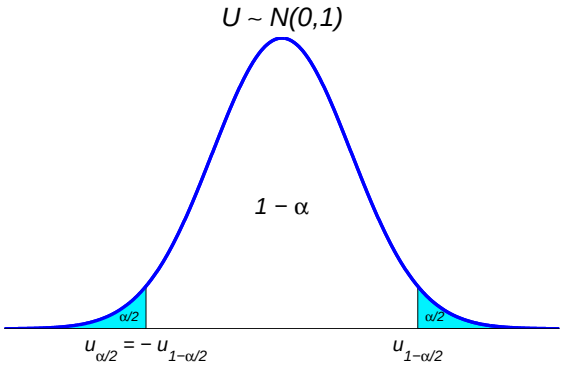
$\langle \bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\rangle$ - *je* $100(1-\alpha)\%$ *interval spolehlivosti pro střední hodnotu* μ *při známém* σ^2

$\bar{X} - u_{1-\alpha}\frac{\sigma}{\sqrt{n}}$ - *je* ***dolní odhad*** *střední hodnoty* μ *při známém* σ^2 *se spolehlivostí* $1-\alpha$

$\bar{X} + u_{1-\alpha}\frac{\sigma}{\sqrt{n}}$ - *je* ***horní odhad*** *střední hodnoty* μ *při známém* σ^2 *se spolehlivostí* $1-\alpha$

Důkaz. Za pivotovou statistiku zvolíme statistiku

$$U = U_{\bar{X}} = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1).$$



Pro lepší čitelnost místo $P_\theta = P_\mu$ budeme psát pouze P.

Počítejme

$$\begin{aligned} 1-\alpha &= P(u_{\frac{\alpha}{2}} \le U \le u_{1-\frac{\alpha}{2}}) \\ &= P(u_{\frac{\alpha}{2}} \le \tfrac{\bar{X}-\mu}{\sigma}\sqrt{n} \le u_{1-\frac{\alpha}{2}}) \\ &= P(\bar{X} - u_{1-\alpha/2}\tfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + u_{1-\alpha/2}\tfrac{\sigma}{\sqrt{n}}) \end{aligned}$$

□

DŮSLEDEK 5.4. *Mějme* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$, *kde* μ *a* σ^2 *jsou* ***neznámé parametry***. *Pak*

(1) pro střední hodnotu $\boxed{\mu}$

$\langle \bar{X} - t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\rangle$ - *je* $100(1-\alpha)\%$ *interval spolehlivosti pro střední hodnotu* μ *při neznámém* σ^2

$\bar{X} - t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}$ - *je* ***dolní odhad*** *střední hodnoty* μ *při neznámém* σ^2 *se spolehlivostí* $1-\alpha$

$\bar{X} + t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}$ - *je* ***horní odhad*** *střední hodnoty* μ *při neznámém* σ^2 *se spolehlivostí* $1-\alpha$

(2) pro rozptyl $\boxed{\sigma^2}$

$\left\langle \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right\rangle$ - *je* $100(1-\alpha)\%$ *interval spolehlivosti pro rozptyl* σ^2

$\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}$ - *je* ***dolní odhad*** *rozptylu* σ^2 *se spolehlivostí* $1-\alpha$

$\frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}$ - *je* ***horní odhad*** *rozptylu* σ^2 *se spolehlivostí* $1-\alpha$

V dalším si budeme všímat intervalů spolehlivosti pro DVA NEZÁVISLÉ VÝBĚRY.

VĚTA 5.5. *Nechť* $\perp\!\!\!\perp\{X_1,\dots,X_{n_1}\} \sim N(\mu_1,\sigma_1^2)$ *je náhodný výběr rozsahu* n_1 *z normálního rozdělení* $N(\mu_1,\sigma_1^2)$, $\bar{X}$ *je jeho výběrový průměr a* S_1^2 *jeho výběrový rozptyl.*

Dále nechť $\perp\!\!\!\perp\{Y_1,\dots,Y_{n_2}\} \sim N(\mu_2,\sigma_2^2)$ *je náhodný výběr rozsahu* n_2 *z normálního rozdělení* $N(\mu_2,\sigma_2^2)$, $\bar{Y}$ *je jeho výběrový průměr a* S_2^2 *jeho výběrový rozptyl.*

Předpokládejme, že oba výběry jsou stochasticky nezávislé, tj. $\mathbf{X} \perp \mathbf{Y}$. *Pak*

(1) Statistika

$$U_{\bar{X}-\bar{Y}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

(2) Pokud $\sigma_1^2=\sigma_2^2=\sigma^2$, *pak statistika*

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_{12}}\sqrt{\frac{n_1n_2}{n_1+n_2}} \sim t(n_1+n_2-2), \text{ kde } S_{12}^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

(3) Statistika

$$F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1,n_2-1).$$

DŮSLEDEK 5.6. *Nechť* $\perp\!\!\!\perp\{X_1,\dots,X_{n_1}\} \sim N(\mu_1,\sigma_1^2)$ *je náhodný výběr rozsahu* n_1 *z normálního rozdělení* $N(\mu_1,\sigma_1^2)$, $\bar{X}$ *je jeho výběrový průměr a* S_1^2 *jeho výběrový rozptyl.*

Dále nechť $\perp\!\!\!\perp\{Y_1,\dots,Y_{n_2}\} \sim N(\mu_2,\sigma_2^2)$ *je náhodný výběr rozsahu* n_2 *z normálního rozdělení* $N(\mu_2,\sigma_2^2)$, $\bar{Y}$ *je jeho výběrový průměr a* S_2^2 *jeho výběrový rozptyl.*

Předpokládejme, že oba výběry jsou stochasticky nezávislé, tj. $\mathbf{X} \perp \mathbf{Y}$. *Pak*

(1) jsou-li $\boxed{\sigma_2^2 \text{ a } \sigma_1^2 \text{ známé}}$, *pak* $100(1-\alpha)\%$ *interval spolehlivosti pro rozdíl středních hodnot* $\mu_1-\mu_2$ *je tvaru*

$$\left\langle \bar{X}-\bar{Y}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \bar{X}-\bar{Y}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right\rangle.$$

(2) Jestliže $\boxed{\sigma_2^2 \text{ a } \sigma_1^2 \text{ nejsou známé}}$ *a platí* $\boxed{\sigma_2^2=\sigma_1^2=\sigma^2}$, *pak* $100(1-\alpha)\%$ *interval spolehlivosti pro rozdíl středních hodnot* $\mu_1-\mu_2$ *je tvaru*

$$\left\langle \bar{X}-\bar{Y}-t_{1-\frac{\alpha}{2}}(n_1+n_2-2)\,S_{12}\sqrt{\tfrac{n_1+n_2}{n_1n_2}}, \bar{X}-\bar{Y}+t_{1-\frac{\alpha}{2}}(n_1+n_2-2)\,S_{12}\sqrt{\tfrac{n_1+n_2}{n_1n_2}} \right\rangle,$$

kde

$$S_{12}^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

(3) Při $\boxed{\text{neznámých } \mu_1,\mu_2,\sigma_1^2,\sigma_2^2}$ *je* $100(1-\alpha)\%$ *interval spolehlivosti pro podíl rozptylů* $\frac{\sigma_1^2}{\sigma_2^2}$ *roven*

$$\left\langle \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)} \right\rangle.$$

Poznámka 5.7. Ve statistických tabulkách bývají uváděny kvantily F-rozdělení pouze pro hodnoty $\alpha \geq 0.5$. Ukážeme, proč není třeba uvádět hodnoty kvantilů pro $\alpha < 0.5$. Uvažujme místo pivotové statistiky F statistiku

$$F^* = \frac{S_2^2}{S_1^2}\frac{\sigma_1^2}{\sigma_2^2} = \frac{1}{F} \sim F(n_2-1,n_1-1).$$

Opět označme $\nu_1 = n_1 - 1$ a $\nu_2 = n_2 - 1$ a počítejme interval spolehlivosti pro takto navrženou pivotovou statistiku

$$1-\alpha = P(F_{\frac{\alpha}{2}}(\nu_2,\nu_1)) \le F^* \le F_{1-\frac{\alpha}{2}}(\nu_2,\nu_1)) = P\left(F_{\frac{\alpha}{2}}(\nu_2,\nu_1)) \le \frac{S_2^2}{S_1^2}\frac{\sigma_1^2}{\sigma_2^2} \le F_{1-\frac{\alpha}{2}}(\nu_2,\nu_1))\right)$$
$$= P\left(\frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2}}(n_2-1,n_1-1) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{1-\frac{\alpha}{2}}(n_2-1,n_1-1)\right)$$

Takže $\boxed{F_{1-\frac{\alpha}{2}}(n_2-1,n_1-1) = \dfrac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}}$ a interval spolehlivosti pro $\boxed{\frac{\sigma_1^2}{\sigma_2^2}}$ lze vyjádřit i takto

$$\left\langle \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}F_{1-\frac{\alpha}{2}}(n_2-1,n_1-1)\right\rangle.$$

V dalším se zaměříme na interval spolehlivosti pro rozdíl středních hodnot u tzv. PÁROVÝCH VÝBĚRŮ.

VĚTA 5.8. *Nechť* $\mathbf{X}_1 = (X_1,Y_1)', \dots, \mathbf{X}_n = (X_n,Y_n)'$ *je náhodný výběr z dvourozměrného normálního rozdělení* $N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *s parametry* $\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$ *a* $\boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$, *kde* $\mu_1,\mu_2 \in \mathbb{R}$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$ *a* $\rho \in (0,1)$.

Pro $i = 1,\dots,n$ *označme*
$$Z_i = X_i - Y_i$$
$$\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$$
$$S_Z^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2.$$

Pak
$$\left\langle \bar{Z} - t_{1-\frac{\alpha}{2}}(n-1)\frac{S_Z}{\sqrt{n}}, \bar{Z} + t_{1-\frac{\alpha}{2}}(n-1)\frac{S_Z}{\sqrt{n}}\right\rangle$$
je intervalový odhad parametrické funkce $\mu_1 - \mu_2$ *o spolehlivosti* $1-\alpha$.

6. Bodové a intervalové odhady založené na centrální limitní větě

Odhady parametrů normálního rozdělení, které jsme doposud zkoumali, mají díky **centrální limitní větě** (CLV) širší použití.

Často lze najít takovou **transformaci** $\boxed{h}$, že náhodná veličina $\boxed{h(\mathbf{X},\gamma(\boldsymbol{\theta}))}$ má pro $n \to \infty$ **asymptoticky** standardizované normální rozdělení $\boxed{N(0,1)}$, tj.

$$h(\mathbf{X},\gamma(\boldsymbol{\theta})) \overset{A}{\sim} N(0,1)$$

Přitom rozdělení, z něhož výběr pochází

- nemusí splňovat požadavky **spojitosti** a **ryzí monotonie** distribuční funkce,
- může být i diskrétní.

Bodové i intervalové odhady lze pak sestrojit stejným způsobem jako v případě normálních náhodných výběrů, jejich **spolehlivost** bude $\boxed{1-\alpha}$ jen přibližně, tj. **asymptoticky**.

VĚTA 6.1. *Mějme* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}),\sigma^2(\boldsymbol{\theta}))$ *a výběrový průměr* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. *Nechť* $S_*^2 = S_*^2(\mathbf{X})$ *je* ***(slabě) konzistentním odhadem*** *rozptylu* $\sigma^2(\boldsymbol{\theta})$. *Pak statistika*

$$U_* = \frac{\bar{X}-\mu(\boldsymbol{\theta})}{S_*}\sqrt{n} \overset{A}{\sim} N(0,1).$$

DŮSLEDEK 6.2. ***(Binární náhodné výběry)***. *Nechť* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq A(p)$ *je náhodný výběr s alternativním (binárním) rozdělením. Potom intervalovým odhadem parametru* $\boxed{p}$ *o asymptotické spolehlivosti* $1-\alpha$ *je interval*

$$\left\langle \bar{X} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right\rangle.$$

DŮSLEDEK 6.3. ***(Poissonovské náhodné výběry)***. *Nechť* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq Po(\lambda)$ *je náhodný výběr s Poisonovým rozdělením. Potom intervalovým odhadem parametru* $\boxed{\lambda}$ *(*$0<\lambda<\infty$*) o asymptotické spolehlivosti* $1-\alpha$ *je interval*

$$\left\langle \bar{X} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}}{n}}, \bar{X} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}}{n}} \right\rangle.$$

7. Testování statistických hypotéz

7.1. Úvod. Mějme náhodný výběr $\mathbf{X} = (X_1,\dots,X_n)'$ rozsahu n z rozdělení o distribuční funkci $F(x;\boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1,\dots,\theta_m)' \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Množina $\boldsymbol{\Theta}$ nechť je neprázdná a otevřená.

Předpokládejme, že o parametru $\boldsymbol{\theta}$ existují dvě konkurující si hypotézy: H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$; H_1: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1 = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$

Tvrzení $\boxed{H_0}$ se nazývá **nulovou hypotézou**. / $\boxed{H_1}$ **alternativní hypotézou**.

Je-li $\boxed{\boldsymbol{\Theta}_0}$ / $\boxed{\boldsymbol{\Theta}_1}$ **jednobodová**, nazývá se **jednoduchou**, v opačném případě **složenou hypotézou**.

O platnosti této hypotézy se má rozhodnout na základě náhodného výběru $\mathbf{X} = (X_1,\dots,X_n)'$, a to tak, že ↗ **zamítneme** *nebo* ↘ **nezamítneme** platnost hypotézy H_0.

Na testování použijeme statistiku $T_n = T(\mathbf{X})$, kterou nazýváme **testovací statistikou**. Množinu hodnot, které může testovací statistika nabýt, rozdělíme na dvě disjunktní oblasti. Jednu označíme W_α , a nazveme ji **kritickou oblastí** (nebo také *oblastí zamítnutí hypotézy*) a druhá je doplňkovou oblastí (*oblast nezamítnutí testované hypotézy*).

Na základě realizace náhodného výběru $\mathbf{x} = (x_1,\dots,x_n)'$ vypočítáme hodnotu testovací statistiky $t_n = T(\mathbf{x})$.

- Pokud hodnota testovací statistiky t_n nabude hodnoty z kritické oblasti, tj. $\boxed{t_n = T(\mathbf{x}) \in W_\alpha}$, pak **nulovou hypotézu zamítáme**.
- Pokud hodnota testovací statistiky nabude hodnoty z oblasti nezamítnutí, tj. $\boxed{t_n = T(\mathbf{x}) \notin W_\alpha}$, tak **nulovou hypotézu nezamítáme**, což ovšem neznamená že přijímáme alternativu.

Toto rozhodnutí nemusí však být správné. V následující tabulce jsou uvedeny možné situace

H_0	PLATÍ	NEPLATÍ
ZAMÍTÁME $t_n = T(\mathbf{x}) \in W_\alpha$	***chyba 1. druhu*** (α_0 je ***hladina testu***) $\alpha_0 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_0) \le \alpha$	O.K. (tzv. ***síla testu*** či ***silofunkce***) $1-\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_1)$ pro $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$
NEZAMÍTÁME $t_n = T(\mathbf{x}) \notin W_\alpha$	O.K.	***chyba 2. druhu*** $\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\mathbf{X}) \notin W_\alpha \mid H_1)$ pro $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$

Volba kritického oboru W_α se řídí požadavky:

(1) Chceme, aby pravděpodobnost chyby 1. druhu byla menší nebo rovna předem zvolenému malému $\alpha \in (0,1)$ (obvykle se volí $\alpha = 0.01$ nebo $\alpha = 0.05$), tj. aby platilo pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$$\alpha_0 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_0) \le \alpha.$$

Pro spojitá rozdělení je vždy možné (i když ne nutné) zvolit test, jehož hladina je právě rovna α. U diskrétních rozdělení jsou možnými hladinami testu jen některé diskrétní hodnoty. Není-li zvolená hladina mezi nimi, rozhodneme se pro hladinu, která je nejbližší nižší (nebo nejbližší vyšší).

(2) Mezi testy na hladině α se pak snažíme zvolit test s co **nejmenší pravděpodobností chyby druhého druhu**, tj. **co nejsilnější test**.

Vidíme, že postavení obou hypotéz je nesymetrické. Za **nulovou hypotézu** volíme tu, jejíž **neoprávněné zamítnutí** (chyba 1. druhu) **je závažnější**.

DEFINICE 7.1. Chybu, která spočívá **v nesprávném zamítnutí nulové hypotézy, i když je správná**, budeme nazývat **chybou prvního druhu**, pravděpodobnost

$$\alpha_0 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_0)$$

nazveme **hladinou významnosti** (též **hladinou testu**).

Chybu, která spočívá **v nesprávném přijetí nulové hypotézy, i když neplatí**, budeme nazývat **chybou druhého druhu** a její pravděpodobnost pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ označíme

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\mathbf{X}) \notin W_\alpha \mid H_1)\ .$$

Pravděpodobnost $1-\beta(\boldsymbol{\theta})$ nazýváme **silou testu** (též **silou kritické oblasti** W_α) a jakožto funkci $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ ji také nazveme **silofunkcí testu**.

7.2. Vztah mezi testy a intervalovými odhady.

Mějme náhodný výběr $\mathbf{X} = (X_1, \dots, X_n)'$ rozsahu n z rozdělení, které závisí na parametru $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)' \in \boldsymbol{\Theta}$ a parametrickou funkci $\gamma(\boldsymbol{\theta})$.

(A) Hypotéza $H_0 : \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti (tzv. *oboustranné*) alternativě $H_1 : \gamma(\boldsymbol{\theta}) \neq \gamma(\boldsymbol{\theta}_0)$:

Mějme **intervalový odhad** $(D_n(\mathbf{X}), H_n(\mathbf{X}))$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $1-\alpha$. Pokud platí nulová hypotéza, pak

$$1 - \alpha = P_{\boldsymbol{\theta}}\left(D_n(\mathbf{X}) \le \gamma(\boldsymbol{\theta}_0) \le H_n(\mathbf{X})\right),$$

takže **kritický obor** tohoto testu má tvar:

$$W_\alpha = \{\mathbf{X} \in \mathbb{R}^n : \gamma(\boldsymbol{\theta}_0) \notin (D_n(\mathbf{X}), H_n(\mathbf{X}))\}\,.$$

Zjistíme-li v konkrétní situaci, že

$$\gamma(\boldsymbol{\theta}_0) \notin (d_n(\mathbf{x}), h_n(\mathbf{x})) \text{ tj. realizace } \mathbf{x} \in W_\alpha ,$$

potom

- buď nastal jev, který má pravděpodobnost α (volí se blízká nule),
- nebo neplatí nulová hypotéza.

Protože při obvyklé volbě $\alpha = 0.05$ nebo $\alpha = 0.01$ je tento jev „prakticky nemožný“, proto nulovou hypotézu H_0 **zamítáme ve prospěch alternativy** H_1.

V opačném případě, tj. pokud

$$\gamma(\boldsymbol{\theta}_0) \in (d_n(\mathbf{x}), h_n(\mathbf{x})) \text{ tj. realizace } \mathbf{x} \notin W_\alpha ,$$

nulovou hypotézu H_0 **nezamítáme**.

(B) Hypotéza $H_0 : \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti (tzv. *jednostranné*) alternativě $H_1 : \gamma(\boldsymbol{\theta}) > \gamma(\boldsymbol{\theta}_0)$:

V tomto případě využijeme **dolní odhad** $D_n(\mathbf{X})$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $1 - \alpha$. Pokud platí nulová hypotéza, pak

$$1 - \alpha = P_{\boldsymbol{\theta}} \left(D_n(\mathbf{X}) \leq \gamma(\boldsymbol{\theta}_0) \right),$$

takže **kritický obor** tohoto testu má tvar:

$$W_\alpha = \{\mathbf{X} \in \mathbb{R}^n : D_n(\mathbf{X}) > \gamma(\boldsymbol{\theta}_0)\}.$$

(C) Hypotéza $H_0 : \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti (tzv. *jednostranné*) alternativě $H_1 : \gamma(\boldsymbol{\theta}) < \gamma(\boldsymbol{\theta}_0)$

V tomto případě využijeme **horní odhad** $H_n(\mathbf{X})$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $1 - \alpha$. Pokud platí nulová hypotéza, pak

$$1 - \alpha = P_{\boldsymbol{\theta}} \left(\gamma(\boldsymbol{\theta}_0) \leq H_n(\mathbf{X}) \right),$$

takže **kritický obor** tohoto testu má tvar:

$$W_\alpha = \{\mathbf{X} \in \mathbb{R}^n : H_n(\mathbf{X}) < \gamma(\boldsymbol{\theta}_0)\}.$$

Předchozí úvahy shrňme do následující tabulky:

H_0	H_1	Hypotézu H_0 zamítáme, pomocí	
		intervalu spolehlivosti	kritické oblasti, tj. pokud $\mathbf{x} \in W_\alpha$, kde $W_\alpha =$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) \neq \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) \notin (d_n(\mathbf{x}), h_n(\mathbf{x}))$	$\{\mathbf{X} \in \mathbb{R}^n : \gamma(\boldsymbol{\theta}_0) \notin (D_n(\mathbf{X}), H_n(\mathbf{X}))\}$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) > \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) < d_n(\mathbf{x})$	$\{\mathbf{X} \in \mathbb{R}^n : D_n(\mathbf{X}) > \gamma(\boldsymbol{\theta}_0)\}$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) < \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) > h_n(\mathbf{x})$	$\{\mathbf{X} \in \mathbb{R}^n : H_n(\mathbf{X}) < \gamma(\boldsymbol{\theta}_0)\}$

7.3. Testy o parametrech normálního rozdělení, testy založené na centrální limitní větě.

Pomocí intervalových (dolních, horních) odhadů, které jsme již dříve odvodili v sekci 5, dostáváme celou řadu kritických oblastí testů o parametrech normálního rozdělení. Poznamenejme, že se shodují s testy podílem věrohodností.

Přehled takto získaných testů pro JEDEN NÁHODNÝ VÝBĚR $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$ podáváme v následující tabulce:

H_0	H_1	Hypotézu H_0 zamítáme, pokud $\mathbf{X}\in W_\alpha$, tj.	Předpoklady
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert\bar X-\mu_0\rvert\sqrt{n}\geq\sigma u_{1-\frac{\alpha}{2}}$	σ^2 známé
$\mu=\mu_0$	$\mu>\mu_0$	$(\bar X-\mu_0)\sqrt{n}\geq\sigma u_{1-\alpha}$	σ^2 známé
$\mu=\mu_0$	$\mu<\mu_0$	$(\bar X-\mu_0)\sqrt{n}\leq-\sigma u_{1-\alpha}$	σ^2 známé
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert\bar X-\mu_0\rvert\sqrt{n}\geq St_{1-\frac{\alpha}{2}}(n-1)$	σ^2 neznámé
$\mu=\mu_0$	$\mu>\mu_0$	$(\bar X-\mu_0)\sqrt{n}\geq St_{1-\alpha}(n-1)$	σ^2 neznámé
$\mu=\mu_0$	$\mu<\mu_0$	$(\bar X-\mu_0)\sqrt{n}\leq-St_{1-\alpha}(n-1)$	σ^2 neznámé
$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2}\notin\left(\chi^2_{\frac{\alpha}{2}}(n-1),\chi^2_{1-\frac{\alpha}{2}}(n-1)\right)$	μ neznámé
$\sigma^2=\sigma_0^2$	$\sigma^2>\sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2}\geq\chi^2_{1-\alpha}(n-1)$	μ neznámé
$\sigma^2=\sigma_0^2$	$\sigma^2<\sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2}\leq\chi^2_{\alpha}(n-1)$	μ neznámé

V případě DVOU NEZÁVISLÝCH VÝBĚRŮ

- první náhodný výběr $\perp\!\!\!\perp\{X_1,\dots,X_{n_1}\}\sim N(\mu_1,\sigma_1^2)$ (s výběrovým průměrem $\bar X$ a výběrový rozptylem S_1^2),
- druhý náhodný výběr $\perp\!\!\!\perp\{Y_1,\dots,Y_{n_2}\}\sim N(\mu_2,\sigma_2^2)$ (s výběrovým průměrem $\bar Y$ a výběrový rozptylem S_2^2),
- a pokud označíme

$$S_{12}^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2},$$

pak následující tabulka se týká testů rovnosti středních hodnot a rozptylů:

H_0	H_1	Hypotézu H_0 zamítáme, pokud $(\mathbf{X}',\mathbf{Y}')'\in W_\alpha$, tj.	Předpoklady
$\mu_1=\mu_2$	$\mu_1\neq\mu_2$	$\lvert\bar X-\bar Y\rvert\geq u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	σ_1^2,σ_2^2 známé
$\mu_1=\mu_2$	$\mu_1\neq\mu_2$	$\lvert\bar X-\bar Y\rvert\geq t_{1-\frac{\alpha}{2}}(n_1+n_2-2)\ S_{12}\sqrt{\frac{n_1+n_2}{n_1n_2}}$	$\sigma_1^2=\sigma_2^2$ neznámé
$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$	$\frac{S_1^2}{S_2^2}\notin\left(F_{\frac{\alpha}{2}}(n_1-1,n_2-1),F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\right)$	μ_1,μ_2 neznámé

Následující tabulka nabízí ASYMPTOTICKÉ TESTY pro náhodné výběry $\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq\mathcal{L}(\mu(\boldsymbol\theta),\sigma^2(\boldsymbol\theta))$ s konečnými druhými momenty (s výběrovým průměrem $\bar X=\frac{1}{n}\sum_{i=1}^n X_i$ a se $S_*^2=S_*^2(\mathbf{X})$, což je (**slabě**) **konzistentní odhad** rozptylu $\sigma^2(\boldsymbol\theta)$):

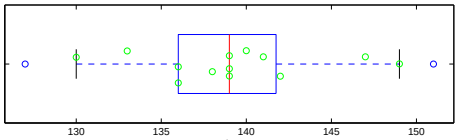
H_0	H_1	Hypotézu H_0 zamítáme, pokud $\mathbf{X}\in W_\alpha$, tj.	Předpoklady
$\mu=\mu_0$	$\mu\neq\mu_0$	$\frac{\lvert\bar X-\mu_0\rvert}{S_*}\sqrt{n}\geq\ u_{1-\frac{\alpha}{2}}$	$0<\sigma^2(\boldsymbol\theta)<\infty$
$\mu=\mu_0$	$\mu\neq\mu_0$	$\frac{\lvert\bar X-\mu_0\rvert}{\sqrt{\bar X}}\sqrt{n}\geq u_{1-\frac{\alpha}{2}}$	$\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq Po(\mu)$
$p=p_0$	$p\neq p_0$	$\frac{\lvert\bar X-p_0\rvert}{\sqrt{p_0(1-p_0)}}\sqrt{n}\geq u_{1-\frac{\alpha}{2}}$	$\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq A(p)$

Příklad 7.2 (Výška desetiletých chlapců)**.** V roce 1961 byla u 15 náhodně vybraných chlapců z populace všech desetiletých chlapců žijících v Československu zjištěna výška

Výšky 15 desetiletých chlapců

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
130	140	136	141	139	133	149	151	139	136	138	142	127	139	147

Je známo, že každá následující generace je v průměru o něco vyšší než generace předcházející. Můžeme se tedy ptát, zda průměr $\bar{x} = 139.133$ zjištěný v náhodném výběru rozsahu



$n = 15$ znamená, že na 5% hladině máme zamítnout nulovou hypotézu $H_0 : \mu = 136.1$ (zjištění z roku 1951) ve prospěch alternativní hypotézy $H_1 : \mu > 136.1$.

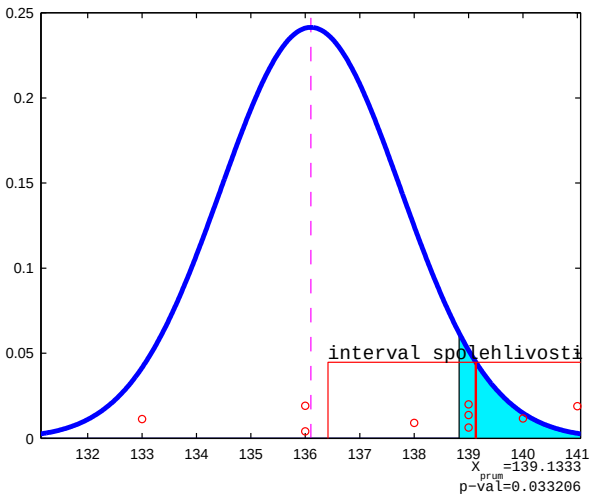
Rozptyl $\sigma^2 = 6.4^2$ cm^2, zjištěný v roce 1951 (kdy se provádělo rozsáhlé šetření), můžeme považovat za známý, neboť variabilita výšek zůstává (na rozdíl od střední výšky) téměř nezměněná.

Řešení. (I) Testování nulové hypotézy pomocí pivotové statistiky $U_{\overline{X}}$ a kritické hodnoty. Protože kritický obor W_0 lze ekvivalentně vyjádřit i takto

$$W_0 = \left\{ \mathbf{x} \in \mathbb{R}^n : \bar{x} - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha} > \mu_0 \right\} = \left\{ \mathbf{x} \in \mathbb{R}^n : u_{\bar{x}} = \tfrac{\bar{x}-\mu_0}{\sigma}\sqrt{n} > u_{1-\alpha} \right\},$$

počítejme $u_{\bar{x}} = \frac{139.133-136.1}{6.4}\sqrt{15} = 1.835$. Protože $u_{\bar{x}} = 1.835$ překračuje kritickou hodnotu $u_{1-\alpha} = u_{0.95} = 1.645$ (získáme pomocí R, a to příkazem „`rnorm(0.95)`“) nulovou hypotézu na 5% hladině **zamítneme** ve prospěch alternativní hypotézy, že se střední výška desetiletých hochů zvětšila.

(II) Testování nulové hypotézy pomocí p-hodnoty



Dosažená hladina odpovídající testové statistice (tj. tzv. p-hodnota, anglicky *P-value*, *significance value*), což je nejmenší hladina testu, při které bychom ještě hypotézu H_0 zamítli, je rovna 0.033 (opět získáme pomocí R příkazem „`1 - pnorm(mean(x),mean=136.1,sd=6.4/sqrt(n))`“), takže například při $\alpha = 2.5\%$ by již dosažený výsledek nebyl statisticky významný.

Protože p-hodnota je menší než zvolená hladina významnosti $\alpha = 0.05$, hypotézu **zamítáme**.

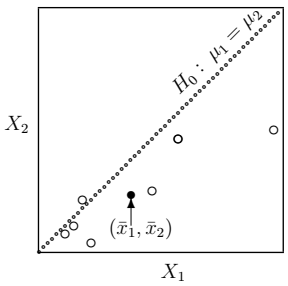
(III) Testování nulové hypotézy pomocí intervalu spolehlivosti $\langle D, +\infty)$

Protože jde o jednostranný test, použijeme **dolní odhad** střední hodnoty μ

$$d = \bar{x} - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha} = 139.133 - \tfrac{6.4}{\sqrt{15}} 1.645 = 136.415$$

Protože interval spolehlivosti $\langle 136.415, +\infty)$ nepokrývá hodnotu 136.1, proto nulovou hypotézu na na hladině významnosti $\alpha = 0.05$ **zamítáme**.

Příklad 7.3. Párový test



Na sedmi rostlinách byl posuzován vliv fungicidního přípravku podle počtu skvrn na listech před a týden po použití přípravku. Otestujte, zdali má přípravek vliv na počet skvrn na listech. Data udávající počet skvrn na listech před a po použití přípravku:

Počet skvrn na listech

před použitím přípravku	X_1	9	17	31	7	8	20	10
po použití přípravku	X_2	10	11	18	6	7	17	5

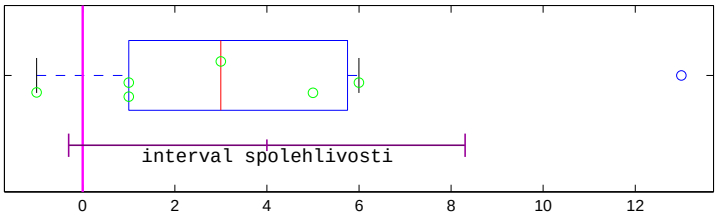
Řešení. Za předpokladu, že náhodný výběr pochází z normálního rozdělení, tj.

$$\perp\!\!\!\perp \left\{ \begin{pmatrix} X_{1,1} \\ X_{2,1} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1,n} \\ X_{2,n} \end{pmatrix} \right\} \sim N_2 \left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \text{ kde } \rho \in (0,1)$$

pak $\begin{matrix} X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2) \end{matrix}$, $Z = X_1 - X_2 \sim N(\mu_z = \mu_1 - \mu_2, \sigma_z^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$

a statistika $T = \frac{\bar{Z}}{S_Z/\sqrt{n}} = \frac{\bar{X}_1 - \bar{X}_2}{S_Z/\sqrt{n}}$ má za platnosti nulové hypotézy $H_0 : \mu_1 - \mu_2 = 0$ Studentovo rozdělení o $n-1$ stupních volnosti.

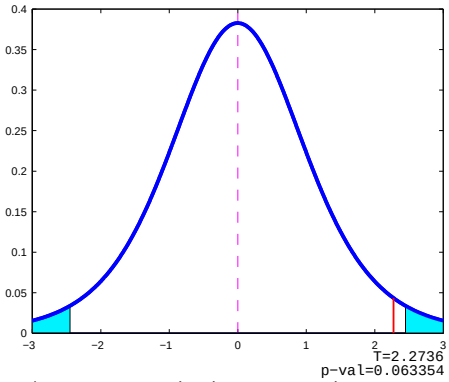
(I) Testování nulové hypotézy pomocí intervalu spolehlivosti



$$[\bar{X}_1 - \bar{X}_2 - t_{1-\alpha/2}(n-1) \cdot S/\sqrt{n};$$
$$\bar{X}_1 - \bar{X}_2 + t_{1-\alpha/2}(n-1) \cdot S/\sqrt{n}] =$$
$$[4 \pm 2.4469 \cdot 4.6547/2.6458] =$$
$$[-0.30492; 8.3049]$$

Protože interval spolehlivosti pokrývá hodnotu $Z = 0$, na dané hladině významnosti **hypotézu nemůžeme zamítnout**.

(II) Testování nulové hypotézy pomocí statistiky T a kritické hodnoty



Vypočítáme-li hodnotu statistiky

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}}$$

a porovnáme s kvantilem Studentova rozdělení, tj.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s/\sqrt{n}} = 2.2736 \not> t_{1-\alpha/2}(n-1) = 2.4469,$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

nezamítáme.

(III) Testování nulové hypotézy pomocí p-hodnoty

Vypočítáme-li p-hodnotu a porovnáme se zvolenou hladinou významnosti $\alpha = 0.05$

$$p = P(|T| > t) = 2(1 - P(|T| \leq t)) = 0.06335 > \alpha$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

nezamítáme.

Shrneme-li předchozí výsledky slovně, pak nulovou hypotézu o tom, že

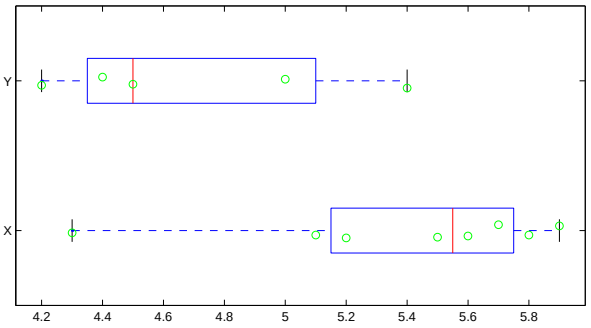
přípravek nemá vliv na počet skvrn

na hladině významnosti $\alpha = 0.05$ **nemůžeme zamítnout** oproti alternativě o jeho vlivu.

Příklad 7.4 (Dva nezávislé náhodné výběry z normálního rozdělení při neznámých ale stejných rozptylech)**.** Bylo vybráno 13 polí stejné kvality. Na 8 z nich se zkoušel nový způsob hnojení, zbývajících 5 bylo ošetřeno běžným způsobem. Výnosy pšenice uvedené v tunách na hektar jsou označeny X_i u nového a Y_i u běžného způsobu hnojení. (převzato z knihy [**2**], str. 82, př. 8.2).

Je třeba zjistit, zda způsob hnojení má vliv na výnos pšenice.

X_i	5.7	5.5	4.3	5.9	5.2	5.6	5.8	5.1
Y_i	5.0	4.5	4.2	5.4	4.4			



Nechť $\perp\!\!\!\perp\{X_1, \ldots, X_{n_1}\} \sim N(\mu_1, \sigma_1^2)$ je náhodný výběr rozsahu n_1 z normálního rozdělení $N(\mu_1, \sigma_1^2)$, $\bar{X}$ je jeho výběrový průměr a S_1^2 jeho výběrový rozptyl.

Dále nechť $\perp\!\!\!\perp\{Y_1, \ldots, Y_{n_2}\} \sim N(\mu_2, \sigma_2^2)$ je náhodný výběr rozsahu n_2 z normálního rozdělení $N(\mu_2, \sigma_2^2)$, $\bar{Y}$ je jeho výběrový průměr a S_2^2 jeho výběrový rozptyl.

Předpokládejme, že oba výběry jsou stochasticky nezávislé, tj. $\mathbf{X} \perp \mathbf{Y}$.

Řešení. Chceme-li testovat hypotézu, že rozdíl středních hodnot je nulový (při neznámém rozptylu $\sigma^2 = \sigma_1^2 = \sigma_2^2$), za pivotovou statistiku zvolíme statistiku

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_{12}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sim t(n_1 + n_2 - 2),$$

kde

$$S_{12}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Chceme-li použít $T_{\bar{X}-\bar{Y}}$, měli bychom být přesvědčeni o tom, že rozptyly obou výběrů se významně neliší. Budeme tedy nejprve testovat hypotézu $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$, že podíl obou rozptylů je roven jedné proti alternativě, že se nerovná $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$. Za pivotovou statistiku zvolíme statistiku

$$F = \frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

(a) Můžeme například vypočítat statistiku F za platnosti nulové hypotézy a porovnat ji s příslušnými oboustrannými kvantily.

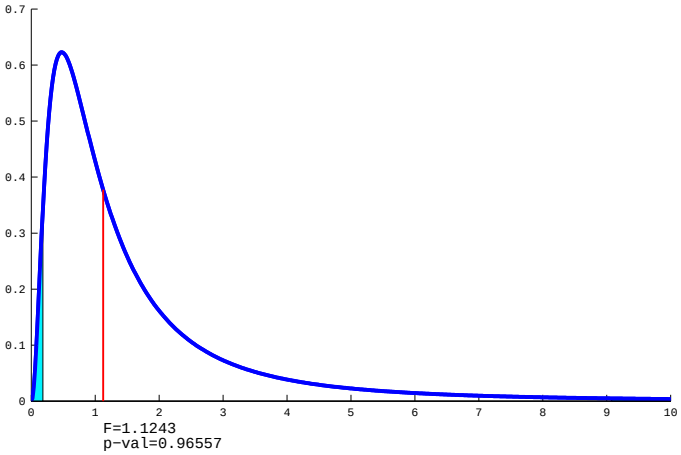
Protože

$f = 1.1243$

$F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = 0.1811$

$F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = 9.0741$

vidíme, že f není ani větší než horní kritický bod, ani menší než dolní kritický bod, takže hypotézu o rovnosti rozptylů proti alternativě nerovnosti **nezamítáme** a můžeme konstatovat, že data nejsou v rozporu s testovanou hypotézou.



(b) Další možností je spočítat dosaženou hladinu významnosti, tj. p-hodnotu (pomocí R: `2*min(1-pf(var(x)/var(y),n1-1,n2-1),pf(var(x)/var(y),n1-1,n2-1))`) a srovnat se zvolenou hladinou testu α:

$$p\text{-value} = 0.9656 \gg 0.05$$

Protože p-hodnota je výrazně větší než zvolená hladina testu, hypotézu o rovnosti rozptylů proti alternativě nerovnosti **nezamítáme**. Můžeme také říci, že data **nejsou v rozporu s testovanou hypotézou**.

(c) A naposledy můžeme ještě zkonstruovat $100(1-\alpha)\%$ interval spolehlivosti pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$

$$\left\langle \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)} \right\rangle.$$

a zjistit, zda pokrývá hodnotu 1. Protože dostáváme interval $\langle 0.1239, 6.2088\rangle$, který pokrývá jedničku, hypotézu nezamítáme.

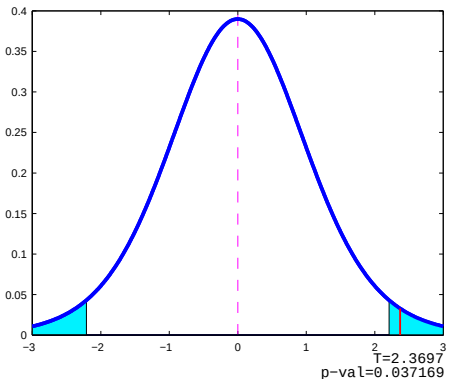
Díky předchozímu zjištění již můžeme bez obav testovat hypotézu $H_0 : \mu_1 - \mu_2 = 0$ proti alternativě $H_1 : \mu_1 - \mu_2 \neq 0$ a provedeme to opět třemi způsoby:

(I) Testování nulové hypotézy pomocí intervalu spolehlivosti

$$\left\langle \bar{X}-\bar{Y}-t_{1-\frac{\alpha}{2}}(\nu)\, S\sqrt{\tfrac{n_1+n_2}{n_1n_2}}; \bar{X}-\bar{Y}+t_{1-\frac{\alpha}{2}}(\nu)\, S\sqrt{\tfrac{n_1+n_2}{n_1n_2}} \right\rangle = \langle 0.6875 \pm 2.201 \cdot 0.5089/1.7541\rangle$$
$$= \langle 0.048958; 1.326\rangle$$

Protože interval spolehlivosti nepokrývá nulu, na dané hladině významnosti **hypotézu zamítáme** ve prospěch alternativy.

(II) Testování nulové hypotézy pomocí statistiky T a kritické hodnoty



Vypočítáme-li hodnotu statistiky

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_{12}}\sqrt{\frac{n_1n_2}{n_1+n_2}}$$

a porovnáme s kvantilem Studentova rozdělení, tj.

$$t_{\bar{x}-\bar{y}} = 2.3697 > t_{1-\alpha/2}(11) = 2.201,$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

zamítáme.

(III) Testování nulové hypotézy pomocí p-hodnoty

Vypočítáme-li p-hodnotu a porovnáme se zvolenou hladinou významnosti $\alpha = 0.05$

$$p = P(|T_{\bar{X}-\bar{Y}}| > t) = 2(1 - P(|T_{\bar{X}-\bar{Y}}| \leq t)) = 0.037169 < \alpha$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

zamítáme.

Shrneme-li předchozí výsledky slovně, pak nulovou hypotézu o tom, že

hnojení je stejně účinné

na hladině významnosti $\alpha = 0.05$ **zamítáme** ve prospěch alternativy, že má rozdílné účinky.

Úlohy k procvičení

Cvičení 7.1. Nechť T_1 a T_2 jsou nezávislé nestranné odhady parametru θ. Předpokládejme, že rozptyl statistiky T_1 je dvakrát větší než rozptyl statistiky T_2. Určete konstanty k_1 a k_2 tak, aby odhad $T = k_1T_1 + k_2T_2$ byl nestranným odhadem parametru θ s nejmenším rozptylem.

$[k_1 = 1/3,\ k_2 = 2/3]$

Cvičení 7.2. Uvažujme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ z rozdělení s konečným rozptylem $\sigma^2 > 0$. Určete konstantu c tak, aby statistika $T = c\sum_{i=2}^{n}(X_i - X_{i-1})^2$ byla nestranným odhadem rozptylu.

$[c = \frac{1}{2(n-1)}]$

Cvičení 7.3. Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rovnoměrného rozdělení $Ro(0, \theta)$, $\theta > 0$. Uvažujme statistiky $T_1 = \max(X_1, \ldots, X_n)$ a $T_2 = \frac{2}{n}\sum_{i=1}^{n} X_i$. Rozhodněte, zda jsou statistiky T_1 a T_2 nestrannými odhady parametru θ, případně je upravte tak, aby byly nestrannými odhady.

$[T_1$ není, $\widetilde{T}_1 = \frac{n+1}{n}\max(X_1, \ldots, X_n)$, T_2 je$]$

Cvičení 7.4. Porovnejte rozptyly obou statistik (upravených tak, aby byly nestranné) z Cvičení 7.3.

$[D\widetilde{T}_1 = \frac{\theta^2}{n(n+2)},\ DT_2 = \frac{\theta^2}{3n}$, $\widetilde{T}_1$ je lepší$]$

Cvičení 7.5. Jsou statistiky T_1 a T_2 konzistentními odhady parametru θ?

[ano]

Cvičení 7.6. Rychlost letadla byla určována v 5 zkouškách a z jejich výsledků byl vypočten odhad $\bar{x} = 870,3\ m/s$. Najděte 95% interval spolehlivosti pro μ, je-li známo, že rozptýlení rychlosti letadla se řídí normálním rozdělením se směrodatnou odchylkou $\sigma = 2,1m/s$.

$[(868,46; 872,14)]$

Cvičení 7.7. Deset balíčků mouky pocházejících z balícího stroje mělo hmotnosti v gramech: 987, 1 001, 993, 994, 993, 1 005, 1 007, 999, 995, 1 002. Sestrojte 95% interval spolehlivosti pro střední hodnotu a rozptyl hmotnosti (předpokládejte normální rozdělení).

$[\mu \in (993,1; 1002,1),\ \sigma^2 \in (18,4; 129,8)]$

Cvičení 7.8. Při zjišťování přesnosti nově zaváděné metody pro stanovení obsahu manganu v oceli bylo rozhodnuto provést 4 nezávislá měření. Stanovte dolní odhad pro σ s rizikem $0,05$, když výsledky měření byly: $0,31\%;\ 0,30\%;\ 0,29\%; 0,32\%$. Údaje o obsahu manganu považujeme za náhodný výběr z normálního rozdělení.

$[0,00799]$

Cvičení 7.9. Ze základního souboru byl proveden náhodný výběr s naměřenými intervalovými hodnotami a jejich četnostmi sledovaného znaku

x_i	$(15, 17\rangle$	$(17, 19\rangle$	$(19, 21\rangle$	$(21, 23\rangle$	$(23, 25\rangle$	$(25, 27\rangle$
n_i	10	30	50	70	60	30

Určete

a) interval ve kterém se nachází střední hodnota μ s pravděpodobností 0,95
b) interval ve kterém se nachází rozptyl σ^2 s pravděpodobností 0,95.

[a) $(21,5094; 22,1706)$, b) $(5,952; 8,464)$]

Cvičení 7.10. Z 42 náhodně vybraných účastníků sportovního odpoledne bylo 16 dívek a 26 chlapců. Určete interval spolehlivosti pro podíl dívek mezi účastníky.

[$(0,2331; 0,5269)$]

Cvičení 7.11. Mezi 160 pracovníky (náhodně vybranými z 8 000 pracujících v závodě) 48 cestuje do práce vlakem. Napište bodový odhad a 95% interval spolehlivosti pro podíl a počet zaměstnanců dopravujících se vlakem.

[podíl: $0,3$; $(0,229; 0,371)$, počet: 2 400; $(1\,832; 2\,968)$]

Cvičení 7.12. Spotřeba téhož auta byla testována u 11 řidičů s výsledky 8,8; 8,9; 9,0; 8,7; 9,3; 9,0; 8,7; 8,8; 9,4; 8,6; 8,9 ($l/100\,km$). Můžeme na hladině významnosti 0,05 zamítnout hypotézu, že je pravdivá výrobcem udávaná spotřeba 8,8 $l/100\,km$? Můžeme na stejné hladině významnosti popřít tvrzení, že rozptyl spotřeby je 0,1?

[ne, ne]

Cvičení 7.13. Na hladině významnosti $\alpha = 0,05$ testujte hypotézu $H_0 : \sigma_0 = 300$ o směrodatné odchylce normálně rozdělené náhodné veličiny, jestliže je zaznamenáno $n = 25$, $\overline{X} = 3118$, $s = 357$.

[nezamítáme]

Cvičení 7.14. Denní přírůstky váhy selat (v dkg) byly při krmení směsí A : 62, 54, 55, 60, 53, 58, u směsi B : 52, 56, 50, 49, 51. Je mezi nimi statisticky významný rozdíl?

[ano]

Cvičení 7.15. U 6 aut bylo zjištěno ojetí předních pneumatik (v mm)

L	$1,8$	$1,0$	$2,2$	$0,9$	$1,5$	$1,6$
P	$1,5$	$1,1$	$2,0$	$1,1$	$1,4$	$1,4$

Ojíždějí se levá a pravá pneumatika stejně?

[ano]

Cvičení 7.16. Pro bavlněnou přízi je předepsána horní mez variability pevnosti vlákna: rozptyl pevnosti (která má normální rozdělení) nemá překročit $\sigma_0^2 = 0,36$. Při zkoušce 16 vzorků byly zjištěny výsledky 2,22, 3,54, 2,37, 1,66, 4,74, 4,82, 3,21, 5,44, 3,23, 4,79, 4,85, 4,05, 3,48, 3,89, 4,90, 5,37. Je důvod k podezření na vyšší nestejnoměrnost než je stanoveno?

[ano]

Cvičení 7.17. Bylo provedeno měření obsahu SiO_2 ve strusce dvěma metodami

analyticky	20,1	19,6	20,0	19,9	20,1	
fotokolorometricky	20,9	20,1	20,6	20,5	20,7	20,5

Je mezi rozptyly výsledků jednotlivých metod podstatný rozdíl?

[není]

Cvičení 7.18. Na základě testu máme na 5% hladině významnosti rozhodnout, zda produkce vajec plemene kornyšek černých je nižší než plemene leghornek bílých. Náhodně jsme vybrali 50 kornyšek a 40 leghornek, u nichž byla zjištěna roční produkce vajec. Byl vypočten roční průměr produkce na slepici – kornyška 275, leghornka 280. Z dřívějška jsou známy rozptyly $\sigma_{kor}^2 = 48$, $\sigma_{leg}^2 = 41$.

[H_0 zamítáme, kornyšky mají horší produkci vajec než leghornky]

KAPITOLA 3

Základy regresní a korelační analýzy

Základní informace

(1) V následující kapitole se budeme zabývat statistickou analýzou vzájemných vztahů náhodných jevů. Popíšeme volbu optimální predikce těchto vztahů a uvedeme různé míry závislosti náhodných veličin.
(2) Předpokládá se znalost základních pojmů z teorie pravděpodobnosti a matematické statistiky – náhodná veličina, náhodný vektor, jejich číselné charakteristiky a jejich výběrové odhady

Výstupy z výukové jednotky

Studenti

- umí vysvětlit pojem regrese a korelace
- pochopí optimální volbu predikční funkce
- vypočítají a interpretují index determinace
- vypočítají a interpretují koeficient mnohonásobné korelace
- vypočítají a interpretují parciální korelační koeficient

1. Motivace

V předcházejících kapitolách jsme zkoumali jednotlivé jevy (statistické znaky) izolovaně; zabývali jsme se tzv. jednorozměrnými soubory, tj. soubory popisujícími pouze jeden statistický znak a nezajímaly nás jeho vazby a vztahy k jiným jevům. V reálném světě (v přírodě, společnosti, ekonomice,...) se ovšem jevy nacházejí ve více nebo méně složitých vzájemných vztazích – navzájem na sobě závisí a podmiňují se. Proto se statistická analýza nemůže omezit pouze na zkoumání izolovaných jevů, ale musí se také zabývat analýzou jejich vzájemných vztahů. Tato analýza se dá obecně rozdělit na dvě části: regresní a korelační. Popišme si podrobněji podstatu obou typů analýz.

1.1. Úloha regresní analýzy. Hlavní úlohou regresní analýzy je provést **predikci** nějaké závisle proměnné náhodné veličiny Y na základě informace, kterou poskytují měření nějakých jiných náhodných veličin, řekněme $X_1, \ldots, X_k$. Veličinám $X_1, \ldots, X_k$ se potom říká **nezávisle proměnné** nebo též **doprovodné proměnné**, nebo také **kovariáty**. Měření nezávislých proměnných jsou pro experimentátora snáze dostupné než měření závisle proměnné Y.

Predikce spočívá v nalezení nějaké funkce $g(X_1, \ldots, X_k)$, která vhodně aproximuje závisle proměnnou Y. Kvalita predikce se obvykle posuzuje pomocí tzv. **střední kvadratické chyby predikce** $E[Y - g(X_1, \ldots, X_k)]^2$. Za optimální se považuje volba takové predikční funkce g, která uvedenou střední kvadratickou chybu **minimalizuje**.

1.2. Úloha korelační analýzy. Vedle průběhu sledované závislosti Y na $X_1, \ldots, X_k$ dané funkcí g je také třeba se zaměřit na měření **těsnosti** tohoto vztahu, tedy je nutné zavést nějaké míry velikosti statistické vazby (závislosti) závisle proměnné Y na nezávisle proměnných $X_1, \ldots, X_k$ s ohledem na vybranou funkci g a případně také s ohledem na závislosti mezi náhodnými veličinami $X_1, \ldots, X_k$. Tato problematika je hlavní úlohou korelační analýzy. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0

do 1 (resp. od -1 do 1). Čím je takový koeficient bližší 1 (resp. -1), tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

2. Optimální volba predikční funkce g

Pomocí regresní a korelační analýzy lze provádět predikce nejrůznějšího typu. Nejzávažnější otázkou je, jak volit vhodnou predikční funkci g.

VĚTA 2.1. *Nechť $Y, X_1, \ldots, X_k$ jsou náhodné veličiny. Označme $\mathbf{X} = (X_1, \ldots, X_k)'$ a nechť $EY^2 < \infty$. Pak pro každou měřitelnou funkci*

$$g : \mathbb{R}^k \to \mathbb{R}$$

platí

$$E(Y - g(\mathbf{X}))^2 \geq E[Y - E(Y|\mathbf{X})]^2$$

a rovnost v uvedené nerovnosti nastává právě když

$$P(g(\mathbf{X}) = E(Y|\mathbf{X})) = 1.$$

Poznámka 2.2 (Podmíněná střední hodnota)**.** V předchozí větě se vyskytl nový výraz $E(Y|\mathbf{X})$ pro tzv. **podmíněnou střední hodnotu**. Nebudeme uvádět přesnou definici, pro jednoduchost vysvětlíme tento pojem pro spojité náhodné veličiny X a Y:
Nechť spojitý náhodný vektor $\mathbf{Z} = (Y, X)'$ má sdruženou hustotu $f(y, x)$ a dále nechť náhodné veličiny X a Y mají marginální hustoty $f_X(x)$, resp. $f_Y(y)$. Označme $M_X = \{x \in \mathbb{R} : f_X(x) > 0\}$, $M_Y = \{y \in \mathbb{R} : f_Y(y) > 0\}$.
Pak **podmíněná distribuční funkce** je v tomto případě definována vztahem

$$F(y|x) = \begin{cases} \int_{-\infty}^{y} \frac{f(t,x)}{f_X(x)} dt & \text{pro } x \in M_X, \\ 0 & \text{pro } x \in \mathbb{R} \setminus M_X \end{cases}$$

a **podmíněná hustota** je rovna

$$f(y|x) = \begin{cases} \frac{f(y,x)}{f_X(x)} & \text{pro } x \in M_X, \\ 0 & \text{pro } x \in \mathbb{R} \setminus M_X. \end{cases}$$

Položme

$$h(x) = E(Y|X = x) = \int_{\mathbb{R}} y dF(y|x) = \int_{\mathbb{R}} y \frac{f(y,x)}{f_X(x)} dy, \quad \text{pro } \forall x \in M_X.$$

Pak náhodnou veličinu

$$E(Y|X) = h(X)$$

nazveme **podmíněnou střední hodnotou** náhodné veličiny Y při daném X. Dá se ukázat, že jsou splněny např. tyto vlastnosti:

- Nechť Y_1, Y_2, X jsou náhodné veličiny a a_0, a_1, a_2 jsou reálné konstanty, pak pokud střední hodnoty EY_1, EY_2 existují, platí
$$E(a_0 + a_1 Y_1 + a_2 Y_2|X) = a_0 + a_1 E(Y_1|X) + a_2 E(Y_2|X). \tag{12}$$
- Nechť X, Y jsou náhodné veličiny a střední hodnota EY existuje, pak
$$E\left[E(Y|X)\right] = EY. \tag{13}$$

Definujeme také **podmíněný rozptyl** náhodné veličiny Y při daném X vztahem

$$D(Y|X) = E\left\{[Y - E(Y|X)]^2 |X\right\}.$$

Platí

$$DY = E\left[D(Y|X)\right] + D\left[E(Y|X)\right]. \tag{14}$$

Poznámka 2.3 (Korelační koeficient)**.** Připomeňme ještě tzv. **Pearsonův**[1] **koeficient korelace** náhodných veličin X, Y (které jsou aspoň intervalového charakteru). Ten je definován vztahem

$$R(X,Y) = \begin{cases} \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} & \text{pro } \sqrt{D(X)},\ \sqrt{D(Y)} > 0, \\ 0 & \text{jinak,} \end{cases}$$

kde $C(X,Y) = E[(X - EX)(Y - EY)]$ je **kovariance** náhodných veličin X a Y.

Připomeneme jeho vlastnosti:

- $R(X,X) = 1$
- $R(X,Y) = R(Y,X)$
- $R(a + bX, c + dY) = \operatorname{sgn}(bd) R(X,Y)$
- $-1 \leq R(X,Y) \leq 1$ a rovnosti je dosaženo tehdy a jen tehdy, když existují reálné konstanty a, b, kde $b \neq 0$ tak, že $P(Y = a + bX) = 1$, přičemž $R(X,Y) = 1$ pro $b > 0$ a $R(X,Y) = -1$ pro $b < 0$.

Z těchto vlastností plyne, že $R(X,Y)$ je vhodnou mírou těsnosti **lineárního** vztahu náhodných veličin X, Y.

Věta 2.4. *Mějme náhodnou veličinu Y s konečným a nenulovým rozptylem a náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)'$. Potom pro libovolnou měřitelnou funkci*

$$g : \mathbb{R}^k \to \mathbb{R}$$

takovou, že existuje korelační koeficient $R(Y, g(\mathbf{X}))$ platí

$$|R(Y, g(\mathbf{X}))| \leq R(Y, E(Y|\mathbf{X})) = \sqrt{\frac{D[E(Y|\mathbf{X})]}{DY}}$$

a rovnost nastává v případě, že $D[E(Y|\mathbf{X})] \neq 0$ právě když $g(\mathbf{X})$ je lineární funkcí $E(Y|\mathbf{X})$ skoro všude vzhledem k P. V případě, že $D[E(Y|\mathbf{X})] = 0$ nastává rovnost při libovolné volbě funkce g.

Výsledky uvedené v předchozích dvou větách ukazují velký význam podmíněné střední hodnoty $E(Y|\mathbf{X})$ v **regresní** a **korelační analýze**.

(1) Z první věty plyne, že nejlepší predikci náhodné veličiny Y pomocí náhodných veličin $X_1, \ldots, X_k$, která minimalizuje střední kvadratickou chybu $E(Y - g(\mathbf{X}))^2$, dostaneme, když položíme

$$g(\mathbf{X}) = E(Y|\mathbf{X}).$$

V této souvislosti potom nejlepší prediktor $g(\mathbf{X}) = E(Y|\mathbf{X})$ nazýváme **regresní funkcí** náhodné veličiny Y na náhodných veličinách $X_1, \ldots, X_k$.

(2) Z druhé věty plyne, že regresní funkce $E(Y|\mathbf{X})$ je prediktor, který má ze všech možných prediktorů $g(\mathbf{X})$ největší korelační koeficient s predikovanou náhodnou veličinou Y. To znamená, že regresní funkce $E(Y|\mathbf{X})$ je optimálním prediktorem v tom smyslu, že má maximální statistickou vazbu (měřenou korelačním koeficientem) s predikovanou náhodnou veličinou Y.

[1] Karl Pearson (1857 – 1936). Britský statistik a matematik. Studoval na Cambridge a poté působil na univerzitě v Londýně. Vychoval řadu vynikajících statistiků.

DEFINICE 2.5. Mějme náhodnou veličinu Y s konečným a nenulovým rozptylem a náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)'$. Potom číslo

$$\eta^2_{Y|\mathbf{X}} = \frac{D[E(Y|\mathbf{X})]}{DY}$$

nazýváme **korelačním poměrem** náhodné veličiny Y na náhodném vektoru $\mathbf{X} = (X_1, \ldots, X_k)'$, nebo též **korelačním poměrem** náhodné veličiny Y na náhodných veličinách $X_1, \ldots, X_k$ a pak jej též značíme $\eta^2_{Y|X_1,\ldots,X_k}$.

Poznámka 2.6. Nyní shrneme předchozí výsledky do několika důležitých poznámek:

(1) Z předchozích vět plyne, že

$$\eta^2_{Y|\mathbf{X}} = [R(Y, E(Y|\mathbf{X}))]^2$$

a tedy pro korelační poměr platí nerovnost

$$0 \leq \eta^2_{Y|\mathbf{X}} \leq 1.$$

(2) Po vydělení rovnosti (14) rozptylem DY a jednoduché úpravě dostaneme

$$1 = \frac{E(Y - E(Y|\mathbf{X}))^2}{DY} + \eta^2_{Y|\mathbf{X}}.$$

Označme symbolem $\sigma^2_{Y|\mathbf{X}}$ **střední kvadratickou chybu** predikce, když prediktorem je regresní funkce $E(Y|\mathbf{X})$, tj.

$$\sigma^2_{Y|\mathbf{X}} = E(Y - E(Y|\mathbf{X}))^2,$$

pak díky předchozímu máme

$$\eta^2_{Y|\mathbf{X}} = 1 - \frac{\sigma^2_{Y|\mathbf{X}}}{DY}.$$

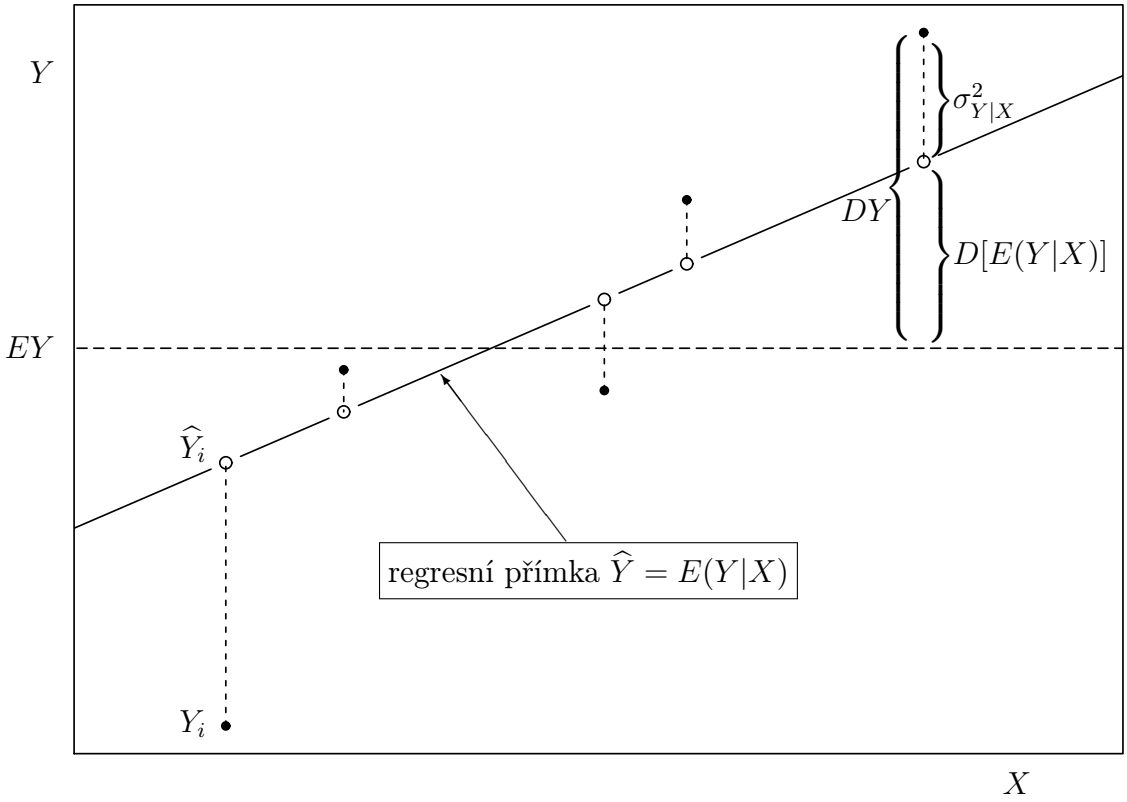
Z tohoto vztahu plyne velice názorná **interpretace** korelačním poměru $\eta^2_{Y|\mathbf{X}}$.

(a) Je-li střední kvadratická chyba predikce $\sigma^2_{Y|\mathbf{X}} = 0$, tedy v případě **ideální predikce**, je korelační poměr $\eta^2_{Y|\mathbf{X}} = 1$.

(b) V druhém krajním případě, když střední kvadratická chyba predikce je rovna DY, tj. $\sigma^2_{Y|\mathbf{X}} = DY$, pak je $\eta^2_{Y|\mathbf{X}} = 0$ a využití informace, kterou o náhodné veličině Y poskytuje náhodný vektor $\mathbf{X}$, nepřináší žádné zmenšení chyby predikce.

Tedy korelační poměr $\eta^2_{Y|\mathbf{X}}$ poskytuje **míru přesnosti predikce** a je velice užitečný při srovnávání různých vektorů doprovodných proměnných.

Poznámka 2.7 (polopatě)**.** Vysvětleme si předchozí pojmy pomocí následujícího obrázku.



Na obrázku je symbolicky znázorněn případ, kdy se zkoumá závislost mezi náhodnými veličinami X a Y. I když jsou vykresleny již konkrétní realizace náhodných veličin (plné kroužky), značení je provedeno velkými písmeny, aby bylo lépe rozumět předchozím vztahům. Přímka představuje predikci v tomto modelu a prázdné kroužky příslušné predikované hodnoty. Symbol DY označuje **celkovou variabilitu** náhodné veličiny Y, tj. odchylku od své střední hodnoty EY (umocněnou na druhou). Symbol $D[E(Y|X)]$ představuje **variabilitu vysvětlenou modelem**, tj. odchylku predikovaných hodnot od střední hodnoty EY. Podíl těchto odchylek (umocněných na druhou) definuje korelační poměr $\eta^2_{Y|\mathbf{X}}$. Symbol $\sigma^2_{Y|X}$ odpovídá tzv. **reziduální variabilitě**, tj. odchylce náhodné veličiny Y od své predikce.

NÁVOD 2.8 (praktický výpočet). Při praktických výpočtech se příslušné rozptyly odhadují výběrovými rozptyly. Odhadnutý korelační poměr $\eta^2_{Y|\mathbf{X}}$ se pak nazývá **index determinace**.

Nechť tedy máme realizace $y_1, \ldots, y_n$ a jejich predikované hodnoty $\hat{y}_1, \ldots, \hat{y}_n$. Koeficient determinace má tvar

$$ID = \frac{s^2_{\widehat{Y}}}{s^2_Y} = 1 - \frac{s^2_{Y\widehat{Y}}}{s^2_Y},$$

kde

$$s^2_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \quad s^2_{Y\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad s^2_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Příklad 2.9. Při laboratorním pokusu bylo získáno následujících 8 výsledků měření

	1	2	3	4	5	6	7	8
x_i	2,2840	2,8170	2,8367	3,5288	4,1031	4,4262	4,5211	4,9446
y_i	4,3046	6,3235	3,7082	7,6835	7,0239	8,7973	10,2961	8,4979

Zvolený model nám predikoval tyto hodnoty

$$\hat{\mathbf{y}} = (4,2614; 5,3352; 5,3750; 6,7694; 7,9264; 8,5774; 8,7685; 9,6217).$$

Určete index determinace a interpretujte ho.

Řešení. Ukážeme oba způsoby výpočtu. Vypočteme nejprve příslušné výběrové rozptyly: $\bar{y} = 7,079$, $s^2_{\widehat{Y}} = \frac{1}{8}\sum_{i=1}^{8}(\hat{y}_i - 7,079)^2 = 3,283$, $s^2_{Y\widehat{Y}} = \frac{1}{8}\sum_{i=1}^{8}(y_i - \hat{y}_i)^2 = 1,131$, $s^2_Y = \frac{1}{8}\sum_{i=1}^{8}(y_i - 7,079)^2 = 4,414$.
Podle definice je

$$ID = \frac{s^2_{\widehat{Y}}}{s^2_Y} = \frac{3,283}{4,414} = 0,7438$$

nebo

$$ID = 1 - \frac{s^2_{Y\widehat{Y}}}{s^2_Y} = 1 - \frac{1,131}{4,414} = 0,7438.$$

Výsledek lze interpretovat tak, že 74,38% celkové variability je vysvětleno zvoleným modelem.

3. Analýza závislosti

Na základě předchozích výsledků můžeme tedy říci, že úloha predikce je teoreticky vyřešena tak, že za nejlepší prediktor stačí zvolit regresní funkci $E(Y|\mathbf{X})$.

Ovšem výpočet podmíněné střední hodnoty $E(Y|\mathbf{X})$ vyžaduje znalost sdruženého rozdělení náhodného vektoru $\mathbf{Z} = (Y, X_1, \ldots, X_k)'$, což činí hlavní potíž při praktickém využití předchozích výsledků.

V praktických situacích nebývá sdružené rozdělení vektoru $\mathbf{Z} = (Y, X_1, \ldots, X_k)'$ známé, proto se, pokud to praktická situace dovolí, uvažují pouze **lineární modely** typu

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \beta_0 + \boldsymbol{\beta}'\mathbf{X},$$

jestliže označíme $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$. Úloha predikce se pak redukuje na nalezení neznámých koeficientů $\beta_0, \ldots, \beta_k$, které minimalizují střední kvadratickou chybu této predikce, tj.

$$(\beta_0, \ldots, \beta_k)' = \underset{(c_0,\ldots,c_k)'\in\mathbb{R}^{k+1}}{\arg\min} E(Y - c_0 - c_1X_1 - \cdots - c_kX_k)^2$$

Označme $\widehat{Y} = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$ nejlepší lineární predikci náhodné veličiny Y. Střední kvadratickou chybu nejlepší **lineární** predikce označíme tentokrát

$$\sigma^2_{Y\cdot\mathbf{X}} = E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2$$

(prosím neplést s označením z minulé kapitoly $\sigma^2_{Y|\mathbf{X}}$ pro střední kvadratickou chybu v případě, že prediktorem je $E(Y|\mathbf{X})$). Poznamenejme, že se někdy střední kvadratická chyba predikce $\sigma^2_{Y\cdot\mathbf{X}}$ také nazývá **reziduální rozptyl**, neboť výraz $Y - \widehat{Y}$ se také nazývá **reziduum**.

3.1. Koeficient mnohonásobné korelace. Dále se budeme zabývat statistickými vazbami mezi predikovanou náhodnou veličinou Y a její nejlepší lineární predikcí $\widehat{Y}$.

DEFINICE 3.1. Pearsonův korelační koeficient $R(Y,\widehat{Y})$ označíme $\rho_{Y\cdot\mathbf{X}}$ a budeme jej nazývat **koeficientem mnohonásobné korelace** náhodné veličiny Y na náhodném vektoru $\mathbf{X}=(X_1,\ldots,X_k)'$ (nebo též na náhodných veličinách $X_1,\ldots,X_k$ a pak budeme podrobněji psát $\rho_{Y\cdot(X_1,\ldots,X_k)}$).

DEFINICE 3.2 (Korelační matice). Nechť $\mathbf{X}=(X_1,\ldots,X_n)'$ a $\mathbf{Y}=(Y_1,\ldots,Y_m)'$ jsou náhodné vektory. Potom matici

$$R(\mathbf{X},\mathbf{Y})=\begin{pmatrix} R(X_1,Y_1) & \cdots & R(X_1,Y_m)\\ \vdots & \ddots & \vdots\\ R(X_n,Y_1) & \cdots & R(X_n,Y_m)\end{pmatrix}=(R(X_i,Y_j))_{\substack{i=1,\ldots,n\\ j=1,\ldots,m}}$$

nazýváme **korelační maticí náhodných vektorů $\mathbf{X}$ a $\mathbf{Y}$**.
Dále matici $R(\mathbf{X},\mathbf{X})$ budeme značit $R(\mathbf{X})$ a budeme ji nazývat **korelační maticí náhodného vektoru $\mathbf{X}$**.

VĚTA 3.3. *Koeficient mnohonásobné korelace $\rho_{Y\cdot\mathbf{X}}$ má následující vlastnosti*
(1) Koeficient mnohonásobné korelace $\rho_{Y\cdot\mathbf{X}}$ je vždy nezáporný.
(2) Pomocí regresních koeficientů $\beta_0,\beta_1,\ldots,\beta_k$ jej lze vyjádřit ve tvaru

$$\rho^2_{Y\cdot\mathbf{X}}=\frac{\boldsymbol{\beta}'D\mathbf{X}\boldsymbol{\beta}}{DY}.$$

(3) Pomocí korelačních matic jej lze vyjádřit ve tvaru

$$\rho^2_{Y\cdot\mathbf{X}}=R(Y,\mathbf{X})(R(\mathbf{X}))^{-1}R(\mathbf{X},Y)$$

(4) Pomocí reziduálního rozptylu lineární predikce jej lze vyjádřit ve tvaru

$$\rho^2_{Y\cdot\mathbf{X}}=1-\frac{\sigma^2_{Y\cdot\mathbf{X}}}{DY}$$

Poznámka 3.4 (polopatě)**.** Z předcházející věty vyplývá:

(1) Vzorec $\boxed{\rho^2_{Y\cdot\mathbf{X}}=\frac{\boldsymbol{\beta}'D\mathbf{X}\boldsymbol{\beta}}{DY}}$ je vhodný pro výpočet koeficientu mnohonásobné korelace v případě, že je k dispozici vektor regresních koeficientů $(\beta_0,\beta_1,\ldots,\beta_k)'$.
(2) Vzorec $\boxed{\rho^2_{Y\cdot\mathbf{X}}=R(Y,\mathbf{X})(R(X))^{-1}R(\mathbf{X},Y)}$ se využívá v případě, že jsou k dispozici korelační koeficienty mezi náhodnými veličinami $Y,X_1,\ldots,X_k$.
(3) Identity $\boxed{\rho^2_{Y\cdot\mathbf{X}}=1-\frac{\sigma^2_{Y\cdot\mathbf{X}}}{DY}}$ a $\boxed{\eta^2_{Y|\mathbf{X}}=1-\frac{\sigma^2_{Y|\mathbf{X}}}{DY}}$ ukazují, že korelační poměr $\eta^2_{Y|\mathbf{X}}$ je roven kvadrátu koeficientu mnohonásobné korelace $\rho^2_{Y\cdot\mathbf{X}}$ v případě, že teoretická regresní funkce $g(X)=E(Y|\mathbf{X})$ je lineární funkcí proměnných $X_1,\ldots,X_k$. Dále je z tohoto vzorce patrné, že pokud se omezíme na lineární predikce, je interpretace koeficientu mnohonásobné korelace stejná jako je interpretace korelačního poměru v obecném případě.

(4) Podle uváděných vzorců lze koeficient mnohonásobné korelace $\rho_{Y\cdot\mathbf{X}}$ počítat i v případě, kdy podmíněná střední hodnota $E(Y|\mathbf{X})$ není lineární. V tomto případě potom

díky vztahu (dokázaném ve větě 2.1)

$$\underbrace{E(Y-\beta_0-\boldsymbol{\beta}'\mathbf{X})^2}_{=\sigma^2_{Y\cdot\mathbf{X}}} \geq \underbrace{E[Y-E(Y|\mathbf{X})]^2}_{=\sigma^2_{Y|\mathbf{X}}}$$

snadno vidíme, že

$$0 \leq \rho^2_{Y\cdot\mathbf{X}} \leq \eta^2_{Y|\mathbf{X}} \leq 1$$

Dá se ukázat, že ve třídě **linerárních** predikčních funkcí má koeficient mnohonásobné korelace analogické vlastnosti jako korelační poměr, tedy že platí analogie Věty 2.4

VĚTA 3.5. *Pro libovolný nenulový vektor* $\mathbf{c}=(c_1,\ldots,c_k)'\in\mathbb{R}^k$ *a* $c_0\in\mathbb{R}$ *platí*

$$\rho^2_{Y\cdot\mathbf{X}} \geq R^2(Y, c_0+\mathbf{c}'\mathbf{X}),$$

tj. koeficient mnohonásobné korelace je maximální korelační koeficient mezi náhodnou veličinou Y *a libovolnou lineární funkcí* $c_0+\mathbf{c}'\mathbf{X}$ *náhodného vektoru* $\mathbf{X}$.

DŮSLEDEK 3.6. *Pro libovolné* $j=1,\ldots,k$ *platí*

$$\rho^2_{Y\cdot\mathbf{X}} \geq R^2(Y, X_j),$$

tj. absolutní hodnota libovolného korelačního koeficientu mezi náhodnou veličinou Y *a libovolnou z náhodných veličin* $X_1,\ldots,X_k$ *je nejvýše rovna koeficientu mnohonásobné korelace mezi náhodnou veličinou* Y *a náhodným vektorem* $\mathbf{X}=(X_1,\ldots,X_k)'$.

DEFINICE 3.7. Mějme náhodný výběr rozsahu n s vektory $\mathbf{X}_1=\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Z}_1\end{pmatrix},\ldots,\mathbf{X}_n=\begin{pmatrix}\mathbf{Y}_n\\ \mathbf{Z}_n\end{pmatrix}$, kde pro $i=1,\ldots,n$ jsou náhodné vektory $\mathbf{Y}_i$ typu $p\times 1$ a $\mathbf{Z}_i$ typu $q\times 1$, přičemž $p+q=k$. Definujme **výběrové kovarianční matice**

$$\mathbf{S}_{YZ}=\tfrac{1}{n-1}\sum_{i=1}^{n}(\mathbf{Y}_i-\overline{\mathbf{Y}})(\mathbf{Z}_i-\overline{\mathbf{Z}})'=(S_{ij}) \qquad (\text{typu } p\times q),$$

kde

$$\overline{\mathbf{Y}}=\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i=\begin{pmatrix}\overline{Y}_1\\ \vdots\\ \overline{Y}_p\end{pmatrix} \qquad \text{a} \qquad \overline{\mathbf{Z}}=\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i=\begin{pmatrix}\overline{Z}_1\\ \vdots\\ \overline{Z}_q\end{pmatrix},$$

a **výběrovou korelační matici**

$$\mathbf{R}_{ZY}=(r_{ij})=\left(\frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}\right)$$

Nyní definujme výběrový protějšek ke koeficientu mnohonásobné korelace.

DEFINICE 3.8. Mějme náhodné vektory $\begin{pmatrix}Y_1\\ \mathbf{X}_1\end{pmatrix},\cdots,\begin{pmatrix}Y_n\\ \mathbf{X}_n\end{pmatrix}$, kde Y_i jsou náhodné veličiny a $\mathbf{X}_i \quad (i=1,\cdots,n)$ jsou náhodné vektory typu $p\times 1$.
Jestliže matice $\mathbf{R}_{XX}$ je regulární, pak **výběrový koeficient mnohonásobné korelace** je definován vztahem:

$$r^2_{Y\cdot\mathbf{X}}=\mathbf{R}_{YX}\mathbf{R}^{-1}_{XX}\mathbf{R}_{XY}.$$

NÁVOD 3.9 (praktický výpočet). V praxi se většinou výběrový koeficient mnohonásobné korelace počítá pomocí nějakého software. Hledání inverzní matice $\mathbf{R}_{XX}^{-1}$ může být obecně složitý proces, proto ještě uvedeme alternativní výpočet. Položme $\mathbf{Z} = (Y, \mathbf{X})$ a $\mathbf{R} = \mathbf{R}_{ZZ}$. Pak

$$r^2_{Y\cdot\mathbf{X}} = 1 - \frac{\det(\mathbf{R})}{\det(\mathbf{R}_{XX})}.$$

Příklad 3.10. Zjišťujeme závislost koncentrace ozónu[2] (proměnná Y) ve spodních vrstvách atmosféry na meteorologických podmínkách, které jsou popsány intenzitou slunečního záření (X_1), rychlosti větru (X_2) a teplotě vzduchu (X_3). Naměřená data udává následující tabulka.

i	Y	X_1	X_2	X_3
1	23	148	8,00	82
2	21	191	14,90	77
3	37	284	20,70	72
4	20	37	9,20	65
5	12	120	11,50	73
6	13	137	10,30	76
7	135	269	4,10	84
8	49	248	9,20	85
9	32	236	9,20	81
10	64	175	4,60	83

Vypočtěte výběrový koeficient mnohonásobné korelace.

Řešení. $\mathbf{R}_{YX} = (0,55; -0,51; 0,54)$.

$$\mathbf{R}_{XX} = \begin{pmatrix} 1,00 & 0,19 & 0,60 \\ 0,19 & 1,00 & -0,52 \\ 0,60 & -0,52 & 1,00 \end{pmatrix}$$

Její inverze je tvaru

$$\mathbf{R}_{XX}^{-1} = \begin{pmatrix} 3,29 & -2,25 & -3,13 \\ -2,25 & 2,91 & 2,85 \\ -3,13 & 2,85 & 4,34 \end{pmatrix}$$

a celkově dostáváme $r^2_{Y\cdot\mathbf{X}} = \mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY} = 0,8557$.
Pokud bychom použili druhý způsob uvedený v Návodu 3.15, je třeba vypočítat matici $\mathbf{R}$, kterou lze z předešlého vyjádřit $\mathbf{R} = \begin{pmatrix} 1 & \mathbf{R}_{YX} \\ \mathbf{R}'_{YX} & \mathbf{R}_{XX} \end{pmatrix}$, tj.

$$\mathbf{R} = \begin{pmatrix} 1,00 & 0,55 & -0,51 & 0,54 \\ 0,55 & 1,00 & 0,19 & 0,60 \\ -0,51 & 0,19 & 1,00 & -0,52 \\ 0,54 & 0,60 & -0,52 & 1,00 \end{pmatrix}.$$

Pak

$$r^2_{Y\cdot\mathbf{X}} = 1 - \frac{\det(\mathbf{R})}{\det(\mathbf{R}_{XX})} = 1 - \frac{0,032}{0,22} = 0,8557.$$

[2] část datového souboru *airquality* implementovaného v jazyce R

Hodnota tohoto koeficientu poukazuje na do jisté míry velkou lineární závislost proměnné Y na ostatních proměnných. Tato hodnota je však značně ovlivněna také korelacemi proměnných X_1, X_2 a X_3 mezi sebou. Při pohledu na prvky matice $\mathbf{R}_{XX}$ vidíme, že je např. významná korelace mezi intenzitou slunečního záření (X_1) a teplotou vzduchu (X_3). Pro vyloučení těchto vlivů je třeba spočítat parciální korelační koeficienty – viz další odstavec.

3.2. Parciální korelační koeficient. Na závěr této kapitoly zavedeme pojem parciální korelační koeficient. Pro tento případ budeme uvažovat náhodné veličiny

$$Y, Z, X_1, \ldots, X_k.$$

Motivací k zavedení tohoto korelačního koeficientu je fakt, že korelační koeficient $R(Y, Z)$ mezi náhodnou veličinou Y a Z může být dosti vysoký proto, že obě náhodné veličiny jsou silně závislé na náhodném vektoru $\mathbf{X} = (X_1, \ldots, X_k)'$. Zajímá nás proto, jaká by byla korelace mezi Y a Z při vyloučení vlivu, který je způsoben náhodným vektorem $\mathbf{X}$.

Toto odstranění vlivu náhodného vektoru $\mathbf{X}$ lze uskutečnit tak, že se sleduje korelace mezi Y a Z při pevných hodnotách náhodného vektoru $\mathbf{X}$.

Protože v praktických situacích není možné uspořádání experimentu takovým způsobem, aby byla provedena eliminace vlivu náhodného vektoru $\mathbf{X}$, je třeba ji provést pomocí vhodného matematického modelu. Obdobně jako v případě koeficientu mnohonásobné korelace se omezíme pouze na lineární vztahy.

Označme $\widehat{Y}$ a $\widehat{Z}$ nejlepší lineární predikce náhodných veličin Y a Z pomocí náhodného vektoru $\mathbf{X}$. Korelaci očištěnou od vlivu náhodného vektoru $\mathbf{X}$ dostaneme, budeme-li počítat korelaci $R(Y - \widehat{Y}, Z - \widehat{Z})$. Definujme proto

DEFINICE 3.11. Nechť existuje korelační koeficient $R(Y - \widehat{Y}, Z - \widehat{Z})$. Potom jej budeme nazývat **parciálním korelačním koeficientem** náhodných veličin Y a Z při pevném $\mathbf{X}$ a budeme jej značit

$$\rho_{Y,Z\cdot\mathbf{X}} = R(Y - \widehat{Y}, Z - \widehat{Z}).$$

VĚTA 3.12. *Pro parciální korelační koeficient náhodných veličin Y a Z při pevném $\mathbf{X}$ platí*

$$\rho_{Y,Z\cdot\mathbf{X}} = \left[R(Y,Z) - R(Y,\mathbf{X})(R(\mathbf{X}))^{-1}R(\mathbf{X},Z)\right]\left[\left(1-\rho^2_{Y\cdot\mathbf{X}}\right)\left(1-\rho^2_{Z\cdot\mathbf{X}}\right)\right]^{-\frac{1}{2}}$$

Poznámka 3.13. Z hodnoty korelačního koeficientu $R(Y, Z)$ nelze usuzovat na velikost parciálního korelačního koeficientu $\rho_{Y,Z\cdot\mathbf{X}}$. Tyto dva koeficienty se od sebe mohou dosti odlišovat, mohou mít i různé znaménko a v případě, že jeden z nich je roven nule, může být druhý různý od nuly a podobně. Jejich vztah je tedy odlišný od vztahu $R(Y, X_j)$ a $\rho_{Y\cdot\mathbf{X}}$, který dává důsledek 3.12.

Pro praktické účely opět definujme výběrový protějšek k parciálnímu korelačnímu koeficientu.

DEFINICE 3.14. Mějme náhodné vektory $\begin{pmatrix} Y_1 \\ Z_1 \\ \mathbf{X}_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_n \\ Z_n \\ \mathbf{X}_n \end{pmatrix}$, kde Y_i, Z_i jsou náhodné veličiny a $\mathbf{X}_i \quad (i = 1, \cdots, n)$ jsou náhodné vektory typu $p \times 1$.
Pak **výběrový parciální korelační koeficient** je definován vztahem

$$r_{Y,Z\cdot\mathbf{X}} = \frac{r^2_{YZ} - r^2_{Y\cdot\mathbf{X}} r^2_{Z\cdot\mathbf{X}}}{\sqrt{\left(1 - r^2_{Y\cdot\mathbf{X}}\right)\left(1 - r^2_{Z\cdot\mathbf{X}}\right)}},$$

kde r^2_{YZ} je výběrový koeficient korelace náhodných veličin Y, Z a $r^2_{Y\cdot\mathbf{X}}$, $r^2_{Z\cdot\mathbf{X}}$ jsou příslušné výběrové koeficienty mnohonásobné korelace.

NÁVOD 3.15 (praktický výpočet). V praxi se pro výpočet parciálního korelačního koeficientu používá následujícího postupu. Položme $\mathbf{W} = (Y, Z, \mathbf{X})$ a $\mathbf{R} = \mathbf{R}_{WW}$. Pak

$$r_{Y,Z\cdot\mathbf{X}} = \frac{\det(\mathbf{R}_{(12)})}{\sqrt{\det(\mathbf{R}_{(11)})\det(\mathbf{R}_{(22)})}},$$

kde $\mathbf{R}_{(ij)}$ je submatice, která vznikne z $\mathbf{R}$ vynecháním i-tého řádku a j-tého sloupce.

Příklad 3.16. Na datech z Příkladu 3.10 vypočtěte parciální korelační koeficient $r_{Y,X_1\cdot(X_2,X_3)}$.

Řešení. Připomeňme matici $\mathbf{R}$, která byla tvaru

$$\mathbf{R} = \begin{pmatrix} 1,00 & 0,55 & -0,51 & 0,54 \\ 0,55 & 1,00 & 0,19 & 0,60 \\ -0,51 & 0,19 & 1,00 & -0,52 \\ 0,54 & 0,60 & -0,52 & 1,00 \end{pmatrix}.$$

Příslušné submatice jsou

$$\mathbf{R}_{(11)} = \begin{pmatrix} 1,00 & 0,19 & 0,60 \\ 0,19 & 1,00 & -0,52 \\ 0,60 & -0,52 & 1,00 \end{pmatrix},$$

$$\mathbf{R}_{(12)} = \begin{pmatrix} 0,55 & 0,19 & 0,60 \\ -0,51 & 1,00 & -0,52 \\ 0,54 & -0,52 & 1,00 \end{pmatrix},$$

$$\mathbf{R}_{(22)} = \begin{pmatrix} 1,00 & -0,51 & 0,54 \\ -0,51 & 1,00 & -0,52 \\ 0,54 & -0,52 & 1,00 \end{pmatrix}.$$

Po dosazení dostáváme

$$r_{Y,X_1\cdot(X_2,X_3)} = \frac{0,2827}{\sqrt{0,2220 \cdot 0,4654}} = 0,8795.$$

Výsledek lze interpretovat jako velikost lineární závislosti ozónu na intenzitě slunečního záření s vyloučením vlivu rychlosti větru a teploty vzduchu. Podobně by šlo zkoumat ostatní vazby mezi proměnnými.

Úlohy k procvičení

Cvičení 3.1. V tabulce jsou uvedeny výsledky měření (x_i, y_i) a predikované hodnoty $\hat{y}_i$, $i = 1, \ldots, 10$

i	1	2	3	4	5	6	7	8	9	10
x_i	1,60	1,86	2,21	2,29	3,38	3,42	3,62	3,65	3,76	4,27
y_i	3,24	3,12	3,81	5,12	6,28	7,15	7,33	7,81	8,08	8,43
$\hat{y}_i$	2,98	3,54	4,31	4,48	6,85	6,94	7,37	7,44	7,68	8,79

Určete index determinace a interpretujte ho.

$[ID = 0.95532]$

Cvičení 3.2. Během 14-ti dní byla měřena polední teplota vzduchu. K predikci teploty byly použity dva modely – model A a model B. Naměřené hodnoty a predikované hodnoty obou modelů jsou uvedeny v následující tabulce.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	0,35	-1,54	0,47	-0,50	-1,99	-2,17	-1,86	-1,37	-1,88	-2,30	-2,13	-2,12	-1,76	-1,06
$\hat{y}_i^A$	-0,62	-0,75	-0,87	-0,99	-1,11	-1,24	-1,36	-1,48	-1,60	-1,73	-1,85	-1,97	-2,09	-2,22
$\hat{y}_i^B$	-0,17	-0,35	-0,52	-0,70	-0,87	-1,05	-1,22	-1,39	-1,57	-1,74	-1,92	-2,09	-2,27	-2,44

Na základě indexu determinace rozhodněte, který z modelů je lepší.

$[ID_A = 0,31;\ ID_B = 0,24]$

Cvičení 3.3. Na datech ze Cvičení 3.2 byla predikována hodnota polední teploty vzduchu v 15. den. Model A tuto hodnotu odhadl $\hat{y}_{15}^A = -2,34$, predikce pomocí modelu B byla $\hat{y}_{15}^B = -2,61$. Ve skutečnosti byla naměřena hodnota $y_{15} = -1,34$. Na nových datech opět porovnejte oba modely pomocí indexu determinace.

$[ID_A = 0,22;\ ID_B = 0,09]$

Cvičení 3.4. Zjišťujeme závislost spotřeby paliva osobních automobilů[3] (proměnná Y, počet mil/galon) na vlastnostech motoru, které jsou popsány objemem válců (X_1, kubické palce), výkonem (X_2, počet koní), hmotností vozidla (X_3, kilolibry) a zrychlením (X_4, počet sekund na 1/4 míle). Naměřená data udává následující tabulka.

Model (r.v. 1974)	Y	X_1	X_2	X_3	X_4
Mazda RX4 Wag	21,00	160,00	110,00	2,88	17,02
Datsun 710	22,80	108,00	93,00	2,32	18,61
Hornet 4 Drive	21,40	258,00	110,00	3,21	19,44
Valiant	18,10	225,00	105,00	3,46	20,22
Merc 280C	17,80	167,60	123,00	3,44	18,90
Cadillac Fleetwood	10,40	472,00	205,00	5,25	17,98
AMC Javelin	15,20	304,00	150,00	3,44	17,30
Fiat X1-9	27,30	79,00	66,00	1,94	18,90
Porsche 914-2	26,00	120,30	91,00	2,14	16,70
Ford Pantera L	15,80	351,00	264,00	3,17	14,50

Vypočtěte závislost spotřeby paliva osobních automobilů na objemu válců, výkonu, hmotnosti a zrychlením vozidla.

$[r^2_{Y\cdot\mathbf{X}} = 0,934]$

Cvičení 3.5. V rámci biometrického výzkumu byl na jednotlivých stromech zjišťován vztah mezi veličinami objem (Y, m^3), výčetní tloušťka (X_1, cm), výška (X_2, m) a délka zelené koruny (X_3, m). Naměřené hodnoty jsou uvedeny v následující tabulce.

[3] část datového souboru *mtcars* implementovaného v jazyce R

Strom	Y	X_1	X_2	X_3
1	0,013	8	9,8	3,6
2	0,021	8	10,2	3,6
3	0,012	7	9,4	3,0
4	0,009	7	7,8	1,4
5	0,065	12	11,2	4,6
6	0,071	12	12,0	5,1
7	0,102	13	13,5	6,9
8	0,048	10	12,1	4,6
9	0,049	11	10,8	4,3
10	0,011	7	8,9	3,9
11	0,017	8	9,3	3,5
12	0,059	11	12,0	4,8

Vyšetřete korelační závislost objemu na tloušťce, výšce a délce zelené koruny.

$[r^2_{Y\cdot\mathbf{X}} = 0,9634]$

Cvičení 3.6. Na datech ze Cvičení 3.4 vypočtěte parciální korelační koeficienty $r_{Y,X_1\cdot(X_2,X_3,X_4)}$, $r_{Y,X_2\cdot(X_1,X_3,X_4)}$, $r_{Y,X_3\cdot(X_1,X_2,X_4)}$, $r_{X_1,X_4\cdot(X_1,X_2,X_3)}$.

$[r_{Y,X_1\cdot(X_2,X_3,X_4)} = 0,2319$; $r_{Y,X_2\cdot(X_1,X_3,X_4)} = -0,5219$; $r_{Y,X_3\cdot(X_1,X_2,X_4)} = -0,7405$; $r_{X_1,X_4\cdot(X_1,X_2,X_3)} = -0,0736.]$

Cvičení 3.7. Na datech ze Cvičení 3.5 vypočtěte všechny parciální korelační koeficienty.

$[r_{Y,X_1\cdot(X_2,X_3)} = 0,8558$; $r_{Y,X_2\cdot(X_1,X_3)} = 0,1938$; $r_{Y,X_3\cdot(X_1,X_2)} = 0,2974$; $r_{X_1,X_2\cdot(Y,X_3)} = 0,1248$; $r_{X_1,X_3\cdot(Y,X_2)} = -0,22$; $r_{X_2,X_3\cdot(Y,X_1)} = 0,6161.]$

KAPITOLA 4

Lineární regresní model

Základní informace

(1) V následující kapitole se budeme zabývat statistickou analýzou vzájemných vztahů náhodných jevů, kde se bude předpokládat, že tyto vztahy lze popsat pomocí lineárních operací. Popíšeme obecně volbu optimálních parametrů lineárního modelu a uvedeme konkrétní příklady.
(2) Předpokládá se znalost základních pojmů z teorie pravděpodobnosti a matematické statistiky – náhodná veličina, náhodný vektor, jejich číselné charakteristiky a jejich výběrové odhady, testování hypotéz

Výstupy z výukové jednotky

Studenti

- definují lineární regresní model
- vypočítají odhady neznámých parametrů a testují hypotézy o těchto parametrech
- ovládají nejdůležitější aplikace: klasické regresní modely – regresní přímka, polynomická regrese

1. Motivace

Často chceme prozkoumat vztah mezi dvěma veličinami, kde jedna z nich, tzv. „nezávisle proměnná“ X, má řídit druhou, tzv. „závisle proměnnou“ Y. Předpokládá se, že obě veličiny jsou spojité. Prvním krokem ve zkoumání by mělo být zakreslení dat do grafu. V řadě případů tento krok napoví mnohé o tom, co nás zajímá: Existuje vztah mezi oběma proměnnými (veličinami)? Pokud ano, pak rostou či klesají obě v jednom směru, nebo jedna klesá, když druhá roste? Je přímka vhodným modelem pro vyjádření vztahu mezi těmito dvěma veličinami? Chceme-li se dostat dále za tuto intuitivní úroveň analýzy, je lineární regrese často užitečným nástrojem. Tato metoda zahrnuje proložení přímky daty a analýzu statistických vlastností takovéto přímky.

2. Lineární regresní model

Předpokládejme, že mezi nějakými nenáhodnými veličinami $y, x_1, \ldots, x_k$ platí lineární vztah

$$y = \beta_1 x_1 + \cdots + \beta_k x_k,$$

ve kterém $\beta_1, \ldots, \beta_k$ jsou neznámé parametry. Informace o neznámých parametrech budeme získávat pomocí experimentu, a to tak, že opakovaně budeme měřit hodnoty veličiny y při vybraných hodnotách proměnných $x_1, \ldots, x_k$. Při měřeních však vznikají chyby, což lze modelovat takto

$$Y = \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

kde ε je náhodná chyba měření.

Opakované hodnoty sledovaných veličin budeme pro $i = 1, \ldots, n$ značit $Y_i, x_{i1}, \ldots, x_{ik}$, obdobně také náhodné chyby ε_i.

Celkově jsme dostali model

$$\begin{aligned} Y_1 &= \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ &\vdots \\ Y_n &= \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \quad \text{vyjádřeme maticově} \quad \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{\mathbf{X}\text{(matice plánu)}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

O náhodných chybách $\varepsilon_1, \ldots, \varepsilon_n$ budeme předpokládat, že jsou

- **nesystematické**, což lze matematicky vyjádřit požadavkem, že $\boxed{E\varepsilon_i = 0}$, $i = 1, \ldots, n$, tj. $\boxed{E\boldsymbol{\varepsilon} = \mathbf{0}}$ a tedy $\boxed{E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}}$
- **homogenní v rozptylu**, tj. že $\boxed{D\varepsilon_i = \sigma^2 > 0}$ pro $i = 1, \ldots, n$;
- jednotlivé náhodné chyby jsou **nekorelované**, tj. že $\boxed{C(\varepsilon_i, \varepsilon_j) = 0}$ pro $i \neq j, i, j = 1, \ldots, n$, tj. $\boxed{D\mathbf{Y} = D\boldsymbol{\varepsilon} = \sigma_2 \mathbf{I}_n}$, takže i měření jsou nekorelovaná.

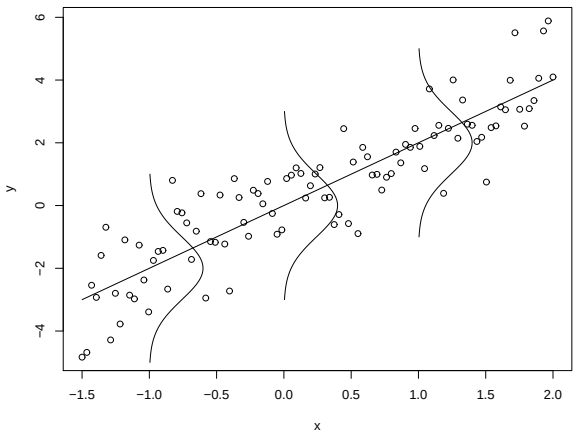
Používá se následující **terminologie** a značení

- parametry $\beta_1, \ldots, \beta_k$ se nazývají **regresní koeficienty**;
- matice $\mathbf{X}$ obsahuje nenáhodné prvky x_{ij} a nazývá se **regresní maticí** nebo **maticí plánu** (*Design Matrix*);
- popsaný model souhrnně zapíšeme jako $\boxed{\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)}$.

Takto zavedený model budeme nazývat **linerární regresní model**. Dále budeme předpokládat, že $\boxed{n > k}$ a o hodnosti matice $\mathbf{X}$ budeme předpokládat, že je rovna k, tj. $\boxed{h(\mathbf{X}) = k}$. Bude-li tento přepoklad splněn, budeme říkat, že jde **linerární regresní model plné hodnosti**. V tom případě jsou sloupce matice $\mathbf{X}$ nezávislé.

V opačném případě, by bylo možné daný sloupec matice $\mathbf{X}$ napsat jako lineární kombinaci ostatních sloupců, což je možné interpretovat tak, že proměnná odpovídající danému sloupci je nadbytečná, protože ji lze vyjádřit jako lineární funkci ostatních proměnných.

Příklad 2.1. Regresní přímka v klasickém lineárním regresním modelu



Klasickým speciálním případem lineárního modelu je **jednoduchá lineární regrese**, kdy předpokládáme, že nezávislé náhodné veličiny Y_i $(i = 1, \ldots, n)$ mají normální rozdělení

$$\boxed{Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)},$$

kde x_i jsou dané konstanty, které nejsou všechny stejné. Rozptyly Y_i jsou stejné, kdežto střední hodnoty lze vyjádřit jako lineární funkci známých konstant x_i pomocí neznámých parametrů β_0, β_1.

V tomto případě

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \text{ matice plánu: } \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

3. Odhady neznámých parametrů

V dalším se budeme věnovat odhadům vektoru neznámých parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$.

DEFINICE 3.1. Řekneme, že odhad $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je **lineárním odhadem** vektoru $\boldsymbol{\beta}$, jestliže existuje matice reálných čísel $\mathbf{B}_{k\times n}$ taková, že $\widehat{\boldsymbol{\beta}} = \mathbf{BY}$.
Dále řekneme, že odhad $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je **nestranným odhadem** vektoru $\boldsymbol{\beta}$, jestliže pro každé $\boldsymbol{\beta} \in \mathbb{R}^k$ platí $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.
Jestliže $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je takový lineární nestranný odhad vektoru parametrů $\boldsymbol{\beta}$, že pro každý jiný lineární nestranný odhad $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\mathbf{Y})$ je rozdíl variančních matic $D\widetilde{\boldsymbol{\beta}}(\mathbf{Y}) - D\widehat{\boldsymbol{\beta}}(\mathbf{Y})$ **pozitivně semidefinitní matice**, potom budeme říkat, že $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Y})$ je **nejlepší nestranný lineární odhad** (*Best Linear Unbiased Estimator*) parametrů $\boldsymbol{\beta}$, zkráceně $BLUE$ odhad.

V další části budeme hledat $BLUE$-odhad parametru $\boldsymbol{\beta}$ a odvodíme jeho vlastnosti. Tento odhad budeme hledat metodou nejmenších čtverců (*Ordinary Least Square Method*).

DEFINICE 3.2. Řekneme, že odhad $\widehat{\boldsymbol{\beta}}_{OLS}$ je odhadem parametru $\boldsymbol{\beta}$ **metodou nejmenších čtverců**, jestliže

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2$$

VĚTA 3.3. *Odhad parametru $\boldsymbol{\beta}$ v modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ je tvaru*

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Důkaz. Nejprve označme symbolem $\mathbf{x}_i'$ i-tý řádek matice plánu $\mathbf{X}$ a symbolem $\mathbf{X}_j$ j-tý sloupec této matice, tj.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \ldots \mathbf{X}_k \end{pmatrix}$$

Nutnou podmínkou pro extrém je, aby parciální derivace byly nulové, tj. pro $s = 1, \ldots, k$

$$0 = \frac{\partial}{\partial\beta_s} S(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_s} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_s} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2.$$

Proto počítejme

$$\begin{aligned} \boxed{\frac{\partial}{\partial\beta_s} S(\boldsymbol{\beta})} &= \frac{\partial}{\partial\beta_s} \sum_{i=1}^{n} \left[Y_i^2 - 2Y_i \sum_{j=1}^{k} x_{ij}\beta_j + \left(\sum_{j=1}^{k} x_{ij}\beta_j \right)^2 \right] \\ &= -2\sum_{i=1}^{n} Y_i x_{is} + 2\sum_{i=1}^{n} \left(\sum_{j=1}^{k} x_{ij}\beta_j \right) x_{is} \\ &= -2\sum_{i=1}^{n} Y_i x_{is} + 2\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}x_{is}\beta_j = \boxed{0} \end{aligned}$$

tj.

$$\boxed{\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}x_{is}\beta_j = \sum_{i=1}^{n} Y_i x_{is}}.$$

Nyní se budeme snažit vyjádřit předchozí rovnost maticově. Upravujme postupně levou a pravou stranu:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}x_{is}\beta_s = \sum_{i=1}^{n} x_{is}\underbrace{\sum_{j=1}^{k} x_{ij}\beta_j}_{=\mathbf{x}_i'\boldsymbol{\beta}} = \sum_{i=1}^{n} x_{is}\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{X}_s'\underbrace{\begin{pmatrix}\mathbf{x}_1'\boldsymbol{\beta}\\ \vdots \\ \mathbf{x}_n'\boldsymbol{\beta}\end{pmatrix}}_{=\mathbf{X}\boldsymbol{\beta}} = \mathbf{X}_s'\mathbf{X}\boldsymbol{\beta}$$

$$\sum_{i=1}^{n} Y_i x_{is} = \mathbf{X}_s'\mathbf{Y}$$

a celkově, zapíšeme-li k rovnic pod sebe a uvažujeme-li obě strany rovnosti, dostaneme

$$\underbrace{\begin{pmatrix}\mathbf{X}_1'\\ \vdots \\ \mathbf{X}_k'\end{pmatrix}\begin{pmatrix}\mathbf{x}_1'\\ \vdots \\ \mathbf{x}_n'\end{pmatrix}\boldsymbol{\beta}}_{=\mathbf{X}'\mathbf{X}\boldsymbol{\beta}} = \underbrace{\begin{pmatrix}\mathbf{X}_1'\\ \vdots \\ \mathbf{X}_k'\end{pmatrix}\mathbf{Y}}_{=\mathbf{X}'\mathbf{Y}} \qquad \ldots \qquad \text{tzv. normální rovnice}$$

Vzhledem k předpokladu, že jde o model plné hodnosti, tj. $h(\mathbf{X}) = h(\mathbf{X}'\mathbf{X}) = k$, řešení normálních rovnic má tvar

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Nyní zbývá dokázat, že tento extrém je také minimem, tj. že matice druhých parciálních derivací je pozitivně semidefinitní matice. Proto počítejme (sh)-tý prvek matice druhých parciálních derivací

$$\begin{aligned}\frac{\partial^2}{\partial\beta_s\partial\beta_h}S(\boldsymbol{\beta}) &= \frac{\partial}{\partial\beta_h}\left[-2\sum_{i=1}^{n} Y_i x_{is} + 2\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}x_{is}\beta_j\right]\\ &= 2\sum_{i=1}^{n} x_{is}\underbrace{\frac{\partial}{\partial\beta_h}\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)}_{=x_{ih}} = 2\sum_{i=1}^{n} x_{is}x_{ih} = 2\mathbf{X}_s'\mathbf{X}_h\end{aligned}$$

Takže matice druhých parciálních derivací je $\left(\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_s\partial\beta_h}\right)_{s,h=1}^{k} = \left(\sum_{i=1}^{n} x_{is}x_{ih}\right)_{s,h=1}^{k} = \mathbf{X}'\mathbf{X} > 0$,

tj. jde o pozitivně definitní matici a tím je věta dokázaná. □

VĚTA 3.4. *(Gaussova-Markovova věta). Odhad $\widehat{\boldsymbol{\beta}}_{OLS}$ v modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ je BLUE-odhad (tj. je nejlepší nestranný lineární odhad) a jeho variační matice je rovna*

$$D\widehat{\boldsymbol{\beta}}_{OLS} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

VĚTA 3.5. *Pro libovolný vektor $\mathbf{c} \in \mathbb{R}^k$ je $\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS}$ BLUE-odhad parametrické funkce $\mathbf{c}'\boldsymbol{\beta}$ a má rozptyl $\sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$.*

VĚTA 3.6. *Platí*

$$S_e = S(\widehat{\boldsymbol{\beta}}_{OLS}) = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_{OLS}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y},$$

kde $\mathbf{H}$ je tzv. „hat“ matice

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

VĚTA 3.7. *Odhad $\boxed{s^2 = \frac{S_e}{n-k}}$ je nestranným odhadem rozptylu σ^2.*

4. Testování hypotéz v lineárním regresním modelu

Díky předchozím větám dokážeme v lineárním regresním modelu plné hodnosti vypočítat nejen OLS-odhady neznámých parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, ale také máme k dispozici odhad neznámého rozptylu σ^2 a známe vlastnosti těchto odhadů.

V dalším se zaměříme na stanovení jejich rozdělení v případě, že náhodný vektor $\boxed{\mathbf{Y}}$ má **vícerozměrné normální rozdělení**. Pak teprve budeme moci přejít k testování hypotéz o neznámých parametrech $\beta_1, \ldots, \beta_k$.

Jestliže náhodný vektor $\mathbf{Y}$ se řídí lineárním regresním modelem plné hodnosti, což zapisujeme $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, a navíc má **vícerozměrné normální rozdělení**, budeme psát

$$\boxed{\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)}.$$

VĚTA 4.1. *Mějme lineární regresní model plné hodnosti, přičemž* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. *Pak platí*

(a) OLS-odhad vektoru neznámých parametrů má normální rozdělení

$$\widehat{\boldsymbol{\beta}}_{OLS} \quad \sim \quad N_k\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

(b) náhodná veličina

$$K = \frac{n-k}{\sigma^2}s^2 \quad \sim \quad \chi^2(n-k)$$

(c) náhodná veličina $\boxed{K = \frac{n-k}{\sigma^2}s^2}$ *a OLS-odhad* $\boxed{\widehat{\boldsymbol{\beta}}_{OLS}}$ *jsou nezávislé.*

Díky tomuto tvrzení lze dokázat následující větu.

VĚTA 4.2. *V modelu* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *plné hodnosti pro každé* $\mathbf{c} \in \mathbb{R}^k$, $\mathbf{c} \neq \mathbf{0}$ *platí*

$$T = \frac{\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}'\boldsymbol{\beta}}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \quad \sim \quad t(n-k).$$

DŮSLEDEK 4.3. *V modelu* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *plné hodnosti má* $100(1-\alpha)\%$ *interval spolehlivosti pro parametrickou funkci* $\mathbf{c}'\boldsymbol{\beta}$ *(kde* $\mathbf{c} \neq \mathbf{0}$*) tvar*

$$(D_T, H_T) = \left(\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} - s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}\, t_{1-\alpha/2}(n-k), \mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} + s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}\, t_{1-\alpha/2}(n-k)\right).$$

Poznámka 4.4. Prakticky lze provést test hypotézy $H_0 : \mathbf{c}'\boldsymbol{\beta} = \gamma_0$ (γ_0 je dané reálné číslo) proti alternativě $H_1 : \mathbf{c}'\boldsymbol{\beta} \neq \gamma_0$ na hladině významnosti α tak, že hypotézu H_0 **zamítáme**, pokud platí

$$\frac{\left|\mathbf{c}'\widehat{\boldsymbol{\beta}}_{OLS} - \gamma_0\right|}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \geq t_{1-\alpha/2}(n-k)$$

Poznámka 4.5. V praktických situacích se nejčastěji volí vektor $\mathbf{c}$ jako jednotkový s jedničkou na j-tém místě $\mathbf{c} = (0, \ldots, 1, 0, \ldots, 0)'$ a v tom případě $\boxed{\mathbf{c}'\boldsymbol{\beta} = \beta_j}$, takže

(a) $100(1-\alpha)\%$ interval spolehlivosti má tvar (při značení $(\mathbf{X}'\mathbf{X})^{-1} = (v_{ij})_{i,j=1}^k$)

$$\left(\widehat{\beta}_{OLS,j} - s\sqrt{v_{jj}}\, t_{1-\alpha/2}(n-k) \;\;,\;\; \widehat{\beta}_{OLS,j} + s\sqrt{v_{jj}}\, t_{1-\alpha/2}(n-k)\right).$$

(b) Test hypotézy $H_0 : \beta_j = \gamma_0$ (γ_0 je dané reálné číslo) proti alternativě $H_1 : \beta_j \neq \gamma_0$ na hladině významnosti α se provede tak, že hypotézu H_0 **zamítáme**, pokud platí

$$\frac{\left|\widehat{\beta}_{OLS,j} - \gamma_0\right|}{s\sqrt{v_{jj}}} \geq t_{1-\alpha/2}(n-k).$$

Před další větou zavedeme následující bloková značení:

$$\boldsymbol{\beta} = (\underbrace{\beta_1, \ldots, \beta_m}_{=\boldsymbol{\beta}_1'}, \underbrace{\beta_{m+1}, \ldots, \beta_k}_{=\boldsymbol{\beta}_2'})',$$

obdobně

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\widehat{\boldsymbol{\beta}}'_{OLS,1}, \widehat{\boldsymbol{\beta}}'_{OLS,2})'$$

a nakonec také pro matici

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix},$$

kde matice $\mathbf{V}_{11}$ je typu $m \times m$.

VĚTA 4.6. *V modelu* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ *plné hodnosti platí, že statistika*

$$F = \frac{1}{s^2(k-m)} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_2\right)' \mathbf{V}_{22}^{-1} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_2\right) \quad \sim \quad F(k-m, n-k).$$

Poznámka 4.7. Díky předcházející větě můžeme testovat nulovou hypotézu

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2,0},$$

(kde $\boldsymbol{\beta}_{2,0}$ je daný vektor reálných čísel, nejčastěji nulový vektor)

proti alternativě

$$H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_{2,0}$$

na hladině významnosti α tak, že hypotézu H_0 **zamítáme**, pokud platí

$$F_0 = \frac{1}{s^2(k-m)} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_{2,0}\right)' \mathbf{V}_{22}^{-1} \left(\widehat{\boldsymbol{\beta}}_{OLS,2} - \boldsymbol{\beta}_{2,0}\right) \geq F_{1-\alpha}(k-m, n-k).$$

5. Speciální modely lineární regrese

Speciální volba matice $\mathbf{X}$ vede ke speciálním modelům lineární regrese, které popisují časté experimentální situace.

MODEL I: **Regresní přímka** $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n;\ n > 2$.

Matice plánu $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$, $\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}$

a model bude plné hodnosti, pokud všechny hodnoty $x_1, \ldots, x_n$ nebudou stejné.
Normální rovnice jsou tvaru:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n Y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i Y_i \end{aligned}$$

MODEL II: **Regrese procházející počátkem** $Y_i = \beta x_i + \varepsilon_i$, $i = 1, \ldots, n;\ n > 1$.

Matice plánu $\mathbf{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $\mathbf{X}'\mathbf{X} = \left(\sum_{i=1}^n x_i^2\right)$, $\mathbf{X}'\mathbf{Y} = \left(\sum_{i=1}^n x_i Y_i\right)$

a model bude plné hodnosti, pokud alespoň jedna z hodnot $x_1, \ldots, x_n$ bude různá od nuly.

Normální rovnice:

$$\beta \sum_{i=1}^{n} x_i^2 \;=\; \sum_{i=1}^{n} x_i Y_i$$

MODEL III: **Kvadratická regrese** $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, $i = 1, \ldots, n;\ n > 3$.

Matice plánu $\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$, $\mathbf{X'X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{pmatrix}$, $\mathbf{X'Y} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \\ \sum_{i=1}^{n} x_i^2 Y_i \end{pmatrix}$

Normální rovnice jsou tvaru:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} Y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i Y_i \\ \beta_0 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 Y_i \end{aligned}$$

MODEL IV: **Polynomická regrese** $Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_m x_i^m + \varepsilon_i$, $i = 1, \ldots, n;\ n > m+1$.

$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}$, $\mathbf{X'X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^m \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^m & \sum_{i=1}^{n} x_i^{m+1} & \cdots & \sum_{i=1}^{n} x_i^{2m} \end{pmatrix}$, $\mathbf{X'Y} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^m Y_i \end{pmatrix}$

Při polynomické regresi vyšších řádů je třeba kontrolovat, zda matice $\mathbf{X'X}$ není špatně podmíněná, což nastává, pokud determinant této matice je blízký nule. Tento jev se také nazývá **multikolinearitou**. Pro posuzování multikolinearity existuje řada orientačních kritérií.

MODEL V: **Dva nezávislé výběry se stejnou variabilitou** $Y_{jk} = \mu + \alpha + \varepsilon_{jk}$, $j = 1, 2$, $k = 1, \ldots, n_j;\ n = n_1 + n_2 > 2$, kde μ budeme chápat jako společnou hladinu a α jako příspěvek druhého výběru.

Matice plánu $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$, $\mathbf{X'X} = \begin{pmatrix} n & n_2 \\ n_2 & n_2 \end{pmatrix}$, $\mathbf{X'Y} = \begin{pmatrix} Y_{\cdot\cdot} \\ Y_{2\cdot} \end{pmatrix}$

s použitím tzv. tečkové notace $Y_{\cdot\cdot} = \sum_{j=1}^{2} \sum_{k=1}^{n_j} Y_{jk}$ a $Y_{j\cdot} = \sum_{k=1}^{n_j} Y_{jk}$ pro $j = 1, 2$.

Normální rovnice jsou tvaru:

$$\begin{aligned} n\mu + n_2\alpha &= Y_{\cdot\cdot} \\ n_2\mu + n_2\alpha &= Y_{2\cdot} \end{aligned}$$

Odečteme-li od první rovnice druhou, dostaneme

$$\mu \underbrace{(n - n_2)}_{=n_1} = \underbrace{Y_{\cdot\cdot} - Y_{2\cdot}}_{=Y_{1\cdot}} \qquad \Rightarrow \qquad \mu = \tfrac{1}{n_1} Y_{1\cdot} = \bar{Y}_1$$ (výběrový průměr 1. výběru).

Pokud obě strany druhé rovnice vydělíme výrazem n_2, můžeme psát

$\mu + \alpha = \frac{1}{n_2} Y_{2\cdot} = \bar{Y}_2 \qquad \Rightarrow \qquad \alpha = Y_{2\cdot} - Y_{1\cdot}$ (rozdíl mezi 1. a 2. výběrovým průměrem)

MODEL VI: **Více nezávislých výběrů s homogenním rozptylem** $Y_{jk} = \mu + \alpha_j + \varepsilon_{ij}$, $j = 1, \ldots, J$, $k = 1, \ldots, n_j$; $n = n_1 + \cdots + n_J > J + 1$.

(a) Pokud bychom jako neznámé parametry uvažovali $\mu, \alpha_1, \ldots, \alpha_J$, pak dostaneme model, který **není plné hodnosti**, neboť první sloupec je součtem všech ostatních. Říkáme, že model je PŘEPARAMETRIZOVÁN (tzv. *"overparametrized model"*).

$$\underbrace{\begin{pmatrix} Y_{11} \\ \vdots \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \vdots \\ \vdots \\ Y_{J1} \\ \vdots \\ \vdots \\ Y_{Jn_J} \end{pmatrix}}_{=\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & 0 & \cdots & \cdots & 0 & 1 & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & \cdots & 0 & 1 & 0 \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}}_{=\mathbf{X}^*} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_J \end{pmatrix}}_{=\boldsymbol{\beta}^*} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{J1} \\ \vdots \\ \vdots \\ \varepsilon_{Jn_J} \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

(b) Protože matice plánu $\mathbf{X}$ není plné hodnosti, proveďme proto následující **reparametrizaci**: $\boxed{a_j = \mu_j - \mu.}$ Pak $Y_{jk} = \mu_j + \varepsilon_{jk}$, $j = 1, \cdots, J \quad i = 1, \cdots, n_j$. Maticově, lze napsat tento regresní model ve tvaru

$$\underbrace{\begin{pmatrix} Y_{11} \\ \vdots \\ \vdots \\ \underline{Y_{1n_1}} \\ Y_{21} \\ \vdots \\ \vdots \\ \underline{Y_{2n_2}} \\ \vdots \\ \vdots \\ \vdots \\ \overline{Y_{J1}} \\ \vdots \\ \vdots \\ Y_{Jn_J} \end{pmatrix}}_{=\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{1} & \underline{0} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{0} \\ 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{0} & \underline{1} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{0} \\ 0 & 0 & \cdots & \cdots & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{0} & \underline{0} & \underline{\cdots} & \underline{\cdots} & \underline{0} & \underline{1} & \underline{0} \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_J \end{pmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \vdots \\ \underline{\varepsilon_{1n_1}} \\ \varepsilon_{21} \\ \vdots \\ \vdots \\ \underline{\varepsilon_{2n_2}} \\ \vdots \\ \vdots \\ \vdots \\ \overline{\varepsilon_{J1}} \\ \vdots \\ \vdots \\ \varepsilon_{Jn_J} \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n_J \end{pmatrix} \text{ a } \mathbf{X}'\mathbf{Y} = \begin{pmatrix} Y_{1.} \\ Y_{2.} \\ \vdots \\ Y_{J.} \end{pmatrix} \Rightarrow \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \bar{Y}_1 \\ \vdots \\ \bar{Y}_J \end{pmatrix}$$

takže dostáváme

$$\begin{aligned} \mu_1 &= \bar{Y}_1 \\ &\vdots \\ \mu_J &= \bar{Y}_J \end{aligned}$$

Model VI: **Dvě regresní přímky** (se stejným rozptylem).

Mějme dva nezávislé náhodné výběry $Y_{11}, \ldots, Y_{1n_1}$ (resp. $Y_{21}, \ldots, Y_{2n_2}$) a k tomu odpovídající hodnoty regresorů $x_{11}, \ldots, x_{1n_1}$ (resp. $x_{21}, \ldots, x_{2n_2}$). Předpokládejme, že platí

$$\begin{aligned} Y_{1i} &= a_1 + b_1 x_{1i} + \varepsilon_{1i}, \quad i = 1, \ldots, n_1, \quad \varepsilon_{1i} \sim N(0, \sigma_1^2) \\ Y_{2i} &= a_2 + b_2 x_{2i} + \varepsilon_{2i}, \quad i = 1, \ldots, n_2, \quad \varepsilon_{2i} \sim N(0, \sigma_2^2) \end{aligned}$$

Vytvořme společný regresní model:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ \underline{Y_{1n_1}} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \underline{1} & \underline{x_{1n_1}} & \underline{0} & \underline{0} \\ 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} a_1 \\ \underline{b_1} \\ a_2 \\ b_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \underline{\varepsilon_{1n_1}} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}.$$

Vyjádřeno blokově:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

Počítejme postupně

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1'\mathbf{Y}_1 \\ \mathbf{X}_2'\mathbf{Y}_2 \end{pmatrix} \text{ a } \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}_2 \end{pmatrix}.$$

Označme

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\varepsilon}}_1 \\ \hat{\boldsymbol{\varepsilon}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 - \widehat{\mathbf{Y}}_1 \\ \mathbf{Y}_2 - \widehat{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 \\ \mathbf{Y}_2 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$$

Pak

$$SSE = \widehat{\boldsymbol{\varepsilon}}\,'\,\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\varepsilon}}_1\,'\,\widehat{\boldsymbol{\varepsilon}}_1 + \widehat{\boldsymbol{\varepsilon}}_2\,'\,\widehat{\boldsymbol{\varepsilon}}_2 = SSE_1 + SSE_2$$

a

$$\begin{aligned} s_1^2 &= \tfrac{SSE_1}{n_1-2} = \tfrac{\widehat{\boldsymbol{\varepsilon}}_1\,'\,\widehat{\boldsymbol{\varepsilon}}_1}{n_1-2} \\ s_2^2 &= \tfrac{SSE_1}{n_2-2} = \tfrac{\widehat{\boldsymbol{\varepsilon}}_2\,'\,\widehat{\boldsymbol{\varepsilon}}_2}{n_2-2} \end{aligned} \quad \Rightarrow \quad s^2 = \frac{SSE}{n_1+n_2-4} = \frac{(n_1-2)s_1^2 + (n_2-2)s_2^2}{n_1+n_2-4}$$

Testování rovnoběžnosti dvou regresních přímek.
Při testování hypotézy $H_0 : b_1 = b_2$ proti alternativě $H_1 : b_1 \neq b_2$ využijeme toho, že statistika

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{s^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \quad \sim \quad t(n-p-1).$$

Položme

$$\mathbf{c} = (0,1,0,-1) \quad \Rightarrow \quad \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = v_{22} + v_{44}, \text{ přičemž } (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & v_{44} \end{pmatrix}.$$

Za platnosti nulové hypotézy statistika

$$T_0 = \tfrac{\hat{b}_1 - \hat{b}_2}{s\sqrt{v_{22}+v_{44}}} \sim t(n_1+n_2-4).$$

Nulovou hypotézu zamítáme na hladině významnosti α, pokud

$$|t_0| > t_{1-\frac{\alpha}{2}}(n_1+n_2-4)$$

nebo pomocí *p-value*

$$p_0 = P(T_0 > |t_0|) < \tfrac{\alpha}{2}.$$

Testování shodnosti dvou regresních přímek.

Budeme testovat hypotézu $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ proti alternativě $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$.

Využijeme vlastnosti

$$\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2 \sim N\Big(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2, \sigma^2\big(\underbrace{(\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}}_{W}\big)\Big).$$

a

$$K_1 = \frac{1}{\sigma^2}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)'W^{-1}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2) \sim \chi^2(2),$$

dále

$$K_2 = \frac{SSE}{\sigma^2} = \frac{(n_1 + n_2 - 4)s^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 4),$$

takže k testování nulové hypotézy použijeme statistiku

$$F_0 = \frac{K_1/2}{K_2/(n_1+n_2-4)} = \frac{1}{2s^2}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)'W^{-1}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2) \sim F(2, n_1 + n_2 - 4)$$

a **nulovou hypotézu zamítáme** na hladině významnosti α, pokud

$$f_0 < F_{\frac{\alpha}{2}}(2, n_1 + n_2 - 4) \text{ nebo } f_0 > F_{1-\frac{\alpha}{2}}(2, n_1 + n_2 - 4)$$

nebo pomocí *p-value*, jestliže

$$p_0 = P(F > f_0) < \tfrac{\alpha}{2} \text{ popřípadě } 1 - p_0 < \tfrac{\alpha}{2}.$$

Ověřování shodnosti rozptylů.

Při testování hypotézy $H_0 : \sigma_1^2 = \sigma_2^2$ proti alternativě $H_1 : \sigma_1^2 \neq \sigma_2^2$ využijeme toho, že statistika

$$F_0 = \frac{\frac{SSE_1}{(n_1-2)\sigma^2}}{\frac{SSE_2}{(n_2-2)\sigma^2}} = \frac{s_1^2}{s_2^2} \sim F(n_1 - 2, n_2 - 2)$$

a **nulovou hypotézu zamítáme** na hladině významnosti α, pokud

$$f_0 < F_{\frac{\alpha}{2}}(n_1 - 2, n_2 - 2) \text{ nebo } f_0 > F_{1-\frac{\alpha}{2}}(n_1 - 2, n_2 - 2)$$

nebo pomocí *p-value*, jestliže

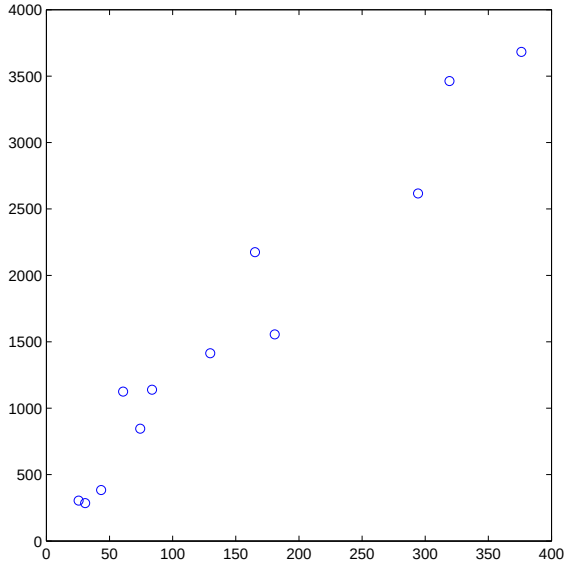
$$p_0 = P(F_0 > f_0) < \tfrac{\alpha}{2} \text{ nebo } 1 - p_0 < \tfrac{\alpha}{2}.$$

Příklad 5.1. Analyzujte data o počtu pracovních hodin za měsíc spojených s provozováním anesteziologické služby v závislosti na velikosti spádové populace nemocnice (viz následující

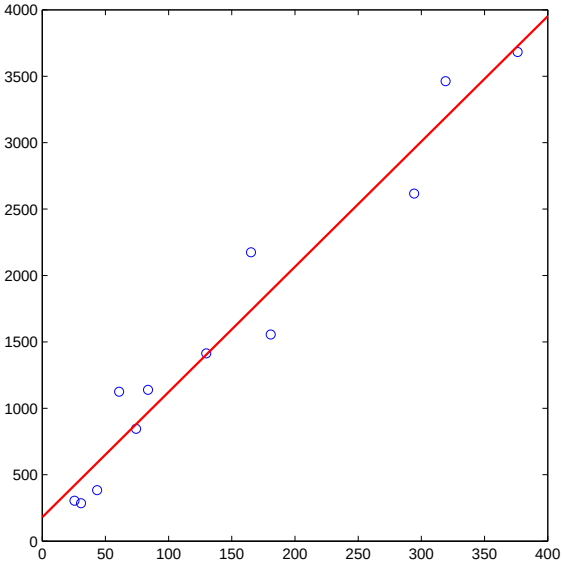
tabulka). Údaje byly získány ve 12 nemocnicích ve Spojených státech.

Poř. číslo	Počet pracovních hodin	Velikost populace spádové oblasti (osoby v tisích)
1	304,37	25,5
2	2616,32	294,3
3	1139,12	83,7
4	285,43	30,7
5	1413,77	129,8
6	1555,68	180,8
7	383,78	43,4
8	2174,27	165,2
9	845,30	74,3
10	1125,28	60,8
11	3462,60	319,2
12	3682,33	376,2

Závislost počtu pracovních hodin na velikosti populace



Řešení. Graf naznačuje lineární vztah mezi pracovní dobou a velikostí populace, a tak budeme pokračovat kvantifikací tohoto vztahu pomocí přímky $y = \beta_0 + \beta_1 x$.



Používáme-li model regresní analýzy pro statistické zpracování našich dat, je dobré ověřit předpoklady, ze kterých model vychází. Shrňme je v následujících třech bodech.

(1) Závisle proměnná Y (pracovní doba) má normální rozdělení pro každou hodnotu nezávisle proměnné x (velikost populace).
(2) Rozptyl závisle proměnné Y je stejný pro každou hodnotu nezávisle proměnné x.
(3) Závislost veličiny Y na x je lineární.

Pro tuto chvíli předpokládejme, že pro náš příklad jsou tyto předpoklady splněny.

Odhad absolutního členu β_0 a směrnice β_1 regresní přímky a jejich statistické charakteristiky jsou uvedeny v další tabulce. Směrodatná chyba koeficientu je výběrová směrodatná odchylka odhadovaného parametru, tj. $\widehat{s}_{\beta_0} = S_{M1}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$ a $\widehat{s}_{\beta_1} = \frac{S_{M1}}{\sqrt{S_{XX}}}$ (Ve statistických programech je obvykle označována anglicky jako Standard Error.)

Statistické charakteristiky lineární regrese

Parametr	Koeficient	Směrodatná chyba koef.	t-statistika	p-hodnota
Absolutní člen β_0	180,658	128,381	1,407	0,1896823
Směrnice β_1	9,429	0,681	13,847	7.520972e-08

Z tabulky tedy dostáváme:

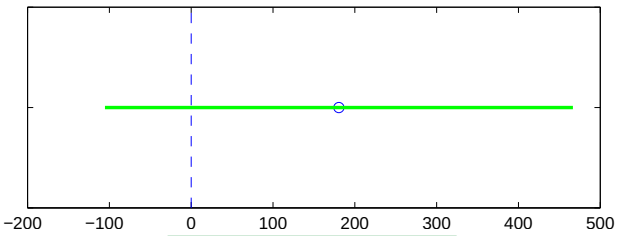
$$\text{pracovní doba } = 180{,}658 + 9{,}429 \cdot \text{velikost populace.}$$

To je třeba interpretovat jako odhad průměrné hodnoty počtu pracovních hodin pro populaci s danou velikostí. Očekáváme, že na každých dalších 1 000 lidí stoupne za měsíc počet pracovních hodin o 9,429, což je směrnice regresní přímky. Uvědomte si, že absolutní člen $(180,658)$ značí průměrný počet pracovních hodin, když je populace rovna nule. To zřejmě nedává smysl a mělo by nám to připomenout, že model by se měl používat pouze v tom rozmezí obou veličin, v němž se pohybovaly pozorované hodnoty. V tomto případě to znamená x od 26 do 370. Je ovšem pravda, že dosažená hladina významnosti pro absolutní člen je přibližně $0,19$, a nelze tedy říci, že by se absolutní člen β_0 významně lišil od nuly.

Připomeňme, že tyto výsledky jsme spočítali pro náhodný výběr 12 nemocnic. Kdybychom teď zvolili jiný náhodný výběr 12 nemocnic, dostali bychom odlišný odhad směrnice a absolutního členu. Určeme proto intervaly spolehlivosti neznámých parametrů β_0 a β_1.

Oboustranný interval spolehlivosti pro β_0

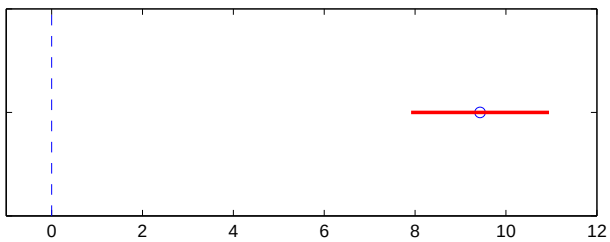
$180{,}6575 \pm 2{,}228 \cdot 128{,}3812 = 180{,}6575 \pm 286{,}051$



$(-105{,}394;\ 466{,}709)$

Oboustranný interval spolehlivosti pro β_1

$9{,}429 \pm 2{,}228 \cdot 0{,}681 = 9{,}429 \pm 1{,}517$



$(7{,}912;\ 10{,}946)$

Na základě výběru 12 nemocnic můžeme říci, že neznámý parametr β_0 leží mezi $-105,394$ a $466,709$ a neznámý parametr β_1, tj. parametr změny průměrného počtu pracovních hodin v závislosti na změně velikosti populace (v tisících), leží mezi $7,912$ a $10,946$ pracovními hodinami za měsíc.

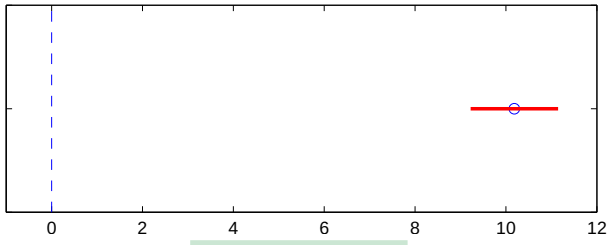
Protože interval spolehlivosti pro β_0 pokrývá nulu, nelze potvrdit, že se významně liší od nuly. Naproti tomu interval spolehlivosti pro β_1 nulu nepokrývá, tedy se významně liší od nuly, jinak řečeno počet pracovních hodin skutečně lineárně závisí na rozsahu spádové populace.

Pokud bychom uvažovali **regresi procházející počátkem** (plná čára) a výsledek srovnali s obecnou regresní přímkou (čárkovaná čára), dostaneme následující odhady

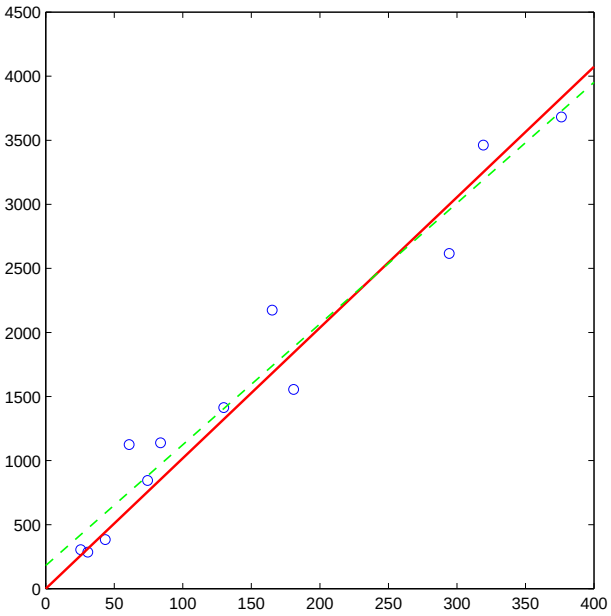
$\widehat{\beta}_1^* = 10,185$ $\widehat{s}_{\beta_1^*} = 0,4371,$
$t^* = 3,30157$, $p^* - hodnota = 1.0318 \times 10^{-10}$

Oboustranný interval spolehlivosti pro β_1^*

$$10{,}185 \pm 2{,}2 \cdot 0{,}4371 = 10{,}185 \pm 0{,}962$$



(9,223; 11,147)



Protože interval spolehlivosti pro β_1^* nulu nepokrývá, opět jsme prokázali, že se významně liší od nuly, tj. počet pracovních hodin skutečně lineárně závisí na rozsahu spádové populace.

$$\text{pracovní doba } = 10{,}185 \cdot \text{velikost populace.}$$

Úlohy k procvičení

Cvičení 5.1.

V lineárním regresním modelu $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$, $\boldsymbol{X} = \begin{pmatrix} 1 & -3 & 9 \\ 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix}$, $\boldsymbol{Y} = \begin{pmatrix} 7 \\ 4 \\ 2 \\ 2 \\ 5 \\ 8 \end{pmatrix}$ spočítejte metodou nejmenších čtverců odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2.

$[\hat{\boldsymbol{\beta}} = (1,5; 0,1786; 0,6786)'$, $\widehat{\boldsymbol{Y}} = (7,0714; 3,8571; 2; 2,3571; 4,5714; 8,1429)'$, $S_e = 0,3571$, $s^2 = 0,119.]$

Cvičení 5.2. Pro data

x	-2	-1	0	1	2
Y	0	2	3	3	1

spočítejte metodou nejmenších čtverců odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2 ve dvou modelech. Který model je vhodnější? (Proč?) Oba modely vykreslete.

(a) model s regresní funkcí $Y = \beta_0 + \beta_1 x + \beta_2 x^2$

(b) model s maticí plánu $\boldsymbol{X} = \begin{pmatrix} 1 & 4 \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 4 \end{pmatrix}$

[(a) $\hat{\boldsymbol{\beta}} = (3,09; 0,3; -0,64)'$, $\widehat{\boldsymbol{Y}} = (-0,086; 2,143; 3,086; 2,743; 1,114)'$, $S_e = 0,114$, $s^2 = 0,057$.
(b) $\hat{\boldsymbol{\beta}} = (3,17; -0,67)'$, $\widehat{\boldsymbol{Y}} = (0,5; 2,5; 0; 2,5; 0,5)'$, $S_e = 10$, $s^2 = 3,33$.]

Cvičení 5.3. Pomocí regresní přímky procházející počátkem spočítejte metodou nejmenších čtverců odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2 v LRM $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ pro data

x	10	20	30	40	50	60
Y	0,18	0,35	0,48	0,65	0,84	0,97

Jedná se o měření teplotní délkové roztažnosti měděné trubky. Rozdíl teploty od referenční 20 °C je x, prodloužení tyče je měřená veličina Y.

[$\hat{\boldsymbol{\beta}} = 0,0164$, $\widehat{\boldsymbol{Y}} = (0,164; 0,328; 0,493; 0,657; 0,821; 0,985)'$, $S_e = 0,0015$, $s^2 = 0,0003$.]

Cvičení 5.4. U 126 podniků řepařské oblasti v České Republice byl sledován hektarový výnos cukrovky ve vztahu ke spotřebě průmyslových hnojiv.
Data jsou uložena v souboru „`cukrovka.Rdata`“ ve 4 sloupcích:

(1) dolní hranice spotřeby K_2O (kg/ha)
(2) horní hranice spotřeby K_2O (kg/ha)
(3) četnosti
(4) průměrné výnosy cukrovky (q/ha)

a) odhadněte parametry regresní funkce tvaru

$$y = \beta_0 + \beta_1 x$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$y = \beta_0 + \beta_1 x^{0,5}$$

Poznámka: Za hodnoty nezávisle proměnné volte střed intervalu.

b) Porovnejte vhodnost tří použitých regresních modelů.

Cvičení 5.5. U 19 vzorků potravinářské pšenice byl zjišťován obsah zinku v zrnu (proměnná Y), v kořenech (proměnná X_1), v otrubách (proměnná X_2) a ve stonku a listech (proměnná X_3). Data jsou uložena v souboru „`psenice.Rdata`“.

a) Předpokládejte, že je vhodný regresní model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

Odhadněte regresní koeficienty a rozptyl, vypočtěte vektor predikce a index determinace. Proveďte celkový F-test a dílčí t-testy. Hladinu významnosti volte 0,05. Normalitu reziduí posuďte graficky pomocí funkce `qqnorm`.

b) Z regresního modelu odstraňte ty proměnné, jejichž regresní koeficienty se ukázaly nevýznamné pro $\alpha = 0,05$. Sestavte nový regresní model a proveďte v něm všechny úkoly z bodu a).

KAPITOLA 5

Ověřování předpokladů v klasickém modelu lineární regrese

Základní informace

(1) V následující kapitole se budeme zabývat ověřováním předpokladů lineárního regresního modelu. Nejprve se zaměříme na ověření normality dat, graficky i pomocí statistických testů. Dále pak popíšeme metody pro testování korelace v datech a konstrukci odhadů parametrů lineárního regresního modelu pro korelovaná data. Nakonec se budeme zabývat multikolinearitou v matici plánu. Nejprve bude popsáno, jak detekovat tento problém, dále pak jak se s ním vyrovnat.

(2) Předpokládá se znalost základních pojmů z teorie matematické statistiky a lineárních regresních modelů – číselné charakteristiky náhodných veličin a jejich výběrové odhady, odhady parametrů v lineárním regresním modelu, testování hypotéz

Výstupy z výukové jednotky

Studenti

- graficky ověří normální rozdělení dat
- aplikují statistické testy pro ověření normálního rozdělení
- detekují přítomnost autokorelace 1. řádu
- umí konstruovat regresní modely pro korelovaná data
- detekují multikolinearitu v matici plánu
- sestaví vhodný model pomocí postupné regrese

1. Motivace

Možnost modelování závislosti pomocí lineárního regresního modelu je podmíněna nějakými předpoklady o datech. Jedním z nich je předpoklad normality. Ten má zcela zásadní vliv na použití testů o parametrech lineárního regresního modelu, neboť většina testů předpokládá právě normalitu. Dalším důležitým předpokladem je nezávislost chyb, která bývá často (zejména v časových řadách) porušena. Také vlastnosti matice plánu, zejména pak prvky matice $\mathbf{X}'\mathbf{X}$ a existence její inverze, hrají velmi důležitou roli při odhadech parametrů modelu. Opomíjení výše zmíněných předpokladů může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

2. Ověřování normality dat

V této části se zaměříme obecně na ověření normality dat. To je pro nás užitečné zejména pro ověření normality residuí, neboť v předešlých kapitolách jsme do konstrukce lineárního regresního modelu zahrnuli předpoklad normality residuí, tj.

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n.$$

V dalším budeme předpokládat $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z nějakého rozdělení a budeme zjišťovat, zda tímto rozdělením je normální rozdělení.

2.1. Grafické posouzení. Jedním z prvních kroků, které bychom měli provést při posuzování normality dat, je vykreslení dat do nějakého obrázku. Z toho pak můžeme rozhodnout, jestli vůbec má smysl pouštět se do dalších analýz. Možností, jak graficky posoudit normalitu dat je mnoho. Uvedeme tři základní typy.

(1) **Histogram**
Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti normálního rozložení. Nejprve vytvoříme třídicí intervaly $\langle u_1, u_2\rangle, \ldots, \langle u_r, u_{r+1}\rangle$, doporučuje se volit r blízké $\sqrt{n}$. **Četnostní hustota** j-tého třídicího intervalu je definována vztahem

$$f_j = \frac{p_j}{d_j}$$

kde $d_j = u_{j+1} - u_j$. Soustava obdélníků sestrojených nad třídicími intervaly, jejichž plochy jsou rovny relativním četnostem, se nazývá **histogram**. Vykreslením histogramu a příslušné hustoty normálního rozdělení můžeme vizuálně posoudit, zda se data řídí normálním rozdělením. Nutno podotknout, že tato metoda je velmi citlivá na volbu třídících intervalů, zvláště pro menší počty dat.

(2) **Quantile - quantile plot (Q–Q plot)**
Q-Q plot konstruujeme tak, že na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na vodorovnou osu kvantily u_{α_j} normálního rozdělení, kde

$$\alpha_j = \frac{j - r_{adj}}{n + n_{adj}},$$

přičemž r_{adj} a n_{adj} jsou korigující faktory $\leq 0,5$. Implicitně se klade $r_{adj} = 0,375$ a $n_{adj} = 0,25$. Protože normální rozložení závisí na parametrech μ a σ^2, tyto parametry se většinou odhadují z dat. Body $(u_{\alpha_j}(X), x_{(j)})$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím lepší je soulad mezi empirickým a normálním rozdělením.

(3) **Graf výběrové distribuční funkce**
Položme

$$z_{(i)} = \frac{x_{(i)} - \bar{x}}{s}, \quad i = 1, \ldots, n, \; s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Na vodorovnou osu vykreslíme hodnoty $z_{(i)}$ a na svislou osu pak hodnoty distribuční funkce standardizovaného normálního rozdělení $\phi(z_{(i)})$, které porovnáme s hodnotami výběrové distribuční funkce $F_n(z_{(i)}) = \frac{i}{n}$, $i = 1, \ldots n$.

2.2. Kolmogorovův – Smirnovův test.

VĚTA 2.1. *Testujeme hypotézu, která tvrdí, že náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ pochází z rozložení s distribuční funkcí $\Phi(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika*

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)|.$$

Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem

$$\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}. \tag{15}$$

Poznámka 2.2. Nulová hypotéza musí specifikovat distribuční funkci zcela přesně, včetně všech jejích případných parametrů. Např. K – S test lze použít pro testování hypotézy, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $Rs(0, 1)$, což se využívá při testování generátorů náhodných čísel. Pokud však parametry distribuční funkce odhadujeme z výběru,

změní se rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

Příklad 2.3. Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K – S testu zjistěte na hladině významnosti 0,05, zda tato data pocházejí z normálního rozložení.

Řešení. Odhadem střední hodnoty je výběrový průměr $m = 11$, odhadem rozptylu je výběrový rozptyl $s^2 = 10$. Uspořádaný náhodný výběr je (8, 9, 10, 12, 16). Vypočteme hodnoty výběrové distribuční funkce:

(1) $x < 8 : F_5(x) = 0$
(2) $8 \leq x < 9 : F_5(x) = \frac{1}{5} = 0,2$
(3) $9 \leq x < 10 : F_5(x) = \frac{2}{5} = 0,4$
(4) $10 \leq x < 12 : F_5(x) = \frac{3}{5} = 0,6$
(5) $12 \leq x < 16 : F_5(x) = \frac{4}{5} = 0,8$
(6) $x \geq 16 : F_5(x) = 1$

Hodnoty teoretické distribuční funkce $\Phi_T(x)$ v bodech 8, 9, 10, 12, 16:

(1) $\Phi_T(8) = \Phi(\frac{8-11}{\sqrt{10}}) = \Phi(-0,95) = 1 - \Phi(0,95) = 1 - 0,82894 = 0,17106$
(2) $\Phi_T(9) = \Phi(\frac{9-11}{\sqrt{10}}) = \Phi(-0,63) = 1 - \Phi(0,63) = 1 - 0,73565 = 0,26435$
(3) $\Phi_T(10) = \Phi(\frac{10-11}{\sqrt{10}}) = \Phi(-0,32) = 1 - \Phi(0,32) = 1 - 0,62552 = 0,37448$
(4) $\Phi_T(12) = \Phi(\frac{12-11}{\sqrt{10}}) = \Phi(0,32) = 0,62552$
(5) $\Phi_T(16) = \Phi(\frac{16-11}{\sqrt{10}}) = \Phi(1,58) = 0,94295$
(6) (Φ je distribuční funkce rozložení $N(0,1)$.)

Rozdíly mezi výběrovou distribuční funkcí $F_5(x)$ a teoretickou distribuční funkcí $\Phi_T(x)$:

(1) $d_1 = 0,2 - 0,17106 = 0,02894$;
(2) $d_2 = 0,4 - 0,26435 = 0,13565$;
(3) $d_3 = 0,6 - 0,37448 = 0,22552$;
(4) $d_4 = 0,8 - 0,62552 = 0,17448$;
(5) $d_5 = 1 - 0,94295 = 0,05705$.

Testová statistika: $D_5 = 0,22552$, modifikovaná kritická hodnota pro $n = 5, \alpha = 0,05$ je $0,343$. Protože $0,22552 < 0,343$, hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

2.3. Shapirův – Wilkův test normality. Testujeme hypotézu, která tvrdí, že náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ pochází z rozložení $N(\mu, \sigma^2)$. Test je založen na zjištění, zda body v Q-Q plotu jsou významně odlišné od regresní přímky proložené těmito body. Shapirův[1] – Wilkův[2] test se používá především pro výběry menších rozsahů, $n < 50$.

2.4. Testy dobré shody.
I když uvádíme test dobré shody v souvislosti s testováním normality dat, uvedeme obecnou formulaci testů dobré shody, neboť mají široké využití v mnoha dalších partiích statistitky.

[1]Samuel Shapiro (1930 –). Americký statistik a epidemiolog. Profesor na univerzitě Johna Hopkinse.
[2]Martin Bradbury Wilk (1922 –). Kanadský matematik.

VĚTA 2.4. *Testujeme hypotézu, která tvrdí, že náhodný výběr* $\mathbf{X} = (X_1, \ldots, X_n)'$ *pochází z rozložení s distribuční funkcí* $\Phi(x)$.

(1) Je-li distribuční funkce spojitá, pak data rozdělíme do r *třídicích intervalů* $(u_j, u_{j+1}\rangle, j = 1, \ldots, r$. *Zjistíme absolutní četnost* n_j *j-tého třídicího intervalu a vypočteme pravděpodobnost* p_j, *že náhodná veličina* X *s distribuční funkcí* $\Phi(x)$ *se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak* $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.

(2) Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídicích intervalů použijeme varianty $x_{[j]}, j = 1, \ldots, r$. *Pro variantu* $x_{[j]}$ *zjistíme absolutní četnost* n_j *a vypočteme pravděpodobnost* p_j, *že náhodná veličina* X *s distribuční funkcí* $\Phi(x)$ *se bude realizovat variantou* $x_{[j]}$. *Platí-li nulová hypotéza, pak*

$$p_j = \Phi\left(x_{[j]}\right) - \lim_{x \to x_{[j]}^-} \Phi(x) = P\left(X = x_{[j]}\right). \tag{16}$$

Testová statistika:

$$K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}. \tag{17}$$

Platí-li nulová hypotéza, pak $K \approx \chi^2(r - 1 - p)$, *kde* p *je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení* $p = 2$, *protože z dat odhadujeme střední hodnotu a rozptyl.) Nulovou hypotézu zamítáme na asymptotické hladině významnosti* α, *když* $K \geq \chi^2_{1-\alpha}(r-1-p)$. *Aproximace se považuje za vyhovující, když* $np_j \geq 5, j = 1, \ldots, r$.

Poznámka 2.5. Hodnota testové statistiky K je silně závislá na volbě třídicích intervalů. Navíc při nesplnění podmínky $np_j \geq 5, j = 1, \ldots, r$ je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace.

Příklad 2.6. Byl zjišťován počet poruch určitého zařízení za 100 hodin provozu ve 150 disjunktních 100 h intervalech. Výsledky měření:

Počet poruch za 100 h provozu	0	1	2	3	4 a víc
Absolutní četnost	52	48	36	10	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že náhodný výběr $X_1, \ldots, X_{150}$ pochází z rozložení $Po(1, 2)$.

Řešení. Pravděpodobnost, že náhodná veličina s rozložením $Po(\lambda)$, kde $\lambda = 1, 2$ bude nabývat hodnot $p_0, \ldots, p_4$ a víc je $p_j = \frac{\lambda^j}{j!}e^{-\lambda} = \frac{1,2^j}{j!}e^{-1,2}, j = 0, 1, 2, 3$, $p_4 = 1 - (p_0 + p_1 + p_2 + p_3)$.
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
0	52	0,301	$150 \cdot 0,301 = 45,15$	1,039
1	48	0,361	$150 \cdot 0,361 = 54,15$	0,698
2	36	0,217	$150 \cdot 0,217 = 32,55$	0,366
3	10	0,087	$150 \cdot 0,087 = 13,05$	0,713
4	4	0,034	$150 \cdot 0,034 = 5,1$	0,237

$K = 1,039 + 0,698 + 0,366 + 0,713 + 0,237 = 3,053, r = 5, \chi^2_{0,95}(4) = 9,488$. Protože $3,053 < 9,488$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Poznámka 2.7. Test dobré shody může být použit i v těch případech, kdy rozložení, z něhož daný náhodný výběr pochází, neodpovídá nějakému známému rozložení (např. exponenciálnímu, normálnímu, Poissonovu, ...), ale je určeno intuitivně nebo na základě zkušenosti.

Příklad 2.8. Ve svých pokusech pozoroval J.G. Mendel 10 rostlin hrachu a na každé z nich počet žlutých a zelených semen. Výsledky pokusu:

č.rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých	25	32	14	70	24	20	32	44	50	44
počet zelených	11	7	5	27	13	6	13	9	14	18
celkem	36	39	19	97	37	26	45	53	64	62

Z genetických modelů vyplývá, že pravděpodobnost výskytu žlutého semene by měla být 0,75 a zeleného 0,25. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že výsledky Mendelových pokusů se shodují s modelem.

Řešení. Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j-np_j)^2}{np_j}$
1	25	0,75	$36 \cdot 0,75 = 27$	0,148148
2	32	0,75	$39 \cdot 0,75 = 29,25$	0,258547
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	44	0,75	$62 \cdot 0,75 = 46,5$	0,134409

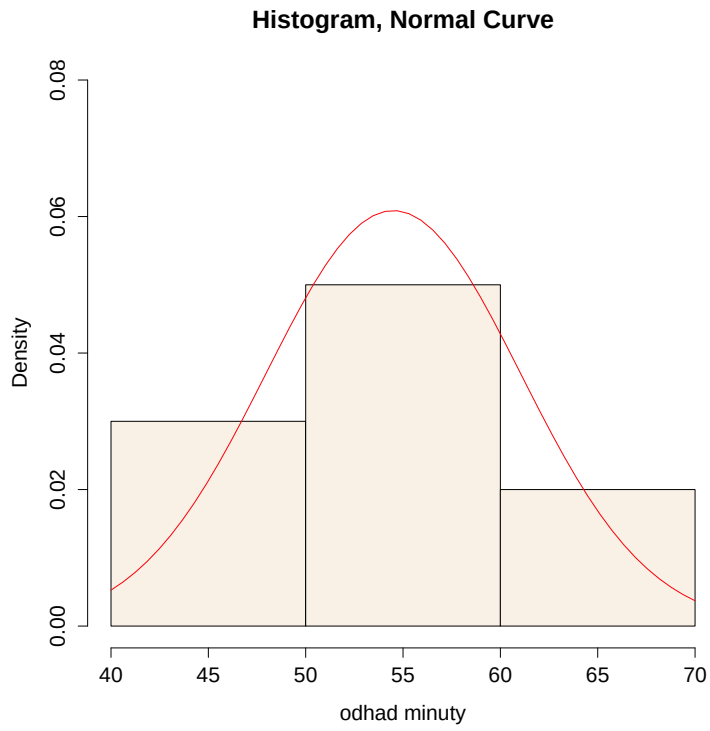
$K = 0,148148 + 0,258547 + \ldots + 0,134409 = 1,797495, r = 10, \chi^2_{0,95}(9) = 16,9$. Protože $1,797495 < 16,9$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Příklad 2.9. Deset pokusných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne jedna minuta. Výsledky pokusu jsou uvedeny v následující tabulce

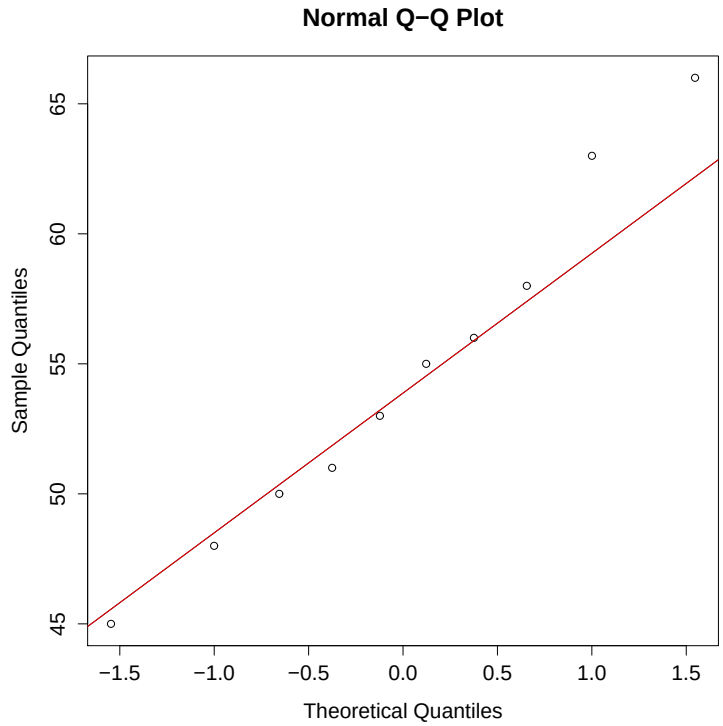
č. osoby	1	2	3	4	5	6	7	8	9	10
odhad minuty	53	48	45	55	63	51	66	56	50	58

Testujte graficky i výpočtem, zda se jedná o výběr z normálního rozdělení.

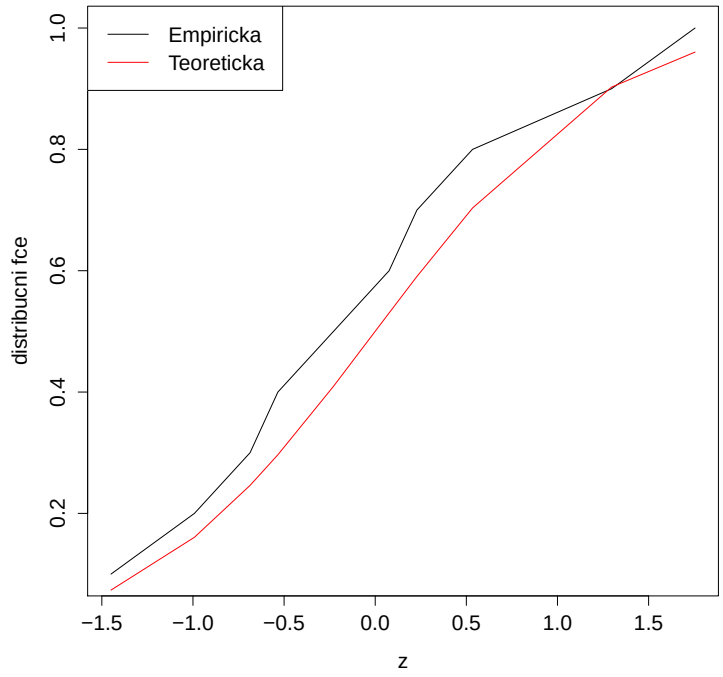
Řešení. Nejprve data rozdělíme do 3 třídících intervalů a vykreslíme histogram a křivku normální hustoty.



Z obrázku je vidět, že by data mohla pocházet z normálního rozdělení. Vykreslíme také Q–Q plot.



I z tohoto obrázku to vypadá na normální rozdělení. Nakonec ještě porovnáme graf výběrové distribuční funkce s grafem distribuční funkce normálního rozdělení.



Také tento obrázek ukazuje na normalitu dat. Závěrem ještě provedeme výše uvedené testy. Všechny přijímají hypotézu o tom, že data pocházejí z normálního rozdělení. Jejich p-hodnoty jsou shrnuty v následující tabulce:

Test	*p-hodnota*
Kolmogorovův – Smirnovův test	$0,9985$
Shapirův – Wilkův test	$0,9164$
Test dobré shody	$0,9189$

3. Autokorelace

V některých případech (často v časových řadách) hodnoty náhodné chyby ε_i závisí na předchozích hodnotách ε_{i-k}, $k = 1, 2, \ldots$, což má za následek, že efekt náhodných chyb není okamžitý, ale je pociťován i v budoucnosti. Tento případ se nazývá **autokorelace**.

Existují různé formy autokorelace, ty pak mají vliv na tvar varianční matice $D\boldsymbol{\varepsilon}$. Omezme se na nejjednodušší a zároveň nejpropracovanější formu autokorelace, která se nazývá **autoregrese 1. řádu**, značíme $AR(1)$.
Předpokládejme tedy, že platí

$$\varepsilon_i = \theta\varepsilon_{i-1} + u_i,$$

kde θ je neznámý parametr, $|\theta| < 1$, $Eu_i = 0$, $E(u_i) = 0$, $cov(u_i, u_j) = \begin{cases} \sigma_u^2 & i = j, \\ 0 & \text{jinak.} \end{cases}$

Vypočtěme postupně

$$\varepsilon_i = \theta\varepsilon_{i-1} + u_i = \theta(\theta\varepsilon_{i-2} + u_{i-1}) + u_i = \theta^2\varepsilon_{i-2} + \theta u_{i-1} + u_i = \cdots = \sum_{j=0}^{\infty} \theta^j u_{i-j}$$

$$E\varepsilon_i = E\sum_{j=0}^{\infty} \theta^j u_{i-j} = \sum_{j=0}^{\infty} \theta^j Eu_{i-j} = 0$$

$$D\varepsilon_i = D\sum_{j=0}^{\infty} \theta^j u_{i-j} = \sum_{j=0}^{\infty} \theta^{2j} Du_{i-j} = \sigma_u^2 \sum_{j=0}^{\infty} \theta^{2j} = \tfrac{\sigma_u^2}{1-\theta^2}$$

$$cov(\varepsilon_i, \varepsilon_{i-j}) = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \theta^r\theta^s cov(u_{i-r}, u_{i-j-s}) = \theta^j\sigma_u^2 \sum_{r=0}^{\infty} \theta^{2r} = \tfrac{\theta^j\sigma_u^2}{1-\theta^2} \text{ pro } j > 0$$

Vektor náhodných chyb již nemá varianční matici diagonální se stejným rozptylem, tj. $D\boldsymbol{\varepsilon} \neq \sigma_2 \mathbf{I}_n$, ale

$$D\boldsymbol{\varepsilon} = \frac{\sigma_u^2}{1-\theta^2} \begin{pmatrix} 1 & \theta & \theta^2 & \dots & \theta^{n-1} \\ \theta & 1 & \theta & \dots & \theta^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \theta \\ \theta^{n-1} & \dots & \theta^2 & \theta & 1 \end{pmatrix} = \underbrace{\frac{\sigma_u^2}{1-\theta^2}}_{\sigma_\varepsilon^2} \mathbf{W}.$$

Náš model je tedy tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ E\boldsymbol{\varepsilon} = \mathbf{0},\ D\boldsymbol{\varepsilon} = \sigma_\varepsilon^2 \mathbf{W},\quad \text{píšeme}\ \ \mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{W}).$$

Vidíme, že uvažovaný model nesplňuje předpoklad homogenity rozptylu náhodných chyb, proto přejdeme k užitečnému zobecnění lineárního regresního modelu.

Uvažujme lineární regresní model s obecnější varianční maticí

$$\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{W}^{-1}), \qquad \mathbf{W} > 0,\ \sigma^2 > 0,\ h(\mathbf{X}) = k.$$

Také v tomto případě jsou $\boldsymbol{\beta}$ a σ^2 neznámé parametry a matice $\mathbf{W}$ je (zpravidla známá) pozitivně definitní matice. Následující věta ukazuje, jakým způsobem lze provést odhad neznámých parametrů v tomto obecnějším případě.

VĚTA 3.1. *(Aitkenův odhad). Mějme regresní model* $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{W}^{-1})$ *plné hodnosti, kde* $\mathbf{W} > 0$. *Pak odhad pomocí metody nejmenších čtverců je roven*

$$\widehat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}.$$

Poznámka 3.2. V případě, že matice $\mathbf{W}$ je **diagonální**, mluvíme o **vážené regresi** a metodě nejmenším čtverců, pomocí které byly provedeny odhady, se v tomto případě říká **vážená metoda nejmenších čtverců**.
Příkladem takového modelu je situace, kdy i-tá složka vektoru $\mathbf{Y}$ je průměrem n_i nezávislých pozorování se stejnou střední hodnotou a stejným rozptylem σ^2. Potom

$$DY_i = \frac{\sigma^2}{n_i}$$

a regresní model je tvaru

$$\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{W}^{-1}),$$

kde

$$\mathbf{W} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n_n \end{pmatrix}.$$

3.1. Detekce autokorelace. Popišme nejprve, jak poznat přítomnost autokorelace v datech.

1) Graficky
   Označme $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$. Do grafu postupně vykreslíme hodnoty $\hat{\varepsilon}_i$ v závislosti na $\hat{\varepsilon}_{i-1}$, $i = 2, \dots, n$. Bude-li z grafu zřejmá přibližná lineární závislost, svědčí to o autokorelaci 1. řádu nebo o špatné volbě modelu.
2) Test hypotézy $H_0 : \theta = 0$ proti $H_1 : \theta \neq 0$

(a) Pokud pracujeme s dostatečně velkým souborem pozorování ($n \geq 30$), přichází v úvahu **asymptotický test** vycházející z přibližné normality $\hat{\theta}$ se střední hodnotou $\hat{\theta} = \theta$ a rozptylem $D\hat{\theta} = \frac{1-\theta^2}{n}$. V takovém případě platí

$$U_{\hat{\theta}} = \frac{\hat{\theta} - \theta}{\sqrt{\frac{1-\theta^2}{n}}} \overset{A}{\sim} N(0,1).$$

Za platnosti hypotézy má tedy statistika

$$\sqrt{n}\hat{\theta} \overset{A}{\sim} N(0,1).$$

Pak nulovou hypotézu zamítáme, pokud $|\sqrt{n}\hat{\theta}| > u_{1-\frac{\alpha}{2}}$, kde $u_{1-\frac{\alpha}{2}}$ je kvantil standardizovaného normálního rozdělení.

(b) Propracovanější je **Durbinův – Watsonův test** založený na statistice

$$D = \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

Pokud budou residua málo korelovaná, hodnota D se bude pohybovat kolem 2. Kladná hodnota způsobí, že $D \in (0,2)$ a záporná korelace způsobí, že $D \in (2,4)$. Přesné hodnoty kritických oborů pro test nalezneme v tabulkách.
Durbinův – Watsonův test se používá zejména u dat, jejichž jednotlivá pozorování byla pořízena postupně v pravidelných časových odstupech. Statisticky významná hodnota D však může svědčit také o nesprávně zvoleném tvaru regresní funkce.

3.2. Odhad parametru θ. Pokud je známý parametr autokorelace θ, není problém spočítat odhad vektoru neznámých parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců. V praktických úlohách však parametr autokorelace θ neznáme a je třeba najít jeho vhodný odhad. Odhady parametru autokorelace θ lze zkonstruovat různým způsobem. Uveďme dva postupy:

1) Jednou z možností je odhadovat ho jako regresní koeficient v modelu

$$\hat{\varepsilon}_i = \theta\hat{\varepsilon}_{i-1} + u_i, \ i = 2, \ldots, n$$

metodou nejmenších čtverců. Odtud pak dostáváme

$$\hat{\theta} = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i\hat{\varepsilon}_{i-1}}{\sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2}.$$

Za určitých předpokladů o limitním chování vysvětlujících proměnných je $\hat{\theta}$ konzistentním odhadem parametru θ.

2) Další možností je využít výše popsané Durbinovy – Watsonovy statistiky D. Odhad je pak tvaru

$$\hat{\theta} = 1 - \frac{D}{2}.$$

Tento odhad je rovněž konzistentní a pro větší n jsou jen malé rozdíly mezi oběma odhady.

Ani v případě prvního či druhého odhadu parametru θ nejsou již odhady $\hat{\boldsymbol{\beta}}$ nejlepší nezkreslené. Konečné vlastnosti těchto odhadů jsou jen obtížně určitelné. Proto se většinou spokojíme s konstatováním, že asymptotické vlastnosti odhadů jsou dobré a předpokládáme, že jsou rozumnou aproximací i v konečných výběrech.

3.3. Odstranění autokorelace 1. řádu. V některých případech nás nezajímají pouze nezkreslené odhady $\hat{\boldsymbol{\beta}}$, ale chceme odstranit autokorelaci z dat nějakou vhodnou transformací. To je samozřejmě možné, je třeba si však uvědomit, že pak dostáváme úplně jiný model a následná interpretace výsledků je obtížně proveditelná.
Uveďme postup pro odstranění autokorelace 1. řádu:

(1) Jednou z výše uvedených metod nalezneme odhad $\hat{\theta}$
(2) Vytvoříme nový model

$$Y_i^* = Y_{i+1} - \hat{\theta} Y_i;\ X_{ij}^* = X_{i+1,j} - \hat{\theta} X_{ij},\ i = 1, \ldots, n-1,\ j = 1, \ldots, k,$$

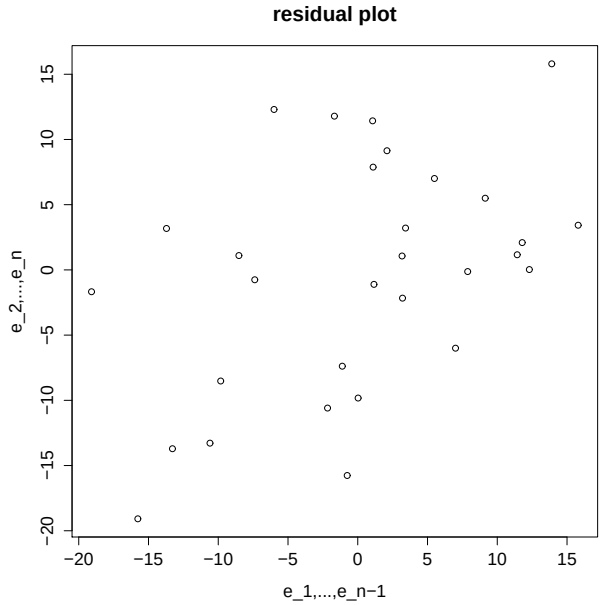
tj. vznikne model

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,\ E\boldsymbol{\varepsilon}^* = \mathbf{0},\ D\boldsymbol{\varepsilon}^* = \sigma_{\varepsilon^*}^2 \mathbf{I}_n,$$

ve kterém již není přítomna autokorelace 1. řádu.
(3) Hledáme odhady $\hat{\boldsymbol{\beta}}^*$ standardním způsobem.

Příklad 3.3. V letech 1953 – 1983 byly měřeny ztráty vody při distribuci do domácností. Výsledky měření jsou uloženy v souboru „`voda.RData`“. Proměnná x označuje množství vyrobené vody, proměnná Y ztrátu. Ověřte, zda se v datech vyskytuje autokorelace 1. řádu a případně ji odstraňte.

Řešení. Nejprve do grafu postupně vykreslíme hodnoty $\hat{\varepsilon}_i$ v závislosti na $\hat{\varepsilon}_{i-1}$, $i = 2, \ldots, 31$.



Z grafu je patrná lineární závislost a tudíž přítomnost autokorelace 1. řádu. Provedeme také oba testy na hladině významnosti $\alpha = 0,05$:

(a) Pro asymptotický test vychází hodnota testové statistiky

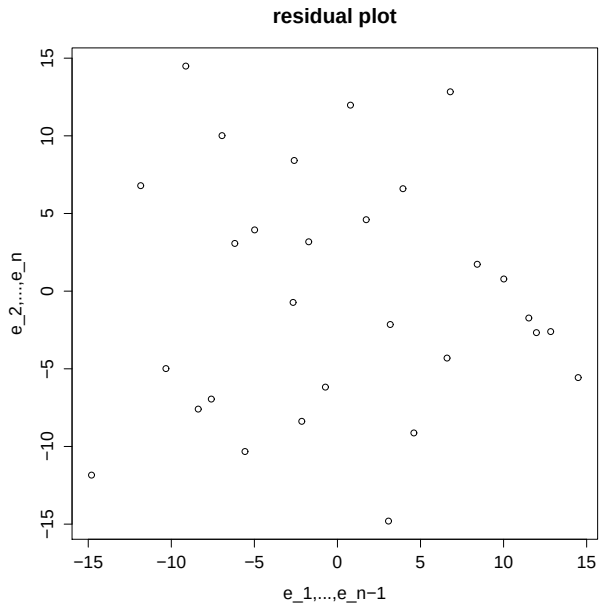
$$U_{\hat{\theta}} = |\sqrt{n}\hat{\theta}| = 2,339.$$

Nulovou hypotézu tedy **zamítáme**, neboť $|\sqrt{n}\hat{\theta}| > u_{1-\frac{\alpha}{2}} = 1,96$.
(b) Pro Durbinův – Watsonův test máme

$$D = \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} = 1,082$$

a p-hodnota testu je $0,0016$, takže také **zamítáme** nulovou hypotézu.

Nyní se pokusíme vhodnou transformací autokorelaci odstranit. Nejprve je třeba odhadnout parametr θ. Použijeme k tomu obě zmíněné metody. Odhady $\hat{\theta}$ jsou velmi podobné. Metodou nejmenších čtverců dostáváme odhad $\hat{\theta} = 0,421$. Odhad pomocí Durbin – Watsonovy statistiky vychází $\hat{\theta} = 0,459$. Pomocí tohoto odhadu vytvoříme nový model a v něm vykreslíme residua.



Z obrázku je patrná nezávislost residuí. Také Durbinův – Watsonův test již **nezamítá** nulovou hypotézu (jeho p-hodnota vychází $0,4$).

4. Multikolinearita

Multikolinearitou se rozumí vzájemná lineární závislost vysvětlujících proměnných. Přesnou multikolinearitou se rozumí případ, kdy jednotlivé sloupce $\mathbf{x}_j,\ j = 1, \ldots, k$ matice plánu $\mathbf{X}$ jsou lineárně závislé, takže pro aspoň jednu nenulovou konstantu c_j platí

$$c_1\mathbf{x}_1 + \cdots + c_k\mathbf{x}_k = \mathbf{0}.$$

V praxi bychom se s tímto případem neměli setkávat, neboť při rozumně sestaveném regresním modelu využijeme lineární kombinaci a zmenšíme počet vysvětlujících proměnných. Podobně nereálný je v praxi případ ortogonálních vysvětlujících proměnných, kdy matice $\mathbf{X}$ je ortogonální a platí, že $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$.
V praxi se tedy **multikolinearitou** rozumí případ, kdy **přibližně** platí rovnice vyjadřující lineární kombinaci vysvětlujících proměnných. V případě silné multikolinearity je determinant informační matice $\mathbf{X}'\mathbf{X}$ blízký nule, nejmenší vlastní číslo je rovněž blízké nule a matice $\mathbf{X}'\mathbf{X}$ je „skoro singulární“. O multikolinearitě svědčí i vysoké hodnoty poměru největšího a nejmenšího vlastního čísla.
Důvody multikolinearity mohou být různé:

- Multikolinearitu způsobuje regresní rovnice obsahující **nadbytečné** vysvětlující proměnné. Statistickými technikami můžeme přebytečné proměnné identifikovat a vyloučit z regresní rovnice.
- Multikolinearitu jen ztěží odstraníme v úlohách, kdy vzájemná spřaženost hodnot vysvětlujících proměnných je způsobena neuvažovanými veličinami nebo **formou statistického zjišťování**. Jde-li např. o údaje z časových řad, je podobný vývoj sledovaných veličin dostatečným důvodem vzniku multikolinearity. Vzhledem k tomu, že multikolinearitu hodnotíme výhradně na základě určitého souboru pozorování, stačí nesprávný výběr kombinací hodnot vysvětlujících proměnných, nereprezentujících obor možných hodnot, k existenci významné multikolinearity.

- Závažným důvodem multikolinearity je **skutečný vztah** vysvětlujících proměnných v rámci sledovaného jevu, procesu nebo systému. V tomto případě je třeba využít všechny informace nevýběrového charakteru k zlepšení kvality regresních odhadů.

4.1. Důsledky multikolinearity. V případě přesné multikolinearity je matice $\mathbf{X}'\mathbf{X}$ singulární a běžnou inverzí nepořídíme odhad neznámých parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců. Pro přibližnou multikolinearitu jsme sice schopni matici $\mathbf{X}'\mathbf{X}$ invertovat, ale kvalita pořízených odhadů je poměrně nízká.
Snížení kvality se projeví

- v kovarianční matici $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- v přesnosti prováděných výpočtů

neboť důsledkem vysokých rozptylů odhadů jsou příliš široké intervaly spolehlivosti, a tedy malá přesnost odhadu.

Logickým důsledkem multikolinearity je obtížné vyjádření individuálního vlivu jednotlivých vysvětlujících proměnných. Projeví se to nízkými hodnotami testových kritérií v t-testech nedovolujícími potvrdit závažnost jednotlivých regresorů v regresní funkci. Závažným důsledkem je značná výpočetní nespolehlivost a nestabilní hodnoty regresních odhadů. Stačí malý zásah do statistických údajů a výsledné odhady jsou odlišné.

Definice 4.1. Diagonální prvky matice $(\mathbf{X}'\mathbf{X})^{-1}$, tj.

$$\mathbf{a} = diag(\mathbf{X}'\mathbf{X})^{-1}$$

se označují jako **VIF – variance inflarion factors**.

Věta 4.2. *Variance inflarion factors úzce souvisí s **vícenásobnými korelačními koeficienty**, vyjadřující vztah j-té vysvětlující proměnné a lineární funkce ostatních vysvětlujících proměnných. Lze je zapsat jako*

$$a_j = \frac{1}{(1 - r_j^2)\mathbf{x}_j'\mathbf{x}_j},$$

kde $r_j = r_{x_j \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_k}$ je koeficient mnohonásobné korelace.

Vysoký stupeň multikolinearity se projevuje vysokými hodnotami korelačních koeficientů r_j, ale i vysokými hodnotami některých jednoduchých korelačních koeficientů.

4.2. Detekce multikolinearity. Jak bylo naznačeno výše, multikolinearita souvisí s korelačními koeficienty. Při detekci přítomnosti multikolinearity se tohoto faktu využívá. V praxi tedy testujeme hypotézu $H_0 : \mathbf{R} = \mathbf{I}_k$ proti $H_1 : \mathbf{R} \neq \mathbf{I}_k$, kde $\mathbf{R}$ je korelační matice proměnných.

Věta 4.3. *Platí-li nulová hypotéza, pak*

$$K = -\left[n - 1 - \frac{1}{6}(2k + 7)\right] \ln |\mathbf{R}| \sim \chi^2\left(\frac{k(k-1)}{2}\right).$$

Hypotézu H_0 tedy na hladině významnosti α zamítáme, pokud $K > \chi^2_{1-\alpha}\left(\frac{k(k-1)}{2}\right)$.

Věta 4.4 (Identifikace proměnných způsobujících multikolinearitu). *Pro identifikaci proměnných způsobujících multikolinearitu se používají statistiky*

$$F_j = \frac{n-k}{k-1}(d_{jj} - 1),$$

kde d_{jj} jsou diagonální prvky matice $\mathbf{D} = \mathbf{R}^{-1}$. V případě, že proměnná X_j nezpůsobuje multikolinearitu, má veličina F_j Fisherovo – Snedecorovo rozdělení $F(k-1, n-k)$.

4.3. Odstranění multikolinearity. Do modelu zařadíme jen ty regresory, které významně přispívají ke zlepšení kvality odhadu $\boldsymbol{\beta}$. K výběru nejlepší podmnožiny regresorů použijeme **metodu postupné regrese**.

Návod 4.5. Algoritmus pro metodu postupné regrese:

(1) Spočteme korelační matici $\mathbf{R}$ a provedeme test hypotézy $H_0 : \mathbf{R} = \mathbf{I}_k$. Je-li korelace prokázána, pokračujeme dalším krokem.
(2) Spočteme korelační koeficienty $r_{Y,X_1}, \ldots, r_{Y,X_k}$ a vybereme ten regresor X_i, jehož r_{Y,X_i} je v absolutní hodnotě největší.
(3) Sestavíme model $Y = \beta_0 + \beta_1 X_i$ a odhadneme jeho parametry. Vypočteme hodnotu statistiky $F = \frac{(n-2)s^2_{\hat{Y}}}{s^2} = \frac{(n-2)ID}{1-ID}$, kde ID značí index determinace. Pokud $F > F_{1-\alpha}(1, n-2)$, ponecháme regresor X_i v modelu.
(4) Spočteme parciální korelační koeficienty $r_{Y,X_1 \cdot X_i}, \ldots, r_{Y,X_{i-1} \cdot X_i}, r_{Y,X_{i+1} \cdot X_i}, \ldots, r_{Y,X_k \cdot X_i}$ a vybereme ten regresor X_j, jehož $r_{Y,X_j \cdot X_i}$ je v absolutní hodnotě největší.
(5) Sestavíme model $Y = \beta_0 + \beta_1 X_i + \beta_2 X_j$ a odhadneme jeho parametry. Vypočteme hodnotu statistiky $F = \frac{(n-3)\Delta ID}{1-ID}$, kde ΔID je přírůstek indexu determinace při zařazení X_j do modelu a ID je index determinace pro model $Y = \beta_0 + \beta_1 X_i + \beta_2 X_j$. Pokud $F > F_{1-\alpha}(2, n-3)$, ponecháme regresor X_j v modelu.
(6) Spočteme parciální korelační koeficienty $r_{Y,X_1 \cdot (X_i,X_j)}, \ldots, r_{Y,X_k \cdot (X_i,X_j)}$ a vybereme ten regresor X_l, jehož $r_{Y,X_l \cdot (X_i,X_j)}$ je v absolutní hodnotě největší a tak pokračujeme dále.

4.4. Zlepšování podmíněnosti matice $\mathbf{X}'\mathbf{X}$. Často se v praxi stává, že matice $\mathbf{X}'\mathbf{X}$ je špatně podmíněná a přesto nemusí být přítomna multikolinearita v modelu. Může to být například způsobeno příliš rozdílnými hodnotami kovariátů. Uveďme některé obecné principy na zlepšení podmíněnosti matice $\mathbf{X}'\mathbf{X}$.

- **Model standardizovaných proměnných**. Místo původních proměnných y_i a x_{ij} pracujeme s proměnnými ve tvaru

$$q_i = \frac{y_i - \bar{y}}{s_y}, \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}},$$

kde s_y a s_{x_j} jsou směrodatné odchylky jednotlivých proměnných. Standardizací vysvětlujících proměnných dostáváme při použití metody nejmenších čtverců místo matice $\mathbf{X}'\mathbf{X}$ korelační matici $\mathbf{R} = \mathbf{Z}'\mathbf{Z}/n$. Vektor $\mathbf{Z}'\mathbf{q}/n$ obsahuje jednoduché korelační koeficienty r_{yx_j}. Standardizací proměnných se zmenšují zaokrouhlovací chyby a zlepšují se možnosti hodnocení individuálního vlivu proměnných pomocí regresních parametrů.

- **Model v kanonickém tvaru**. Místo modelu ve tvaru

$$\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pracujeme s modelem

$$\mathbf{Y} = \mathbf{U}'\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

kde matice $\mathbf{U} = \mathbf{XV}$, vektor $\boldsymbol{\gamma} = \mathbf{V}'\boldsymbol{\beta}$ a $\mathbf{V}$ je matice standardizovaných vlastních vektorů odpovídajících vlastním číslům matice $\mathbf{X}'\mathbf{X}$. Odhady parametrů v kanonickém tvaru:

$$\hat{\boldsymbol{\gamma}} = \mathbf{L}^{-1}\mathbf{U}'\mathbf{Y},$$

kde $\mathbf{L}$ je diagonální matice s vlastními čísly matice $\mathbf{X}'\mathbf{X}$. Kovarianční matice odhadů $var(\hat{\boldsymbol{\gamma}}) = \sigma^2\mathbf{L}^{-1}$ ukazuje, že i v tomto případě jsou odhady nezávislé. Residuální součet čtverců se transformací nemění.

- **Hřebenová regrese** (ridge regression) – nebudeme podrobně popisovat tuto metodu, více informací lze najít např. v [**8**]. Pro praktickou aplikaci této metody lze v jazyce *R* použít proceduru `lm.ridge` z balíku `MASS`.

Příklad 4.6. V souboru „`vydaje.Rdata`“ jsou uložena data o 20 náhodně vybraných domácnostech. Sloupce proměnné „`domacnosti`“ obsahují postupně tyto údaje: výdaje za potraviny a nápoje (Y), počet členů domácnosti (X_1), počet dětí (X_2), průměrný věk výdělečně činných (X_3) a příjem domácnosti (X_4). Metodou postupné regrese zkonstruujte model s nejlepší podmíněností regresorů.

Řešení. Uvažujme nejdřív model se všemi regresory. Spočtěme nejprve pro ilustraci determinant $det((\mathbf{X}'\mathbf{X})^{-1}) = 4,65 \times 10^{-16}$. Také hodnoty VIF jsou pro první dva regresory vysoké:

`clenu`	`deti`	`vek`	`prijem`
$21,23$	$16,18$	$1,31$	$3,4$

Testujeme-li hypotézu $H_0 : \mathbf{R} = \mathbf{I}_4$, hodnota testové statistiky

$$K = -\left[19 - \frac{15}{6}\right] \ln|\mathbf{R}| = 64,94$$

výrazně převyšuje kritickou hodnotu $\chi^2_{0,95}(6) = 12,59$. Hypotézu H_0 tedy na hladině významnosti $0,05$ zamítáme.

Pro identifikaci proměnných způsobujících multikolinearitu můžeme spočítat dílčí statistiky F_j

`clenu`	`deti`	`vek`	`prijem`
$107,88$	$80,94$	$1,66$	$12,83$

které porovnáme s kritickou hodnotou $F_{0,95}(3,16) = 3,24$.
Metodou postupné regrese sestavíme model:

(1) Spočteme korelační koeficienty $r_{Y,X_1}, \ldots, r_{Y,X_4} = (0,77; 0,67; 0,18; 0,73)$. Vybereme regresor X_1, neboť jeho korelace je v absolutní hodnotě největší.
(2) Sestavíme model $Y = \beta_0 + \beta_1 X_1$. Vypočteme hodnotu statistiky $F = \frac{(n-2)ID}{1-ID} = \frac{18 \cdot 0,5897}{1-0,5897} = 25,87$. Tato hodnota je větší než $F_{0,95}(1,18) = 4,41$, takže regresor X_1 ponecháme v modelu.
(3) Spočteme parciální korelační koeficienty $r_{Y,X_2 \cdot X_1}, r_{Y,X_3 \cdot X_1}, r_{Y,X_4 \cdot X_1} = (0,36; 0,499; 0,32)$. Vybereme regresor X_3, jehož parciální korelační koeficient je v absolutní hodnotě největší.
(4) Sestavíme model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3$. Vypočteme hodnotu statistiky $F = \frac{(n-3)\Delta ID}{1-ID} = \frac{17 \cdot 0,102}{1-0,69} = 5,64$. Tato hodnota je větší než $F_{0,95}(2,17) = 3,59$, tedy ponecháme regresor X_3 v modelu.
(5) Spočteme parciální korelační koeficienty $r_{Y,X_2 \cdot (X_1,X_3)}, r_{Y,X_4 \cdot (X_1,X_3)} = (0,19; 0,17)$. Vybereme regresor X_2, jehož parciální korelační koeficient je v absolutní hodnotě největší.

(6) Sestavíme model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_2$. Vypočteme hodnotu statistiky $F = \frac{(n-4)\Delta ID}{1-ID} = \frac{16 \cdot 0{,}012}{1-0{,}704} = 0,63$. Tato hodnota je menší než $F_{0,95}(3,16) = 3,24$, a tedy regresor X_2 již nezahrneme do modelu.

Výsledný model je tedy tvaru

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3.$$

Úlohy k procvičení

Cvičení 4.1. V souboru „`studenti.RData`“ jsou uloženy údaje o 96 studentech VŠE v Praze. Hodnoty v prvním sloupci značí hmotnost studentů v kg (proměnná Y), ve druhém sloupci je výška studentů v cm (proměnná X_1) a ve třetím sloupci je indikátor pohlaví studenta (proměnná X_2, 0 – žena, 1 – muž). Předpokládejte regresní model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Odhadněte parametry modelu a ověřte normalitu residuí. Dále pak testujte přítomnost autokorelace 1. řádu, případně ji odstraňte.

[Odhady parametrů: $\hat{\beta}_0 = -53,67,\ \hat{\beta}_1 = 0,6648,\ \hat{\beta}_2 = 6,3323$, normalita se nezamítá, autokorelace 1. řádu se zamítá.]

Cvičení 4.2. V proměnné „`LakeHuron`“[3] jsou uloženy roční údaje o hloubce jezera Huron (ve stopách) v letech 1875 – 1972. Nalezněte vhodný regresní model a ověřte, zda se v datech vyskytuje autokorelace 1. řádu. Případně se ji pokuste odstranit. Zkoumejte také normalitu residuí.

[Vhodný model: polynom 7. stupně, autokorelace 1. řádu se nezamítá, normalita residuí u nového modelu se nezamítá.]

Cvičení 4.3. V souboru „`cement.RData`“ jsou uloženy údaje, které se týkají chemického složení portlandského cementu:

y	množství tepla v kaloriích na gram cementu
x_1	Tricalcium aluminate 3CaO.Al2O3 v %
x_2	Tricalciam silicate 3CaO.SiO2 v %
x_3	Tetracalcium alumino ferrite 4CaO.Al2O3.Fe2O3 v %
x_4	Dicalcium silicate 2CaO.SiO2 v %

Testujte multikolinearitu v daném modelu. Metodou postupné regrese nalezněte vhodný model. Poté ověřte normalitu residuí.

[Multikolinearita se nezamítá, vhodný model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4$, normalita residuí se nezamítá.]

Cvičení 4.4. V proměnné „`mtcars`“[3] jsou uložena data pro modelování závislosti spotřeby paliva osobních automobilů (proměnná `mpg`, počet mil/galon) na vlastnostech motoru, které jsou popsány následujícími proměnnými:

[3] datový soubor implementovaný v jazyce R

`cyl`	počet válců
`disp`	objem válců (kubické palce)
`hp`	výkon (počet koní)
`drat`	převodový poměr zadní nápravy
`wt`	hmotnost vozidla (kilolibry)
`qsec`	zrychlení (počet sekund z 0 na 1/4 míle)
`vs`	uspořádání válců (1 – „V“, 0 – za sebou)
`am`	převodovka (0 – automat, 1 – manuál)
`gear`	počet převodových stupňů
`carb`	počet karburátorů

Testujte multikolinearitu v daném modelu. Metodou postupné regrese nalezněte vhodný model. Ověřte také normalitu residuí.

[Multikolinearita se nezamítá, vhodný model: $\texttt{mpg} = \beta_0 + \beta_1\texttt{wt} + \beta_2\texttt{cyl}$, normalita residuí se nezamítá.]

Cvičení 4.5 (pro náročné)**.** Naprogramujte funkci „`multicol.R`“, která pro zadaný model zjistí přítomnost multikolinearity v datech. V případě, že je multikolinearita přítomna, metodou postupné regrese nalezne vhodný model.

KAPITOLA 6

Analýza rozptylu

Základní informace

(1) V následující kapitole se budeme zabývat obecně problematikou jednofaktorové analýzy rozptylu. Popíšeme lineární regresní model, ze kterého analýza vychází a aplikujeme výsledky tohoto modelu při testování hypotézy o shodě středních hodnot. Dále také uvedeme testy shody rozptylů a metody mnohonásobného porovnávání.
(2) Předpokládá se znalost základních pojmů z teorie pravděpodobnosti a matematické statistiky – známá rozdělení náhodné veličiny, dále se předpokládá znalost lineárního regresního modelu a testování hypotéz o parametrech tohoto modelu.

Výstupy z výukové jednotky

Studenti

- pochopí konstrukci modelu a testování hypotézy o shodě středních hodnot
- umí sestrojit tabulku analýzy rozptylu
- umí testovat shodu rozptylů
- ovládají metody mnohonásobného porovnávání
- aplikují úvahy analýzy rozptylu na výběrech z alternativního rozdělení

1. Motivace

Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny Y, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina Y).

Předpokládáme, že faktor A má $a \geq 3$ úrovní a i-té úrovni odpovídá n_i výsledků $Y_{i1}, \ldots, Y_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \ldots, a$ a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy $Y_{ij} = \mu_i + \varepsilon_{ij}$, kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, kde $i = 1, \ldots, a$ a $j = 1, \ldots, n_i$.

Na hladině významnosti α testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné oproti alternativní hypotéze, která tvrdí, že alespoň jedna dvojice středních hodnot se liší. Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $r(r-1)/2$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Tento postup však nelze použít, neboť nezaručuje splnění podmínky, že pravděpodobnost chyby 1. druhu je α. Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA[1] (analýza rozptylu, v popsané situaci analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.

Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží metoda mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.

Poznámka 1.1. Většina textu v této kapitole byla převzata z [**3**]. Pro podrobnější studium tohoto tématu proto odkazujeme na tento zdroj.

[1] Z anglického ANalysis Of VAriance

1.1. Označení. Výsledky pokusu popíšeme pomocí spojité náhodné veličiny Y a to tak, že sledujeme výsledky tohoto pokusu při všech úrovních faktoru A. Zjištěné hodnoty $\mathbf{Y} = (Y_1, \dots, Y_n)'$ roztřídíme do $\boxed{a}$ skupin podle úrovní do následující tabulky:

Úroveň faktoru	*Počet pozorování*	*Naměřené hodnoty*	*Součet úrovně*	*Průměr úrovně*	*Rozdělení úrovně*
1.	n_1	$\mathbf{Y}_1 = (Y_{11}, \dots, Y_{1n_1})'$	$Y_{1\cdot} = \sum_{i=1}^{n_1} Y_{1i}$	$\overline{Y}_{1\cdot} = \frac{1}{n_1} Y_{1\cdot}$	$Y_{1i} \sim \mathcal{L}(\mu_1, \sigma^2)$
2.	n_2	$\mathbf{Y}_2 = (Y_{21}, \dots, Y_{2n_2})'$	$Y_{2\cdot} = \sum_{i=1}^{n_2} Y_{2i}$	$\overline{Y}_{2\cdot} = \frac{1}{n_2} Y_{1\cdot}$	$Y_{2i} \sim \mathcal{L}(\mu_2, \sigma^2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a-tá	n_a	$\mathbf{Y}_a = (Y_{a1}, \dots, Y_{an_a})'$	$Y_{a\cdot} = \sum_{i=1}^{n_a} Y_{ai}$	$\overline{Y}_{a\cdot} = \frac{1}{n_a} Y_{a\cdot}$	$Y_{ai} \sim \mathcal{L}(\mu_a, \sigma^2)$
Součet	n		$Y_{\cdot\cdot} = \sum_{j=1}^{a} \sum_{i=1}^{n_j} Y_{ji}$	$\overline{Y}_{\cdot\cdot} = \frac{1}{n} Y_{\cdot\cdot}$	

2. Testování hypotézy o shodě středních hodnot

Definujme nejprve obecně základní model, kterým se řídí pozorované náhodné veličiny. Z tohoto modelu budeme vycházet v dalších úvahách.

DEFINICE 2.1 (model M). Náhodné veličiny Y_{ij} se řídí modelem M:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

pro $i = 1, \dots, a$ a $j = 1, \dots, n_i$, přičemž ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, μ je společná část střední hodnoty proměnné veličiny, α_i je efekt faktoru A na úrovni i.

Při zkoumání vlivu jednoho faktoru A testujeme hypotézu

$$\boxed{H_0}: \alpha_1 = \dots = \alpha_a = 0 \quad \text{proti alternativě} \quad \boxed{H_1}: \exists\, i : \alpha_i \neq 0$$

Jinými slovy hypotéza říká, že $\mathbf{Y} = (\mathbf{Y}_1', \dots, \mathbf{Y}_a')'$ tvoří jeden náhodný výběr, alternativa znamená, že táž pozorování představují obecně a náhodných výběrů lišících se střední hodnotou. To odpovídá situaci, kdy pozorování byla pořízena za a různých podmínek, jejichž vliv, pokud existuje, se dá vyjádřit aditivní změnou střední hodnoty.

Pokud hypotézu zamítneme, považujeme vliv zkoumaného faktoru za významný, v opačném případě za bezvýznamný.

Pokud tedy platí nulová hypotéza H_0, dostáváme následující minimální submodel.

DEFINICE 2.2 (model M_0). Náhodné veličiny Y_{ij} se řídí modelem M_0:

$$Y_{ij} = \mu + \varepsilon_{ij},$$

pro $i = 1, \dots, a$ a $j = 1, \dots, n_i$, přičemž ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$.

Poznámka 2.3 (odvození pro zvídavé)**.** Vyjděme ze základního modelu M.

Matice plánu je $\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \vdots & & \ddots & \mathbf{1}_{n_{a-1}} & \mathbf{0} \\ \mathbf{1}_{n_a} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}_{n_a} \end{pmatrix}$ a vektor parametrů $\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \vdots \\ \alpha_a \end{pmatrix}$,

vektor $\mathbf{1}_k$ značí sloupcový vektor složený z k jedniček. Matice $\mathbf{X}$ má $(a+1)$ sloupců a není plné hodnosti, neboť první sloupec dostaneme, pokud sečteme zbývajících a sloupců.

Vyjádřeme nejprve systém normálních rovnic $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n_1 & n_2 & \cdots & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & \cdots & 0 \\ n_2 & 0 & n_2 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ n_{a-1} & \vdots & & \ddots & n_{a-1} & 0 \\ n_a & 0 & \cdots & \cdots & 0 & n_a \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \cdots & \mathbf{1}'_{n_{a-1}} & \mathbf{1}'_{n_a} \\ \mathbf{1}'_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{n_2} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{1}'_{n_{a-1}} & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}'_{n_a} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \vdots \\ \mathbf{Y}_{a-1} \\ \mathbf{Y}_a \end{pmatrix} = \begin{pmatrix} Y_{\cdot\cdot} \\ Y_{1\cdot} \\ \vdots \\ \vdots \\ Y_{a-1\cdot} \\ Y_{a\cdot} \end{pmatrix}.$$

Jednou z pseudoinverzních matic k matici $\mathbf{X}'\mathbf{X}$ je matice

$$(\mathbf{X}'\mathbf{X})^- = \begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{n_1} & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \frac{1}{n_2} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \vdots & & \ddots & \frac{1}{n_{a-1}} & 0 \\ 0 & 0 & \cdots & \cdots & 0 & \frac{1}{n_a} \end{pmatrix} \quad \Rightarrow \quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' = \begin{pmatrix} \frac{1}{n_1}\mathbf{E}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{n_a}\mathbf{E}_{n_a} \end{pmatrix},$$

kde $\mathbf{E}_k = \mathbf{1}_k\mathbf{1}'_k$ je matice typu $(k \times k)$ samých jedniček. Odtud

$$\widehat{\mathbf{Y}} = \begin{pmatrix} (\widehat{\mu} + \widehat{\alpha}_1) \cdot \mathbf{1}_{n_1} \\ \vdots \\ \vdots \\ (\widehat{\mu} + \widehat{\alpha}_a) \cdot \mathbf{1}_{n_a} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{Y}}_1 \\ \vdots \\ \vdots \\ \widehat{\mathbf{Y}}_a \end{pmatrix} = \mathbf{H}\mathbf{Y} = \begin{pmatrix} \frac{1}{n_1}\mathbf{E}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{n_a}\mathbf{E}_{n_a} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \vdots \\ \mathbf{Y}_a \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1\cdot}\,\mathbf{1}_{n_1} \\ \vdots \\ \vdots \\ \overline{Y}_{a\cdot}\,\mathbf{1}_{n_a} \end{pmatrix}$$

takže odhad střední hodnoty je tvaru $\widehat{\mu}+\widehat{\alpha}_j = \overline{Y}_{j\cdot}$. Přidáním dodatečné podmínky $\sum_{j=1}^{a} n_j\alpha_j = 0$, dostaneme odhad společné střední hodnoty $\widehat{\mu} = \overline{Y}_{\cdot\cdot}$ a pro $j = 1, \ldots, a$ odhad příspěvku j-té skupiny $\widehat{\alpha}_j = \overline{Y}_{j\cdot} - \overline{Y}_{\cdot\cdot}$

Pokud platí nulová hypotéza H_0, dostáváme minimální submodel M_0, který vznikne ze základního modelu vypuštěním posledních a sloupců matice plánu, takže pokud vše vyjádříme maticově, máme $\mathbf{Y} = \mathbf{X}_0\,\beta_0 + \boldsymbol{\varepsilon}$. Vidíme, že jde o model plné hodnosti, ve kterém

$$\mathbf{X}_0 = \mathbf{1}_n, \quad \mathbf{X}'_0\mathbf{X}_0 = \mathbf{1}'_n\mathbf{1}_n = n, \quad \mathbf{X}'_0\mathbf{Y} = \mathbf{1}'_n\mathbf{Y} = Y_{\cdot\cdot} \quad \text{a} \quad \hat{\beta}_0 = (\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}'_0\mathbf{Y} = \tfrac{1}{n}Y_{\cdot\cdot} = \overline{Y}_{\cdot\cdot}$$

Pak $\mathbf{H}_0 = \mathbf{X}_0(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}'_0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n = \frac{1}{n}\mathbf{E}_n$ a $\widehat{\mu}_0 = \widehat{\mathbf{Y}}_0 = \mathbf{H}_0\mathbf{Y} = \frac{1}{n}\mathbf{E}_n\mathbf{Y} = \overline{Y}_{\cdot\cdot}\mathbf{1}_n$.
Tedy součty kvadrátů odchylek

$$\boxed{S_e} = \|\widehat{\boldsymbol{\varepsilon}}\|^2 = (\mathbf{Y}-\widehat{\boldsymbol{\mu}})'(\mathbf{Y}-\widehat{\boldsymbol{\mu}}) = (\mathbf{Y}-\widehat{\mathbf{Y}})'(\mathbf{Y}-\widehat{\mathbf{Y}})$$
$$= \sum_{j=1}^{a}(\mathbf{Y}_j - \overline{Y}_{j\cdot}\mathbf{1}_{n_j})'(\mathbf{Y}_j - \overline{Y}_{j\cdot}\mathbf{1}_{n_j}) = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_{j\cdot})^2 \qquad \textit{reziduální}$$

$$S_{e_0} = \boxed{S_T} = \|\widehat{\boldsymbol{\varepsilon}}_0\|^2 = (\mathbf{Y}-\widehat{\boldsymbol{\mu}}_0)'(\mathbf{Y}-\widehat{\boldsymbol{\mu}}_0) = \sum_{j=1}^{a}(\mathbf{Y}_j - \overline{Y}_{\cdot\cdot}\mathbf{1}_{n_j})'(\mathbf{Y}_j - \overline{Y}_{\cdot\cdot}\mathbf{1}_{n_j}) = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_{\cdot\cdot})^2 \qquad \textit{celkový}$$

$$S_{\Delta_0} = \boxed{S_A} = \|\boldsymbol{\Delta}_0\|^2 = (\widehat{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}_0)'(\widehat{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}_0) = \sum_{j=1}^{a}(\overline{Y}_{j\cdot}\mathbf{1}_{n_j} - \overline{Y}_{\cdot\cdot}\mathbf{1}_{n_j})'(\overline{Y}_{j\cdot}\mathbf{1}_{n_j} - \overline{Y}_{\cdot\cdot}\mathbf{1}_{n_j})$$
$$= \sum_{j=1}^{a}(\overline{Y}_{j\cdot} - \overline{Y}_{\cdot\cdot})^2\mathbf{1}'_{n_j}\mathbf{1}_{n_j} = \sum_{j=1}^{a} n_j(\overline{Y}_{j\cdot} - \overline{Y}_{\cdot\cdot})^2 \qquad \textit{mezi třídami}$$
$$= S_{e_0} - S_e \quad \Rightarrow \quad \boxed{S_T = S_A + S_e}$$

takže pokud platí model $\boxed{M}$, pak statistika $F_A = \frac{(S_{e_0}-S_e)/(a-1)}{S_e/(n-a)} \sim F(a-1, n-a)$.

DEFINICE 2.4. Zavedeme součty čtverců:

- **Celkový součet čtverců** (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru), počet stupňů volnosti $df_T = n - 1$:

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2$$

- **Skupinový součet čtverců** (charakterizuje variabilitu mezi jednotlivými náhodnými výběry), počet stupňů volnosti $df_A = a - 1$:

$$S_A = \sum_{j=1}^{a} n_j \left(\overline{Y}_{j\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$$

- **Reziduální součet čtverců** (charakterizuje variabilitu uvnitř jednotlivých výběrů), počet stupňů volnosti $df_e = n - a$:

$$S_e = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{j\cdot}\right)^2 .$$

VĚTA 2.5. *Lze dokázat, že*

$$S_T = S_A + S_E.$$

VĚTA 2.6. *Rozdíl mezi modely M a M_0 ověřujeme pomocí testové statistiky*

$$F_A = \frac{S_A/df_A}{S_e/df_e},$$

která se řídí rozložením $F(a-1, n-a)$, je-li model M_0 správný.

Předcházející pojmy se shrnují v **tabulce analýzy rozptylu**

Zdroj variability	*Součet čtverců* SS	*Stupně volnosti* df	*Podíl* $MS = \frac{SS}{df}$	$F = \frac{MS}{s^2}$
Třídy	S_A	$df_a = a - 1$	$MS_A = \frac{S_A}{df_a}$	$F_A = \frac{MS_A}{MS_e}$
Reziduální	S_e	$df_e = n - a$	$MS_e = \frac{S_e}{df_e}$	–
Celkový	S_T	$df_T = n - 1$	–	–

Je-li konkrétní realizace statistiky F_A (značíme malými písmeny f_A) větší než $(1-\alpha)$-kvantil F–rozdělení se stupni volnosti $a - 1$ a $N - a$, tj. $f_A > F_{1-\alpha}(a-1, n-a)$, pak zamítáme nulovou hypotézu na hladině významnosti α.

Bývá zvykem označovat v tabulce překročení kvantilu $F_{0.95}(a-1, n-a)$ (tj. $\alpha = 0.05$) označovat jednou hvězdičkou, dvě hvězdičky u $\alpha = 0.01$ a tři hvězdičky u $\alpha = 0.001$. Někdy se přidává sloupec s p-hodnotou, což je $P(F_A > f_A)$.

3. Bartlettův a Levenův test shody rozptylů

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných a výběrech. Uvedeme zde dva testy – Levenův[2] a Bartlettův[3].

VĚTA 3.1 (Levenův test). *Položme* $Z_{ij} = |Y_{ij} - \overline{Y}_{i\cdot}|$. *Označme:*

- $\overline{Z}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$
- $\overline{Z}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} Z_{ij}$
- $S_{Ze} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_{i\cdot}\right)^2$
- $S_{ZA} = \sum_{i=1}^{a} n_i \left(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}\right)^2$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_Z = \frac{S_{ZA}/(a-1)}{S_{Ze}/(n-a)} \sim F(a-1, n-a).$$

H_0 *tedy zamítáme na hladině významnosti* α, *když* $F_Z \geq F_{1-\alpha}(a-1, n-a)$.

VĚTA 3.2 (Bartlettův test). *Platí-li hypotéza o shodě rozptylů, pak statistika*

$$B = \frac{1}{C} \left[(n-a) \ln S_*^2 - \sum_{j=1}^{a} (n_j - 1) \ln S_j^2 \right] \approx \chi^2(a-1),$$

kde

$$C = 1 + \frac{1}{3(a-1)} \left(\sum_{j=1}^{a} \frac{1}{n_j - 1} - \frac{1}{n-a} \right), \quad S_*^2 = \frac{S_e}{n-a}.$$

H_0 *zamítáme na asymptotické hladině významnosti* α, *když* $B \geq \chi^2_{1-\alpha}(a-1, n-a)$.

4. Metody mnohonásobného porovnávání

Zamítneme-li na hladině významnosti α hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti α. Mají-li všechny výběry týž rozsah $\boxed{p}$ (říkáme, že třídění je vyvážené), použijeme Tukeyovu[4] metodu, nemají-li všechny výběry stejný rozsah, použijeme Scheffého[5] metodu.

[2]Howard Levene (1914 — 2003). Profesor matematické statistiky a genetiky na Kolumbijské univerzitě.
[3]Maurice Stevenson Bartlett (1910 – 2002). Anglický matematik
[4]John Wilder Tukey (1915 – 2000). Americký matematik.
[5]Henry Scheffé (1907 – 1977). Americký matematik.

VĚTA 4.1 (Tukeyova metoda). *Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když:*

$$|\overline{Y}_{k\cdot} - \overline{Y}_{l\cdot}| \geq q_{1-\alpha}(a, n-a)\frac{S_*}{\sqrt{p}},$$

kde $q_{1-\alpha}(a, n-a)$ jsou kvantily studentizovaného rozpětí, které najdeme ve statistických tabulkách.

VĚTA 4.2 (Scheffého metoda). *Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když:*

$$|\overline{Y}_{k\cdot} - \overline{Y}_{l\cdot}| \geq S_* \sqrt{(a-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(a-1, n-a)}.$$

Poznámka 4.3. Může nastat situace, kdy při zamítnutí nulové hypotézy nenajdeme významný rozdíl u žádné dvojice středních hodnot. Pak je významně rozdílná některá složitější kombinace středních hodnot.

Příklad 4.4. U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky uvádí tabulka:

odrůda	hmotnost (v kg)				
A	0,9	0,8	0,6	0,9	
B	1,3	1,0	1,3		
C	1,3	1,5	1,6	1,1	1,5
D	1,1	1,2	1,0		

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Řešení. Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.
Výpočtem získáme: $\overline{y}_{1\cdot} = 0,8$, $\overline{y}_{2\cdot} = 1,2$, $\overline{y}_{3\cdot} = 1,4$, $\overline{y}_{4\cdot} = 1,1$, $\overline{y}_{\cdot\cdot} = 1,14$, $S_e = 0,3$, $S_A = 0,816$, $S_T = 1,116$, $F_A = 9,97$. Ze statistických tabulek získáme $F_{0,95}(3,11) = 3,59$. Protože testová statistika se realizuje v kritickém oboru, zamítáme nulovou hypotézu na hladině významnosti 0,05.
Výsledky zapíšeme do tabulky ANOVA:

Zdroj variability	Součet čtverců	Stupně volnosti	Podíl	F_A
třídy	$S_A = 0,816$	3	$S_A/3 = 0,272$	$\frac{S_A/3}{S_E/11} = 9,97$
reziduální	$S_E = 0,3$	11	$S_E/11 = 0,02727$	—
celkový	$S_T = 1,116$	14	—	—

Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $\lvert m_k - m_l \rvert$	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,67	0,36
A, D	0,3	0,41
B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

Na hladině významnosti 0,05 se liší odrůdy A a C.

Poznámka 4.5. Význam předpokladů v analýze rozptylu

- Nezávislost jednotlivých náhodných výběrů - velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.
- Normalita – ANOVA není příliš citlivá na porušení normality, zvlášť pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení se doporučuje Kruskalův – Wallisův test (viz např. [**3**]).
- Shoda rozptylů – mírné porušení nevadí, při větším se doporučuje Kruskalův – Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.

5. Kruskalův – Wallisův test

Kruskalův[6] – Wallisův[7] test je neparametrická obdoba analýzy rozptylu jednoduchého třídění.

Formulace problému

Nechť je dáno a nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_a$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme $n = n_1 + \ldots + n_a$. Chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

VĚTA 5.1 (Kruskalův – Wallisův test). *Všech n hodnot seřadíme do rostoucí posloupnosti a určíme pořadí každé hodnoty. Označme T_j součet pořadí těch hodnot, které patří do j-tého výběru, $j = 1, \ldots, a$ (kontrola: musí platit $T_1 + \ldots + T_a = n(n+1)/2$).*
Testová statistika má tvar:

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{a} \frac{T_j^2}{n_j} - 3(n+1). \tag{18}$$

Platí-li H_0, má statistika Q asymptoticky rozložení $\chi^2(a-1)$, rostou-li rozsahy výběrů nade všechny meze. H_0 tedy zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi^2_{1-\alpha}(a-1)$.

6. Více nezávislých náhodných výběrů z alternativních rozložení

6.1. Test homogenity binomických rozložení. Nechť $Y_{j1}, \ldots, Y_{jn_j} \sim A(\theta_j)$, $j = 1, 2, \ldots, a$ jsou nezávislé náhodné výběry z alternativního rozložení. Testujeme hypotézu H_0: $\theta_1 = \cdots = \theta_a$ proti alternativní hypotéze H_1: „alespoň jedna dvojice parametrů je různá“.

[6] William Kruskal (1919 – 2005). Americký matematik.
[7] Wilson Allen Wallis (1912 – 1988). Americký matematik

VĚTA 6.1. *Statistika*

$$Q = \frac{1}{\overline{Y}_{\cdot\cdot}(1-\overline{Y}_{\cdot\cdot})} \sum_{j=1}^{a} n_j \left(\overline{Y}_{j\cdot} - \overline{Y}_{\cdot\cdot}\right)^2,$$

má v případě platnosti nulové hypotézy asymptoticky rozložení $\chi^2(a-1)$. *H_0 tedy zamítáme na asymptotické hladině významnosti* α, *když* $Q \geq \chi^2_{1-\alpha}(a-1)$.

Poznámka 6.2. Test lze použít, pokud $n_j \overline{y}_{\cdot\cdot} > 5$ pro všechna $j = 1, \ldots, a$.

Poznámka 6.3. Statistiku Q lze snadno upravit do Brandtova[8] – Snedecorova[9] výpočetního tvaru

$$Q = \frac{1}{\overline{Y}_{\cdot\cdot}(1-\overline{Y}_{\cdot\cdot})} \sum_{j=1}^{a} n_j \overline{Y}^2_{j\cdot} - n\frac{\overline{Y}_{\cdot\cdot}}{1-\overline{Y}_{\cdot\cdot}}. \tag{19}$$

6.2. Test homogenity binomických rozložení založený na arkussinusové transformaci. Není-li splněna podmínka $n_j \overline{y}_{\cdot\cdot} > 5$ pro všechna $j = 1, \ldots, a$, doporučuje se následující postup:

VĚTA 6.4. *Označme*

- $A_j = \arcsin \sqrt{\overline{Y}_{j\cdot}}$.
- $B = \frac{1}{n} \sum_{j=1}^{a} n_j A_j$.

Pak statistika

$$Q = 4 \sum_{j=1}^{a} n_j \left(A_j - B\right)^2 \approx \chi^2(a-1).$$

H_0 tedy zamítáme na asymptotické hladině významnosti α, *když* $Q \geq \chi^2_{1-\alpha}(a-1)$.

6.3. Mnohonásobné porovnávání. Zamítneme-li nulovou hypotézu na asymptotické hladině významnosti α, chceme zjistit, které dvojice parametrů θ_k a θ_l se liší.

VĚTA 6.5. *Platí-li nerovnost*

$$|A_k - A_l| \geq \sqrt{\frac{1}{8}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)} \cdot q_{1-\alpha}(a, \infty),$$

pak na hladině významnosti α *zamítáme hypotézu o shodě parametrů* θ_k *a* θ_l.

Poznámka 6.6. Hodnoty $q_{1-\alpha}(a, \infty)$ jsou kvantily studentizovaného rozpětí. Najdeme je ve statistických tabulkách.

Příklad 6.7. Na gymnázium bylo přijato 142 studentů. Ti byli náhodně rozděleni do tříd A, B, C, D. V každé třídě byla matematika vyučována jinou metodou. Na konci školního roku psali všichni studenti stejnou písemnou práci a byl zaznamenán počet těch studentů, kteří vyřešili všechny zadané úkoly.

Třída	A	B	C	D
Počet studentů	35	36	37	34
Počet úspěšných studentů	5	8	17	15

[8] Alva Esmond Brandt. Americký matematik. Pomohl vzniku katedry statistiky na Zemědělské fakultě Floridské univerzity.

[9] George Waddell Snedecor (1882 – 1974). Americký matematik.

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozdíly v podílech studentů v jednotlivých třídách, kteří správně vyřešili všechny zadané úlohy, jsou způsobeny pouze náhodnými vlivy.

Řešení. Máme čtyři nazávislé náhodné výběry, j-tý pochází z rozložení $A(\theta_j)$, $j = 1, 2, 3, 4$. Testujeme hypotézu H_0: $\theta_1 = \theta_2 = \theta_3 = \theta_4$. Ze zadání a výpočtem zjistíme: $n_1 = 35$, $n_2 = 36$, $n_3 = 37$, $n_4 = 34$, $\overline{y}_{1.} = 5/35$, $\overline{y}_{2.} = 8/36$, $\overline{y}_{3.} = 17/37$, $\overline{y}_{4.} = 15/34$, $\overline{y}_{..} = 45/142$, $Q = 12,288$, $\chi^2_{0,95}(3) = 7,81$.
Protože testové kritérium se realizuje v kritickém oboru, H_0 zamítáme na asymptotické hladině významnosti 0,05.
Spočteme arkussinusové transformace výběrových průměrů. Vyjde: $A_1 = 0,3876$, $A_2 = 0,4909$, $A_3 = 0,7448$, $A_4 = 0,7264$ Nyní metodou mnohonásobného porovnávání zjistíme, které dvojice parametrů se od sebe liší na hladině významnosti 0,05.

Srovnávané třídy	Rozdíly $\lvert A_k - A_l \rvert$	Pravá strana vzorce
A, B	0,1033	0,30
A, C	0,3572	0,30
A, D	0,3388	0,31
B, C	0,2539	0,30
B, D	0,2356	0,31
C, D	0,0184	0,30

Na hladině významnosti 0,05 se liší třídy A, C a A, D.

Úlohy k procvičení

Cvičení 6.1. Jsou známy měsíční tržby (v tisících Kč) tří prodavačů za dobu půl roku.

1. prodavač	12	10	9	10	11	9
2. prodavač	10	12	11	12	14	13
3. prodavač	19	18	16	16	17	15

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty tržeb všech tří prodavačů jsou stejné. Pokud zamítneme nulovou hypotézu, zjistěte, tržby kterých dvou prodavačů se liší na hladině významnosti 0,05.

[Na hladině významnosti 0,05 se liší tržby prodavačů 1, 3 a 2, 3.]

Cvičení 6.2. Naprogramujte funkci „`anovabinom.R`“, která pro vstupní vektory `nj` (počet pozorování ve skupinách) a `pj` (počet „úspěchů“ ve skupinách) provede analýzu rozptylu pro binomická data. V případě zamítnutí nulové hypotézy vypíše indexy skupin, které se od sebe významně liší.

Cvičení 6.3. 104 náhodně vybraných matek bylo dotázáno, zda jejich kojenec dostává dudlík. Zjišťoval se též nejvyšší stupeň dosaženého vzdělání matky.

Vzdělání matky	Počet matek	Počet dětí s dudlíkem
základní	39	27
středoškolské	47	34
vysokoškolské	18	15

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že podíly dětí s dudlíkem nezávisí na vzdělání matky.

[nezávisí]

Cvičení 6.4. Je dáno pět nezávislých náhodných výběrů o rozsazích 5, 7, 6, 8, 5, přičemž i-tý výběr pochází z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \ldots, 5$. Byl vypočten celkový součet čtverců $S_T = 15$ a reziduální součet čtverců $S_e = 3$. Na hladině významnosti 0,05 testujte hypotézu o shodě středních hodnot.

[$n = 31$, $a = 5$, $S_A = 12$, $f_A = 26$, $F_{0,95}(4, 26) = 2,7426$ Protože $f_A \geq F_{0,95}(4, 26)$, H_0 zamítáme na hladině významnosti 0,05.]

Cvičení 6.5. V proměnné „`LakeHuron`“[10] jsou uloženy roční údaje o hloubce jezera Huron (ve stopách) v letech 1875 – 1972. Data proložte polynomem 8. stupně. Pomocí analýzy rozptylu zkoumejte možnosti zmenšení stupně regresního polynomu.

[Možno jít na stupeň 7.]

Cvičení 6.6. U 126 podniků řepařské oblasti v České Republice byl sledován hektarový výnos cukrovky ve vztahu ke spotřebě průmyslových hnojiv.
Data jsou uložena v souboru „`cukrovka.Rdata`“ ve 4 sloupcích:

(1) dolní hranice spotřeby K_2O (kg/ha)
(2) horní hranice spotřeby K_2O (kg/ha)
(3) četnosti
(4) průměrné výnosy cukrovky (q/ha)

a) odhadněte parametry regresní funkce tvaru

$$y = \beta_0 + \beta_1 x$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Poznámka: Za hodnoty nezávisle proměnné volte střed intervalu.

b) Porovnejte vhodnost použitých regresních modelů pomocí analýzy rozptylu.

[Kvadratický model je významný.]

[10] datový soubor implementovaný v jazyce R

KAPITOLA 7

Zobecněné lineární modely

Základní informace

(1) V následující kapitole se budeme zabývat statistickou analýzou vzájemných vztahů náhodných jevů, kde již nebude stačit předpokládat, že tyto vztahy lze popsat pomocí lineárních operací, ale budeme používat obecnějších modelů. Popíšeme obecně volbu optimálních parametrů modelu a uvedeme konkrétní příklady modelů.
(2) Předpokládá se znalost základních pojmů z teorie lineárních regresních modelů a také některých pojmů z teorie odhadu – matice plánu, metoda nejmenších čtverců, testování hypotéz o parametrech modelu, analýza rozptylu, metoda maximální věrohodnosti

Výstupy z výukové jednotky

Studenti

- definují hustotu exponenciálního typu
- definují zobecněný lineární regresní model
- vypočítají odhady neznámých parametrů metodou maximální věrohodnosti a testují hypotézy o těchto parametrech
- definují škálovou deviaci modelu
- testují hypotézy o vhodnosti submodelu

1. Motivace

V reálném světě má mnoho procesů jiný, než lineární vztah závislosti. Např. v ekonomii se ukazuje, že mnoho vztahů má logaritmickou závislost, k vysvětlení procesů v přírodních vědách se užívají reciproké, mocninné i další vztahy. Vysvětlovaná veličina popisující pravděpodobnost přežití člověka, v případě určité nemoci a určitého způsobu léčby, může z definice pravděpodobnosti nabývat hodnot pouze z intervalu $[0, 1]$, což by v případě klasického lineárního modelu bylo možné zajistit jen za přijetí určitých omezení na parametry modelu. Také normalita chyb je často nesplněným předpokladem klasického lineárního regresního modelu. Připomeňme, že normalita se vyznačuje nezávislosti střední hodnoty a rozptylu. Typicky např. u ekonomických veličin s rostoucí střední hodnotou obvykle roste rozptyl náhodné veličiny, přičemž náhodné chyby mají v těchto případech často nesymetrická, kladně sešikmená rozdělení.

Klasický lineární regresní model je tedy sice velmi důležitým stochastickým modelem, avšak má celou řadu omezení. Je omezen pouze na třídu normálních rozdělení a předpokládá striktní rovnost mezi střední hodnotou náhodné veličiny Y a lineární kombinací prediktorů. Je však možné provést určitá zobecnění tohoto klasického lineárního modelu a tím se bude zabývat následující kapitola.

Poznámka 1.1. Většina textu v této kapitole byla převzata z [**5**]. Pro podrobnější studium tohoto tématu proto odkazujeme na tento zdroj.

2. Základní pojmy a definice

2.1. Maximálně věrohodné odhady. Uveďme nejprve pár nezbytných definic a pojmů nezbytných k dalším úvahám.

DEFINICE 2.1. Mějme parametrický prostor $\boldsymbol{\Theta} \subset \mathbb{R}^m$. Řekneme, že systém m-parametrických hustot

$$\mathcal{F}^m_{\text{reg}} = \{f(\mathbf{y};\boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1,\dots,\theta_m)' \in \boldsymbol{\Theta}\}$$

je **regulární**, jestliže platí

(1) $\boldsymbol{\Theta} \subset \mathbb{R}^m$ je otevřená borelovská množina.

(2) Množina $M = \{\mathbf{y} \in \mathbb{R}^n : f(\mathbf{y};\boldsymbol{\theta}) > 0\}$ nezávisí na parametru $\boldsymbol{\theta}$.

(3) Pro každé $\mathbf{y} \in M$ existuje konečná parciální derivace

$$f'_i(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} \quad (i = 1,\dots,m).$$

(4) Pro všechny $\boldsymbol{\theta} = (\theta_1,\dots,\theta_m)' \in \boldsymbol{\Theta}$ platí

$$\int_M \frac{f'_i(\mathbf{y};\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta})} dF(\mathbf{y};\boldsymbol{\theta}) = \int_M \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} dF(\mathbf{y};\boldsymbol{\theta}) = 0 \qquad i = 1,\dots,m,$$

kde $F(\mathbf{y};\boldsymbol{\theta})$ je odpovídající distribuční funkce.

(5) Pro všechny $\boldsymbol{\theta} = (\theta_1,\dots,\theta_m)' \in \boldsymbol{\Theta}$ je integrál

$$J_{ij}(\boldsymbol{\theta}) = \int_M \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_j} dF(\mathbf{y};\boldsymbol{\theta}) \qquad i,j = 1,\dots,m$$

konečný a matice $\mathbf{J} = \mathbf{J}(\boldsymbol{\theta}) = (J_{ij}(\boldsymbol{\theta}))_{i,j=1}^m$ je pozitivně definitní. Matice $\mathbf{J}$ se nazývá **Fisherova informační matice** o parametru $\boldsymbol{\theta}$.

Poznámka 2.2. Pro jednoduchost někdy hovoříme o regularitě $f(\mathbf{y};\boldsymbol{\theta})$, ne o regularitě systému hustot.

DEFINICE 2.3. Nechť $f \in \mathcal{F}^m_{\text{reg}}$. Pak náhodný vektor

$$\mathbf{U} = \mathbf{U}(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}),\dots,U_m(\boldsymbol{\theta}))' \qquad \text{se složkami} \qquad U_i = U_i(\boldsymbol{\theta}) = \frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \theta_i}$$

se nazývá **skórový vektor** příslušný hustotě f.

VĚTA 2.4.

(1) Je-li $f \in \mathcal{F}^m_{reg}$ a pro $i,j = 1,\dots,m$ existují

$$f''_{ij}(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j},$$

pak

$$E\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0} \qquad a \qquad D\mathbf{U}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta})\,.$$

(2) Platí-li navíc pro $i,j = 1,\dots,m$

$$E\left(\frac{f''_{ij}(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y};\boldsymbol{\theta})}\right) = 0,$$

pak

$$\mathbf{J}(\boldsymbol{\theta}) = -E(\mathbf{U}'(\boldsymbol{\theta})),$$

kde

$$\mathbf{U}'(\boldsymbol{\theta}) = \left(\frac{\partial U_i(\boldsymbol{\theta})}{\partial \theta_j}\right)_{i,j=1}^m.$$

V dalším budeme uvažovat **náhodný výběr** $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$ z rozdělení s regulární hustotou $f \in \mathcal{F}^m_{\text{reg}}$. Označme $M = \{y \in \mathbb{R} : f(y; \boldsymbol{\theta}) > 0\}$. Pak sdružená (simultánní) hustota náhodného vektoru $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$ je rovna

$$f_{\mathbf{Y}_n}(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}), \qquad \mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n,$$

neboť náhodný výběr je tvořen systémem nezávislých náhodných veličin.

Zaveďme následující značení pro:

funkce:

$$\begin{aligned} l_k &= l(\boldsymbol{\theta}; y_k) = \ln f(y_k; \boldsymbol{\theta}) \\ l^*_n &= l^*_n(\boldsymbol{\theta}; \mathbf{y}) = \ln f_{\mathbf{Y}_n}(\mathbf{y}; \boldsymbol{\theta}) \end{aligned}$$

náhodné vektory:

$$\begin{aligned} \mathbf{U}_k &= \mathbf{U}_k(\boldsymbol{\theta}) = (U_{1,k}(\boldsymbol{\theta}), \ldots, U_{m,k}(\boldsymbol{\theta}))' = \left(\tfrac{\partial \ln f(Y_k;\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \tfrac{\partial \ln f(Y_k;\boldsymbol{\theta})}{\partial \theta_m} \right)' \\ \mathbf{U}^*_n &= \mathbf{U}^*_n(\boldsymbol{\theta}) = (U^*_1(\boldsymbol{\theta}), \ldots, U^*_m(\boldsymbol{\theta}))' = \left(\tfrac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{Y};\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \tfrac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{Y};\boldsymbol{\theta})}{\partial \theta_m} \right)' \end{aligned}$$

maticové funkce:

$$\begin{aligned} \mathbf{J} &= \mathbf{J}(\boldsymbol{\theta}) = (J_{ij}(\boldsymbol{\theta}))^m_{i,j=1} = \left(\int_M \tfrac{\partial \ln f(y;\boldsymbol{\theta})}{\partial \theta_i} \tfrac{\partial \ln f(y;\boldsymbol{\theta})}{\partial \theta_j} dF(y;\boldsymbol{\theta}) \right)^m_{i,j=1} \\ \mathbf{J}_n &= \mathbf{J}_n(\boldsymbol{\theta}) = \left(\int_M \cdots \int_M \tfrac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} \tfrac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_j} dF_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) \right)^m_{i,j=1} \end{aligned}$$

VĚTA 2.5. *Uvažujme náhodný výběr* $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$ *z rozdělení s hustotou* $f \in \mathcal{F}^m_{reg}$.
(1) Pokud pro $i, j = 1, \ldots, m$ *existují*

$$f''_{ij}(y; \boldsymbol{\theta}) = \frac{\partial^2 f(y; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j},$$

pak

$$E\mathbf{U}^*_n(\boldsymbol{\theta}) = \mathbf{0} \qquad a \qquad D\mathbf{U}^*_n(\boldsymbol{\theta}) = n\mathbf{J}(\boldsymbol{\theta}) \ .$$

(2) Platí-li navíc pro $i, j = 1, \ldots, m$

$$E\left(\frac{f''_{ij}(Y; \boldsymbol{\theta})}{f(Y; \boldsymbol{\theta})} \right) = 0,$$

(tj. f je regulární i v 2. derivacích), pak

$$E(\mathbf{U}^{*'}_n(\boldsymbol{\theta})) = -n\mathbf{J}(\boldsymbol{\theta}),$$

kde

$$\mathbf{U}^{*'}_n(\boldsymbol{\theta}) = \left(\frac{\partial U^*_i(\boldsymbol{\theta})}{\partial \theta_j} \right)^m_{i,j=1}.$$

Následující věta uvádí asymptotické vlastnosti skórových vektorů náhodných výběrů.

VĚTA 2.6. *Mějme náhodný výběr* $\mathbf{Y}_n = (Y_1, \dots, Y_n)'$ *z rozdělení s regulární hustotou* $f \in \mathcal{F}^m_{reg}$. *Označme* $M = \{y \in \mathbb{R} : f(y;\boldsymbol{\theta}) > 0\}$. *Nechť pro všechna* $y \in M$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *a* $i, j = 1, \dots, m$ *existují druhé parciální derivace hustoty* $f(y;\boldsymbol{\theta})$.

(1) Pak platí

$$\frac{1}{\sqrt{n}}\mathbf{U}^*_n(\boldsymbol{\theta}) \quad \overset{A}{\sim} \quad N_m(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}))$$

nebo ekvivalentně

$$\mathbf{U}^*_n(\boldsymbol{\theta}) \quad \overset{A}{\sim} \quad N_m(\mathbf{0}, \mathbf{J}_n(\boldsymbol{\theta})).$$

Dále platí

$$\frac{1}{n}\mathbf{U}^*_n(\boldsymbol{\theta})'\mathbf{J}(\boldsymbol{\theta})^{-1}\mathbf{U}^*_n(\boldsymbol{\theta}) \quad \overset{A}{\sim} \quad \chi^2(m)$$

nebo

$$\mathbf{U}^*_n(\boldsymbol{\theta})'\mathbf{J}_n(\boldsymbol{\theta})^{-1}\mathbf{U}^*_n(\boldsymbol{\theta}) \quad \overset{A}{\sim} \quad \chi^2(m).$$

(2) Platí-li navíc, že f *je regulární i v 2.derivacích, tj.*

$$E\left(\frac{f''_{ij}(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta})}\right) = 0,$$

pak matice náhodných veličin

$$\frac{1}{n}\mathbf{U}^{*'}_n(\boldsymbol{\theta}) = \frac{1}{n}\left(\frac{\partial U^*_i(\boldsymbol{\theta})}{\partial\theta_j}\right)^m_{i,j=1} = \frac{1}{n}\left(\frac{\partial^2 l_n(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i\partial\theta_j}\right)^m_{i,j=1} \quad \xrightarrow{s.j.} \quad -\mathbf{J}(\boldsymbol{\theta}),$$

nebo ekvivalentně

$$\mathbf{U}^{*'}_n(\boldsymbol{\theta}) \quad \xrightarrow{s.j.} \quad -\mathbf{J}_n(\boldsymbol{\theta}).$$

V dalším budeme uvažovat pouze regulární hustoty, tj. $f(\mathbf{y};\boldsymbol{\theta}) \in \mathcal{F}^m_{\text{reg}}$, $\mathbf{y} \in \mathbb{R}^n$.

DEFINICE 2.7.

(a) **Věrohodnostní funkcí** rozumíme funkci vektorového parametru $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta})$$

(b) **logaritmickou věrohodnostní funkcí** nazýváme funkci

$$l(\boldsymbol{\theta};\mathbf{y}) = \ln L(\boldsymbol{\theta};\mathbf{y})$$

(c) Řekneme, že odhad $\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \widehat{\boldsymbol{\theta}}_{\text{MLE}}(\mathbf{Y})$ je **maximálně věrohodný odhad (MLE)** vektorového parametru $\boldsymbol{\theta}$, pokud platí

$$L(\widehat{\boldsymbol{\theta}}_{\text{MLE}};\mathbf{Y}) \geq L(\boldsymbol{\theta};\mathbf{Y})$$

pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

V poslední větě této části uveďme důležité vlastnosti, které se týkají asymptotického rozdělení maximálně věrohodných odhadů.

VĚTA 2.8. *Mějme náhodný výběr* $\mathbf{Y}_n = (Y_1, \dots, Y_n)'$ *z rozdělení s regulární hustotou* $f \in \mathcal{F}^m_{reg}$. *Označme* $M = \{y \in \mathbb{R} : f(y;\boldsymbol{\theta}) > 0\}$. *Nechť pro všechna* $y \in M$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *a* $i, j = 1, \dots, m$ *existují druhé parciální derivace hustoty* $f(y;\boldsymbol{\theta})$ *a platí* $E\left(\frac{f''_{ij}(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta})}\right) = 0$. *Pak*

(1) $\widehat{\boldsymbol{\theta}}_{\text{MLE}} \quad \overset{A}{\sim} \quad N_m(\boldsymbol{\theta}, \mathbf{J}_n(\boldsymbol{\theta})^{-1})$ *nebo ekvivalentně* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}) \quad \overset{A}{\sim} \quad N_m(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta})^{-1})$

(2) $W = (\widehat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta})'\mathbf{J}_n(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}) \quad \overset{A}{\sim} \quad \chi^2(m)$ *, tzv.* ***Waldova statistika.***

2.2. Exponenciální třída rozdělení pravděpodobností. Přirozenou třídou hustot, se kterými budeme dále pracovat, je třída hustot exponenciálního typu. Uveďme nejprve její definici.

DEFINICE 2.9. Řekneme, že pozorování pochází z rozdělení **exponenciálního typu**, pokud jeho *pravděpodobnostní funkce* (v případě diskrétních rozdělení) či *hustota* (v případě spojitých rozdělení) je tvaru

$$f(y) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

kde

θ je (neznámý) tzv. **přirozený parametr**

a

$a(y), b(\theta), c(\theta), d(y)$ jsou známé funkce.

Pokud

- $a(y) = y$, říkáme že pravděpodobnostní funkce, popř. hustota je v **kanonické formě**.
- v konkrétním rozdělení figurují další neznámé parametry, nazveme je tzv. **rušivými parametry**.

V dalším budeme uvažovat pouze **regulární** a **kanonické** formy spolu s podmínkou $b(\theta) = \theta$ a přitom zavedeme do označení jeden rušivý parametr ϕ :

$$\boxed{f(y) = \exp\left\{\frac{y\theta - \gamma(\theta)}{\psi(\phi)} + d(y, \phi)\right\}},$$

kde θ a ϕ jsou parametry
$\gamma(\theta),\ \psi(\phi) > 0,\ d(y)$ jsou známé funkce,

a pokud

$\psi(\phi) = \frac{\phi}{\omega} > 0$, $\phi > 0$ je tzv. **faktor měřítka** (scale factor)
$\omega > 0$ je známá **apriorní váha**.

Tato forma se také nazývá **škálovou formou hustoty exponenciálního typu**.

Poznámka 2.10. Jestliže neplatí $b(\theta) = \theta$, stačí provést jednoduchou reparametrizaci a zavést případně nový parametr

$$\theta^* = b(\theta),$$

který se pak nazývá **kanonickým parametrem**.

LEMMA 2.11. *Mějme náhodnou veličinu Y z rozdělení s regulární hustotou f exponenciálního typu:*

$$f(y) = \exp\left\{\frac{y\theta - \gamma(\theta)}{\psi(\phi)} + d(y, \phi)\right\}. \tag{20}$$

Pak

$$EY = \gamma'(\theta)$$

Nechť navíc platí

$$E\left(\frac{f''(Y;\theta)}{f(Y;\theta)}\right) = 0, \tag{21}$$

kde $f''(Y;\theta) = \frac{d^2 f(Y;\theta)}{d\theta^2}$, pak

$$DY = \gamma''(\theta)\psi(\phi)$$

*Funkce $\gamma''(\theta) = \frac{DY}{\psi(\phi)}$ se nazývá **rozptylovou funkcí** (variance function).*

Příklady rozdělení exponenciálního typu

Příklad 2.12 (Normální rozdělení)**.** Mějme

$$Y \sim N(\mu, \sigma^2), \qquad \mu \in \mathbb{R},\ \sigma^2 > 0.$$

Pak $$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} = \exp\left\{ \underbrace{\frac{y\mu - \overbrace{\frac{1}{2}\mu^2}^{\gamma(\theta)}}{\sigma^2}}_{\psi(\phi)} \underbrace{- \frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)}_{d(y,\phi)} \right\}$$

a
$$\begin{array}{lll} \gamma(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2 & \Rightarrow & \gamma'(\theta) = \theta = \mu \\ & \Rightarrow & \gamma''(\theta) = 1 \\ \psi(\phi) = \sigma^2 & \Rightarrow & \phi = \sigma^2. \end{array}$$

Skutečně platí

$$EY = \gamma'(\theta) = \mu$$

a

$$DY = \gamma''(\theta)\psi(\phi) = \sigma^2.$$

Tedy ***přirozený parametr*** $\theta = \mu$ ***scale factor*** $\phi = \sigma^2$
rozptylová funkce $V(\mu) = 1$ ***váhy*** $\omega = 1.$

<u>Příklad</u> 2.13 (Binomické rozdělení)**.** Mějme

$$Z = nY \sim Bi(n, \pi), \qquad n \in \mathbb{N}, \pi \in (0,1).$$

pak $f_Z(z) = \binom{n}{z}\pi^z(1-\pi)^{n-z} = \exp\left\{z\ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + \ln\binom{n}{z}\right\}$
pro $z = 0, \ldots, n,$
přičemž $EZ = \mu = n\pi$ a $DZ = n\pi(1-\pi).$

Pravděpodobnostní funkce **není** ve škálové formě, proveďme reparametrizaci

$$\theta = \ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{n\pi}{n-n\pi}\right) = \ln\left(\frac{\mu}{n-\mu}\right) \quad \Rightarrow \quad \pi = \frac{e^\theta}{1+e^\theta} \text{ a } 1-\pi = \frac{1}{1+e^\theta}.$$

Tedy $$f_Z(z) = \exp\left\{ z\theta - \underbrace{n\ln\left(1+e^\theta\right)}_{\gamma(\theta)} + \underbrace{\ln\binom{n}{z}}_{d(y,\phi)} \right\}$$

a
$$\begin{array}{lll} \gamma(\theta) = n\ln\left(1+e^\theta\right) & \Rightarrow & \gamma'(\theta) = n\frac{e^\theta}{1+e^\theta} = n\pi = \mu \\ & \Rightarrow & \gamma''(\theta) = n\frac{e^\theta}{(1+e^\theta)^2} = n\pi(1-\pi) = \mu\left(1-\frac{\mu}{n}\right) \\ \psi(\phi) = \frac{\phi}{\omega} = 1 & \Rightarrow & \omega = 1 \quad \phi = 1 \end{array}$$

Skutečně platí

$$EZ = \gamma'(\theta) = \mu$$

a

$$DZ = \gamma''(\theta)\psi(\phi) = n\pi(1-\pi).$$

Tedy ***přirozený parametr*** $\theta = \ln\left(\frac{\mu}{n-\mu}\right)$
rozptylová funkce $V(\mu) = \mu\left(1-\frac{\mu}{n}\right)$
scale factor $\phi = 1$
váhy $\omega = 1.$

Poznámka 2.14. Je třeba poznamenat, že ve vztahu

$$\gamma(\theta) = n \ln\left(1 + e^{\theta}\right)$$

se vedle parametru θ vyskytuje i parametr n, který je však vždy znám. Abychom dostáli úvodní definici (20), proto ho nebudeme považovat za rušivý parametr, ale za známou konstantu. Tomuto problému se lze vyhnout, když přejdeme od absolutních četností Z k relativním četnostem Y. V případě, že uvažuje místo absolutních **relativní četnosti**: $Y = \frac{Z}{n}$

je pravděpodobnostní funkce nenulová pro $y \in \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ a je tvaru

$$f_Y(y) = \binom{n}{ny}\pi^{ny}(1-\pi)^{n-ny} = \exp\left\{\frac{y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)}{1/n} + \ln\binom{n}{ny}\right\}$$

a $\quad EY = \mu = \pi \quad$ a $\quad DY = \frac{\pi(1-\pi)}{n}$.

Pravděpodobnostní funkce **není** ve škálové formě, proveďme reparametrizaci

$$\theta = \ln\left(\tfrac{\pi}{1-\pi}\right) \quad \Rightarrow \quad \pi = \tfrac{e^{\theta}}{1+e^{\theta}} \text{ a } 1 - \pi = \tfrac{1}{1+e^{\theta}}$$

Tedy $\quad f_Y(y) = \exp\left\{\dfrac{y\theta - \overbrace{\ln\left(1+e^{\theta}\right)}^{\gamma(\theta)}}{\underbrace{1/n}_{\psi(\phi)}} + \underbrace{\ln\binom{n}{ny}}_{d(y,\phi)}\right\}$

a $\quad \gamma(\theta) = \ln\left(1+e^{\theta}\right) \quad \Rightarrow \quad \gamma'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \pi = \mu$

$\qquad\qquad\qquad\quad \Rightarrow \quad \gamma''(\theta) = \frac{e^{\theta}}{(1+e^{\theta})^2} = \pi(1-\pi)$

$\qquad \psi(\phi) = \frac{\phi}{\omega} = \frac{1}{n} \quad \Rightarrow \quad \omega = n \quad \phi = 1$

Skutečně platí $\quad EY = \gamma'(\theta) = \mu$

$$DY = \gamma''(\theta)\psi(\phi) = \tfrac{\pi(1-\pi)}{n},$$

Tedy	***přirozený parametr***	$\theta = \ln\left(\frac{\mu}{1-\mu}\right)$
	rozptylová funkce	$V(\mu) = \mu(1-\mu)$
	scale factor	$\phi = 1$
	váhy	$\omega = n.$

Příklad 2.15 (Poissonovo rozdělení)**.** Mějme

$$Y \sim Po(\lambda), \qquad \lambda > 0$$

pak $\quad f(y) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp\{y \ln\lambda - \lambda - \ln y!\} \qquad y = 0, 1, 2, \ldots$

přičemž $\quad EY = \mu = \lambda \quad$ a $\quad DY = \lambda.$

Protože pravděpodobnostní funkce **není** ve škálové formě, proveďme reparametrizaci

$$\theta = \ln\lambda \quad \Rightarrow \quad \lambda = e^{\theta}$$

a $$f(y) = \exp\left\{ y\theta - \underbrace{e^\theta}_{\gamma(\theta)} - \underbrace{\ln y!}_{d(y,\phi)} \right\}$$

$$\gamma(\theta) = e^\theta \quad \Rightarrow \quad \gamma'(\theta) = e^\theta = \lambda = \mu$$
$$\Rightarrow \quad \gamma''(\theta) = e^\theta = \lambda = \mu$$

$$\psi(\phi) = \frac{\phi}{\omega} = 1 \quad \Rightarrow \quad \omega = 1 \quad \phi = 1$$

Skutečně platí

$$EY = \gamma'(\theta) = \lambda = \mu$$

a

$$DY = \gamma''(\theta)\psi(\phi) = e^\theta = \lambda = \mu,$$

Tedy ***přirozený parametr*** $\theta = \ln \lambda$

rozptylová funkce $V(\mu) = \mu$

scale factor $\phi = 1$

váhy $\omega = 1.$

<u>Příklad</u> 2.16 (Gamma rozdělení)**.** Mějme

$$Y \sim G(\alpha, \beta), \qquad \alpha > 0, \beta > 0.$$

Pak $f(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}} \qquad y > 0$

přičemž $EY = \alpha\beta$ a $DY = \alpha\beta^2.$

Tento tvar hustoty je však pro nás nevhodný (ve střední hodnotě máme rušivý parametr α), proto uvažujme reparametrizaci

$$\mu = \alpha\beta \quad \Rightarrow \quad \beta = \frac{\mu}{\alpha},$$

pak $$f(y) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha y^{\alpha-1} e^{-\frac{\alpha}{\mu}y} = \exp\left\{ \frac{y\left(-\frac{1}{\mu}\right) - \ln \mu}{\frac{1}{\alpha}} + \alpha \ln \alpha + (\alpha - 1)\ln y - \ln \Gamma(\alpha) \right\}$$

přičemž $EY = \mu$ a $DY = \frac{\mu^2}{\alpha}.$

Hustota není v kanonickém tvaru, proto parametrizujme $\theta = -\frac{1}{\mu} \quad \Rightarrow \quad \mu = -\frac{1}{\theta}$

pak $$f(y) = \exp\left\{ \frac{y\theta - \overbrace{\ln(-\frac{1}{\theta})}^{\gamma(\theta)}}{\frac{1}{\alpha}} + \underbrace{\alpha \ln \alpha + (\alpha - 1)\ln y - \ln \Gamma(\alpha)}_{d(y,\phi)} \right\}$$

a $$\gamma(\theta) = \ln(\mu) = \ln(-\tfrac{1}{\theta}) \quad \Rightarrow \quad \gamma'(\theta) = -\frac{\theta}{\theta^2} = -\frac{1}{\theta} = \mu$$
$$\Rightarrow \quad \gamma''(\theta) = \frac{1}{\theta^2} = \mu^2$$

$$\psi(\phi) = \frac{\phi}{\omega} = \frac{1}{\alpha} \quad \Rightarrow \quad \omega = 1 \quad \phi = \frac{1}{\alpha}$$

Skutečně platí

$$EY = \gamma'(\theta) = \mu$$

a

$$DY = \gamma''(\theta)\psi(\phi) = \frac{\mu^2}{\alpha} = \frac{\alpha^2\beta^2}{\alpha} = \alpha\beta^2$$

Tedy ***přirozený parametr*** $\theta = -\frac{1}{\mu}$

rozptylová funkce	$V(\mu) = \mu^2$
scale factor	$\phi = \frac{1}{\alpha}$
váhy	$\omega = 1.$

Příklad 2.17 (Exponenciální rozdělení)**.** Exponenciální rozdělení je speciálním případem gamma rozdělení

$$Y \sim Ex(\lambda) \equiv G(1, \lambda).$$

Pak $f(y) = \frac{1}{\lambda} e^{-\frac{y}{\lambda}} = \exp\{-\frac{1}{\lambda} y - \ln \lambda\} \qquad y > 0,$

přičemž $EY = \mu = \lambda \quad \text{a} \quad DY = \lambda^2.$

Hustota není v kanonickém tvaru, proto parametrizujme

$$\theta = -\frac{1}{\lambda} \quad \Rightarrow \quad \lambda = -\frac{1}{\theta}$$

Tedy

$$f(y) = \exp\left\{ y\theta - \overbrace{\ln(-\frac{1}{\theta})}^{=\gamma(\theta)} \right\}$$

a

$$\begin{aligned} \gamma(\theta) = \ln(\mu) = \ln(-\tfrac{1}{\theta}) \quad &\Rightarrow \quad \gamma'(\theta) = -\tfrac{1}{\theta} = \lambda = \mu \\ &\Rightarrow \quad \gamma''(\theta) = \tfrac{1}{\theta^2} = \lambda^2 = \mu^2 \\ \psi(\phi) = \tfrac{\phi}{\omega} = 1 \quad &\Rightarrow \quad \omega = 1 \quad \phi = 1 \end{aligned}$$

Skutečně platí

$$EY = \gamma'(\theta) = \mu$$

a

$$DY = \gamma''(\theta)\psi(\phi) = \mu^2 = \lambda^2$$

tj. jde o regulární systém hustot a navíc platí podmínka (21).

Tedy

přirozený parametr	$\theta = -\frac{1}{\mu}$
rozptylová funkce	$V(\mu) = \mu^2$
scale factor	$\phi = 1$
váhy	$\omega = 1.$

3. Definice jednorozměrného GLM

3.1. Omezení klasického lineárního regresního modelu. Mějme klasický lineární regresní model plné hodnosti

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ \wedge \ h(\mathbf{X}) = h(\mathbf{X}'\mathbf{X}) = k \ \wedge \ n > k \ \wedge \ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

kde

$\mathbf{Y} = (Y_1, \ldots, Y_n)'$	je vektor závisle proměnných,
$\mathbf{X} = (x_{ij})$	je matice plánu, $(i = 1, \ldots, n;\ j = 1, \ldots, k)$
$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$	je vektor chyb, přičemž $E\boldsymbol{\varepsilon} = \mathbf{0}$; $D\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}_n$.

Když se podíváme na tento model blíže, zjistíme, že se skládá ze dvou částí:

Systematická (signální) část vyjadřuje lineární vztah pro střední hodnotu a neznámé parametry β_j, tj.

$$EY_i = \mu_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

Tato část je obvykle cílem zkoumání, snažíme se pomocí ní maximálně možně vysvětlit chování náhodné veličiny Y_i a zjistit skrze parametry β_j velikost a znaménko závislosti na vysvětlujících veličinách $\mathbf{x}_i$.

V reálném světě má mnoho procesů jiný, než lineární vztah závislosti. Např. v ekonomii se ukazuje, že mnoho vztahů má logaritmickou závislost, k vysvětlení procesů v přírodních vědách se užívají reciproké, mocninné i další vztahy. Vysvětlovaná veličina popisující pravděpodobnost přežití člověka, v případě určité nemoci a určitého způsobu léčby, může z definice pravděpodobnosti nabývat hodnot pouze z intervalu $[0, 1]$, což by v případě klasického lineárního modelu bylo možné zajistit jen za přijetí omezení na estimátor $\boldsymbol{\beta}$.

Náhodná část je reprezentovaná náhodnými chybami ε_i, které shrnují v sobě všechny ostatní vlivy, působící na Y_i, kromě již uvedených v systematické části. Rozdělení náhodných veličin ε_i je závislé na rozdělení Y_i a má tvar

$$\varepsilon_i \sim N(0, \sigma^2),$$

kde ε_i jsou nezávislé. Právě **normalita chyb** je často nesplněným předpokladem klasického lineárního regresního modelu. Připomeňme, že normalita se vyznačuje nezávislosti střední hodnoty a rozptylu. Typicky např. u ekonomických veličin s rostoucí střední hodnotou obvykle roste rozptyl náhodné veličiny, přičemž náhodné chyby mají v těchto případech často nesymetrická, kladně sešikmená rozdělení.

Shrneme-li předchozí, můžeme říci, že **klasický lineární regresní model** je sice velmi důležitým stochastickým modelem, avšak má celou řadu omezení:

- Je **omezen pouze na třídu normálních rozdělení**: $Y_i \sim N(\mu_i, \sigma^2) \quad i = 1, \ldots, n$, kde $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ tvoří náhodný výběr.
- Předpokládá **striktní rovnost mezi střední hodnotou náhodné veličiny Y_i a lineární kombinací prediktorů**: $EY_i = \mu_i = \mathbf{x}_i'\boldsymbol{\beta}$, kde

  $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ je vektor prediktorů a
  $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ je vektor neznámých parametrů.

Je však možné provést **zobecnění** tohoto klasického lineárního modelu dvěma směry:

(1) Zobecnění na nenormální rozdělení, a to na tzv. **třídu exponenciálních rozdělení**

(2) Zobecnění na nelineární funkce, které **spojují** neznámé střední hodnoty výchozího rozdělení náhodné veličiny Y_i s prediktivními proměnnými.

3.2. Definice jednorozměrného GLM. Předchozí pasáž nám poskytla motivaci pro hledání obecnějšího modelu, než je model lineární. Uveďme nyní již samotnou definici zobecněného lineárního modelu a poté několik příkladů pro lepší názornost.

DEFINICE 3.1 (Zobecněný lineární model). Mějme náhodný výběr

$$\mathbf{Y} = (Y_1, \ldots, Y_n)'$$

a nechť rozdělení Y_i závisí na pevných vektorech

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})' \in \mathbb{R}^k$$

prostřednictvím neznámého vektoru parametrů

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'.$$

Matice

$$\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$$

má rozměr $n \times k$ a hodnost $k < n$.
Říkáme, že $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ se řídí **zobecněným lineárním modelem** (Generalized Linear Model), jestliže dále platí:

(1) rozdělení $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ je exponenciálního typu s **regulární** hustotou tvaru

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \theta_i) = \exp\left\{ \sum_{i=1}^{n} \left[\frac{y_i\theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right] \right\} \tag{22}$$

(2) parametr θ_i závisí na $\mathbf{x}_i$ a $\boldsymbol{\beta}$ prostřednictvím parametru

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}\ , \tag{23}$$

který nazveme **lineární prediktor**.

(3) Existuje známá ryze monotónní diferencovatelná funkce g , tzv. **linkovací funkce** (link function), a platí

$$\eta_i = g(\mu_i) \qquad \mu_i = g^{-1}(\eta_i)\ , \quad \text{kde} \quad \mu_i = \mu(\theta_i) = EY_i. \tag{24}$$

Řekneme, že linkovací funkce je **kanonická**, pokud $\theta_i = \eta_i = g(\mu_i)$.

Matici $\mathbf{X} = (\mathbf{x}_i')_{i=1}^n$ nazýváme **maticí plánu**.

Příklad 3.2. Regresní přímka v klasickém lineárním regresním modelu:

$$Y_i \sim N(\mu_i, \sigma^2)$$

jsou pro $i = 1, \ldots, n$ nezávislé náhodné veličiny,

$$g(\mu_i) = \mu_i = \beta_1 + \beta_2 x_i$$

je identická linkovací funkce, β_1, β_2 a σ^2 jsou neznámé parametry (přičemž σ^2 je rušivým parametrem) a x_i jsou známé kovariáty.

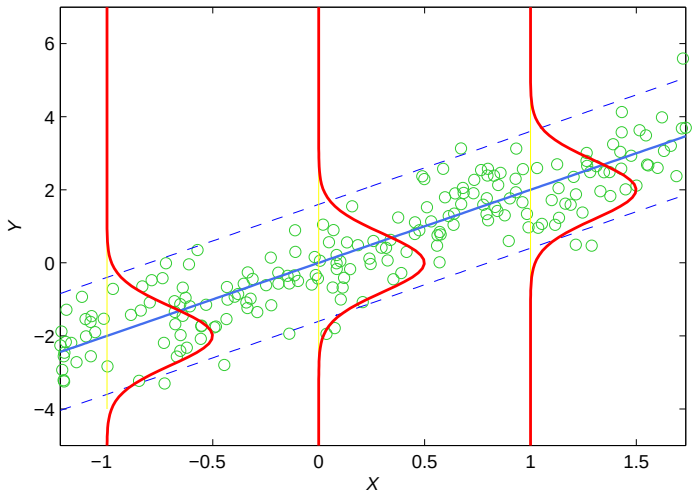
Příklad 3.3. Regresní modely s logaritmickou linkovací funkcí pro exponenciálně a gamma rozdělené závisle proměnné:

$$Y_i \sim Ex(\lambda_i) \equiv G(1, \lambda_i)$$

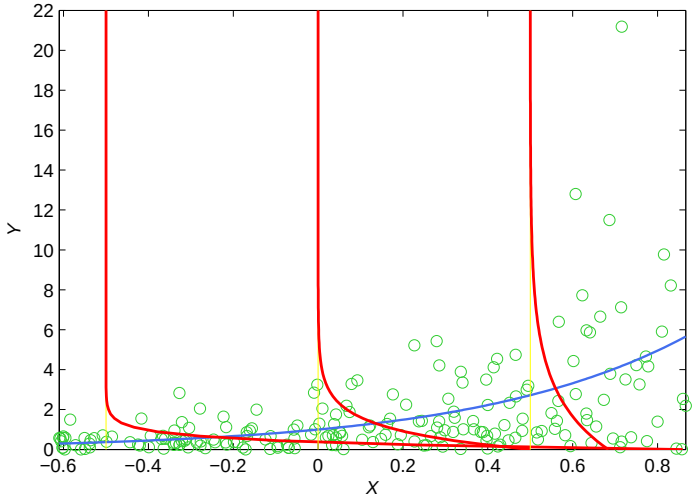
jsou pro $i = 1, \ldots, n$ nezávislé náhodné veličiny ($EY_i = \mu_i = \lambda_i$),

$$g(\mu_i) = \ln \mu_i = \beta_1 + \beta_2 x_i$$

je logaritmická linkovací funkce, β_1, β_2 jsou neznámé parametry a x_i jsou známé kovariáty.



Obrázek 1: Ukázka klasického regresního modelu s homogenním rozptylem.



Obrázek 2: Ukázka GLM modelu s linkovací funkcí $g(\mu) = \ln \mu$ pro exponenciálně rozdělenou náhodnou veličinu Y.

Jestliže $Y_i \sim G(\alpha, \beta_i = \frac{\mu_i}{\alpha})$ jsou pro $i = 1, \ldots, n$ nezávislé náhodné veličiny ($EY_i = \mu_i = \alpha\beta_i$), $g(\mu_i) = \ln \mu_i = \beta_1 + \beta_2 x_i$ je logaritmická linkovací funkce, β_1, β_2 a $\alpha = \frac{1}{\phi}$ jsou neznámé parametry (α je rušivý parametr) a x_i jsou známé kovariáty.

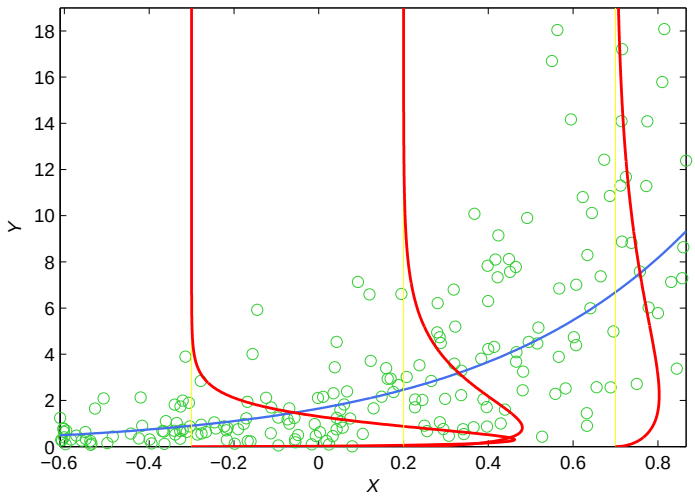
<u>Příklad</u> 3.4. Poissonovská regrese:

$$Y_i \sim Po(\mu_i)$$

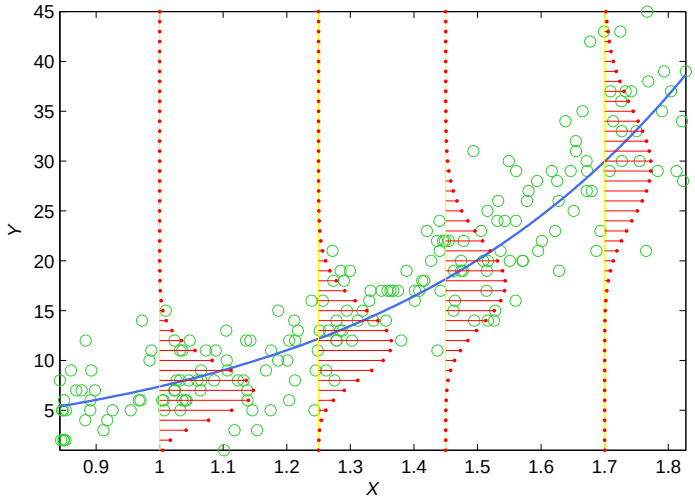
jsou pro $i = 1, \ldots, n$ nezávislé náhodné veličiny ($EY_i = \mu_i$),

$$g(\mu_i) = \ln \mu_i = \beta_1 + \beta_2 x_i$$

je logaritmická linkovací funkce, β_1, β_2 jsou neznámé parametry a x_i jsou známé kovariáty.



Obrázek 3: Ukázka GLM modelu s linkovací funkcí $g(\mu) = \ln \mu$ pro náhodnou veličinu Y s gamma rozdělením.



Obrázek 4: Ukázka poissonovské regrese s linkovací funkcí $g(\mu) = \ln \mu$.

<u>Příklad</u> 3.5. Binomická regrese:

$$Y_i \sim Bi(n_i, \pi_i)$$

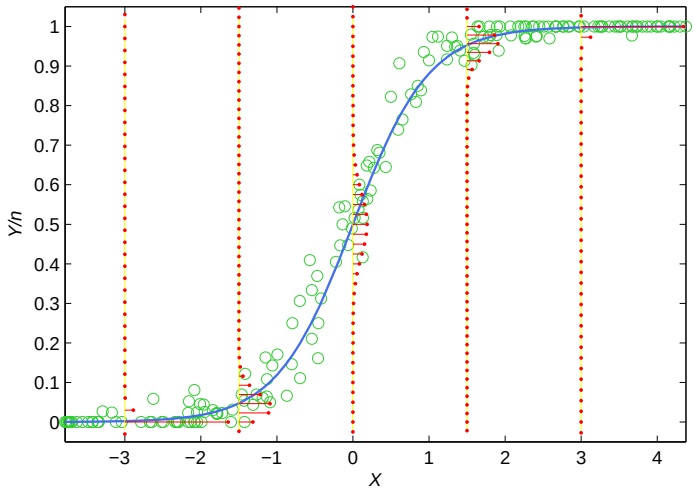
jsou pro $i = 1, \ldots, n$ nezávislé náhodné veličiny, kde

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

je logistická linkovací funkce, β_1, β_2 jsou neznámé parametry a x_i jsou známé kovariáty.

Například ve farmaceutickém experimentu může být n_i počet pacientů, kterým byla podána dávka x_i nového léku a Y_i počet pacientů dávající pozitivní odpověď na danou dávku x_i nového léku.

Jestliže pozorujeme, že $\frac{Y_i}{n_i}$ roste spolu s x_i, hledáme model, ve kterém π_i je funkcí x_i, hodnot $0 < \pi_i < 1$. Proto model $\pi_i = \beta_1 + \beta_2 x_i$ není vhodný, avšak $\beta_1 + \beta_2 x_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ obvykle pracuje dobře.



Obrázek 5: Ukázka binomické regrese s linkovací funkcí $g(\pi) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$.

Příklad 3.6. Kontingenční tabulky:

$$Y_{ij} \sim Mn(n, \pi_{ij})$$

jsou pro $i = 1, \ldots, I,\ j = 1, \ldots, J,\ n = I \cdot J,\ \left(\sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = 1\right)$ nezávislé náhodné veličiny, například počet lidí i-té etnické skupiny, kteří volí politickou stranu j.

Snahou bude testovat hypotézu $H_0 : \pi_{ij} = \alpha_i\beta_j$ pro všechna i, j, kde α_i, β_j jsou neznámé parametry, $\sum_{i=1}^{I} \alpha_i = 1$ a $\sum_{j=1}^{J} \beta_j = 1$, tj. chceme testovat hypotézu, že volba strany a etnická příslušnost jsou nezávislé.

Připomeňme, že	$\mu_{ij} = EY_{ij} = n\pi_{ij}$
takže	$\ln \frac{Y_{ij}}{n} = \ln \pi_{ij}$
a za platnosti hypotézy H_0	$\ln \pi_{ij} = \ln \alpha_i + \ln \beta_j$
ekvivalentně	$g(\mu_{ij}) = \ln \mu_{ij} = const_{ij} + a_i + b_j$ pro nějaké $const_{ij}, a_i, b_j$.

4. Odhady neznámých parametrů v GLM

4.1. Maximálně věrohodné odhady. Všimněme si, že rozdělení náhodných veličin Y_i jsou stejného typu a **logaritmus sdružené věrohodnostní funkce** má tvar

$$l^*(\boldsymbol{\theta}; \mathbf{y}) = l^*(\theta_1, \ldots, \theta_n; y_1, \ldots, y_n) = \sum_{i=1}^{n} l_i(\theta_i; y_i) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right).$$

VĚTA 4.1. *Mějme náhodný výběr* $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, *který se řídí* ***zobecněným lineárním modelem*** *s linkovací funkcí*

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \eta_i \qquad i = 1, \ldots, n.$$

Předpokládejme, že pro $i = 1, \ldots, n$ *existují příslušné derivace* $\gamma'(\theta_i), \gamma''(\theta_i)$ *a platí*

$$EY_i = \mu_i = \gamma'(\theta_i) \qquad DY_i = \gamma''(\theta_i)\psi_i(\phi).$$

Pak

$$U_j^* = U_j^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{x_{ij}(Y_i - \mu_i)}{DY_i} \frac{\partial \mu_i}{\partial \eta_i} \tag{25}$$

a

$$J_{jk}^* = J_{jk}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{DY_i} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \tag{26}$$

což lze zapsat maticově

$$\mathbf{U}_n^* = \mathbf{U}_n^*(\boldsymbol{\beta}) = (U_1^*, \ldots, U_m^*)' = \mathbf{X}'\mathbf{W}(\boldsymbol{\beta})\mathbf{Q}(\boldsymbol{\beta})\mathbf{r}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{Q}\mathbf{r} \tag{27}$$

a

$$\mathbf{J}_n = \mathbf{J}_n(\boldsymbol{\beta}) = \left(J_{jk}^*\right)_{j,k=1}^{m} = \mathbf{X}'\mathbf{W}(\boldsymbol{\beta})\mathbf{X} = \mathbf{X}'\mathbf{W}\mathbf{X}\,, \tag{28}$$

kde

$$\mathbf{r} = \mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))' \qquad r_i = r_i(\boldsymbol{\beta}) = Y_i - \mu_i = Y_i - g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

$$\mathbf{W} = \mathbf{W}(\boldsymbol{\beta}) = diag\{w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta})\} \qquad w_i = w_i(\boldsymbol{\beta}) = \frac{1}{DY_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

$$\mathbf{Q} = \mathbf{Q}(\boldsymbol{\beta}) = diag\{q_1(\boldsymbol{\beta}), \ldots, q_n(\boldsymbol{\beta})\} \qquad q_i = q_i(\boldsymbol{\beta}) = \frac{\partial \eta_i}{\partial \mu_i}.$$

Odhad neznámých parametrů metodou **maximální věrohodnosti** dostaneme řešením rovnic typu

$$\frac{\partial l^*}{\partial \boldsymbol{\beta}} = \mathbf{U}_n^*(\boldsymbol{\beta}) = \mathbf{0}$$

Aby šlo o maximum, je nutné, aby matice druhých parciálních derivací logaritmické věrohodnostní funkce podle složek parametru $\boldsymbol{\beta}$ byla negativně definitní.

Podle věty 2.6 konverguje matice druhých parciálních derivací skoro jistě k matici $-\mathbf{J}_n$, která je při regularitě systému hustot negativně definitní.

Aproximujeme-li proto matici $\frac{\partial \mathbf{U}_n^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ maticí $-\mathbf{J}_n$, je řešení systému předešlých rovnic maximálně věrohodným odhadem parametru $\boldsymbol{\beta}$.

Nyní se vraťme k řešení věrohodnostních rovnic. Protože obecně rovnice

$$U_j^* = \frac{\partial l^*}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}(Y_i - \mu_i)}{DY_i} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad j = 1, \ldots, m.$$

nejsou lineární vzhledem k neznámým parametrům, musí se řešit numerickou iterací.

4.2. Newtonova – Raphsonova metoda. Chceme-li najít řešení systému nelineárních rovnic $\mathbf{U}_n^*(\boldsymbol{\beta}) = \mathbf{0}$, lze použít následující iterativní postup:

(1) Nejprve provedeme **linearizaci pomocí Taylorova rozvoje** v okolí bodu $\boldsymbol{\beta}_0$, kde $\boldsymbol{\beta}_0$ je nějaký počáteční odhad: $\mathbf{U}_n^*(\boldsymbol{\beta}) \approx \mathbf{U}_n^*(\boldsymbol{\beta}_0) + \mathbf{U}_n^{*\prime}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Protože $\mathbf{U}_n^*(\boldsymbol{\beta}) = 0$, pak po jednoduchých úpravách dostaneme $\boldsymbol{\beta} \approx \boldsymbol{\beta}_0 - \left[\mathbf{U}_n^{*\prime}(\boldsymbol{\beta}_0)\right]^{-1} \mathbf{U}_n^*(\boldsymbol{\beta}_0)$.

(2) Odhady parametrů v s-tém kroku jsou získány ze vztahu

$$\widehat{\boldsymbol{\beta}}^{(s)} = \widehat{\boldsymbol{\beta}}^{(s-1)} - \left[\mathbf{U}_n^{*\prime}\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right)\right]^{-1}\mathbf{U}_n^*\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right).$$

(3) Iterační proces popsaný v předchozím bodě pokračuje tak dlouho, dokud

$$\widehat{\boldsymbol{\beta}}^{(s+1)} - \widehat{\boldsymbol{\beta}}^{(s)} \approx 0.$$

4.3. Metoda skórování. Alternativní procedurou k Newtonově – Raphsonově metodě je tzv. **metoda skórování**, kdy se matice druhých parciálních derivací $\mathbf{U}_n^{*\prime}(\boldsymbol{\beta})$ nahradí její střední hodnotou, tj. maticí $-\mathbf{J}_n(\boldsymbol{\beta})$, kde $\mathbf{J}_n(\boldsymbol{\beta})$, je **informační matice**.

Druhý iterační krok pak upravíme takto:

$$\widehat{\boldsymbol{\beta}}^{(s)} = \widehat{\boldsymbol{\beta}}^{(s-1)} + \left[\mathbf{J}_n(\widehat{\boldsymbol{\beta}}^{(s-1)})\right]^{-1}\mathbf{U}_n^*(\widehat{\boldsymbol{\beta}}^{(s-1)}).$$

Jednoduchou úpravou dostaneme

$$\mathbf{J}_n(\widehat{\boldsymbol{\beta}}^{(s-1)})\widehat{\boldsymbol{\beta}}^{(s)} = \mathbf{J}_n(\widehat{\boldsymbol{\beta}}^{(s-1)})\widehat{\boldsymbol{\beta}}^{(s-1)} + \mathbf{U}_n^*(\widehat{\boldsymbol{\beta}}^{(s-1)}).$$

Využijme vztahů:

$$\mathbf{U}_n^*(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}(\boldsymbol{\beta})\mathbf{Q}(\boldsymbol{\beta})\mathbf{r}(\boldsymbol{\beta}) \quad \text{a} \quad \mathbf{J}_n(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}(\boldsymbol{\beta})\mathbf{X}$$

a dále upravujme

$$\begin{aligned}\mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s)} &= \mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s-1)} + \mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{Q}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{r}(\widehat{\boldsymbol{\beta}}^{(s-1)}) \\ &= \mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\underbrace{\left[\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s-1)} + \mathbf{Q}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{r}(\widehat{\boldsymbol{\beta}}^{(s-1)})\right]}_{\mathbf{z}(\widehat{\boldsymbol{\beta}}^{(\mathbf{s-1})})}\end{aligned}$$

přičemž pro $i = 1, \ldots, n$

$$Z_i(\widehat{\boldsymbol{\beta}}^{(s-1)}) = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}^{(s-1)} + r_i^{(s-1)}q_i^{(s-1)} = \sum_{j=1}^{m} x_{ij}\hat{\beta}_j^{(s-1)} + \left(Y_i - \hat{\mu}_i^{(s-1)}\right)\frac{\partial\hat{\eta}_i^{(s-1)}}{\partial\hat{\mu}_i^{(s-1)}}.$$

Můžeme psát

$$\mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s)} = \mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{Z}(\widehat{\boldsymbol{\beta}}^{(s-1)}) \text{ nebo } \mathbf{X}'\mathbf{W}^{(s-1)}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s)} = \mathbf{X}'\mathbf{W}^{(s-1)}\mathbf{Z}^{(s-1)},$$

kde

$$\mathbf{W}^{(s-1)} = diag\{w_1^{(s-1)}, \ldots, w_n^{(s-1)}\}$$

a

$$w_i^{(s-1)} = \frac{1}{\widehat{DY}_i^{(s-1)}}\left(\frac{\partial\hat{\mu}_i^{(s-1)}}{\partial\hat{\eta}_i^{(s-1)}}\right)^2 = \frac{\omega_i}{\phi V(\hat{\mu}_i^{(s-1)})\left(q_i^{(s-1)}\right)^2},$$

kde

$$\widehat{DY}_i^{(s-1)} = \psi(\phi)V\left(\hat{\mu}_i^{(s-1)}\right) = \frac{\phi}{\omega_i}V(\hat{\mu}_i^{(s-1)}).$$

Jde o obdobu vážené metody nejmenších čtverců a v tomto případě mluvíme o **iterační vážené metodě nejmenších čtverců**.

5. Testování hypotéz v GLM modelech

Statistické modely jsou obvykle konstruovány s cílem rozhodnout o předem definované hypotéze a tuto přijmout či vyvrátit.

VĚTA 5.1. *Mějme náhodný výběr* $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$, *který se řídí* ***zobecněným lineárním modelem*** *s maticí vysvětlujících proměnných* $\mathbf{X}_{n\times k}$. *Předpokládejme, že pro* $i = 1, \ldots, n$ *existují příslušné derivace* $\gamma'(\theta_i), \gamma''(\theta_i)$ *a platí*

$$EY_i = \mu_i = \gamma'(\theta_i) \qquad DY_i = \gamma''(\theta_i)\psi_i(\phi).$$

Dále mějme matici $\mathbf{C}_{k\times q}$ *s hodností* $h(\mathbf{C}) = q < k$. *Platí-li hypotéza:* $H_0 : \mathbf{C}'\boldsymbol{\beta} = 0$, *pak* ***Waldova statistika***

$$W = \widehat{\boldsymbol{\beta}}'_{\text{MLE}}\mathbf{C}\left(\mathbf{C}'\mathbf{J}_n(\boldsymbol{\beta})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\widehat{\boldsymbol{\beta}}_{\text{MLE}} \overset{A}{\sim} \chi^2(q),$$

kde $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ *je maximálně věrohodným odhadem vektorového parametru* $\boldsymbol{\beta}$.

Poznámka 5.2. Hypotézu

$$H_0 : \mathbf{C}'\boldsymbol{\beta} = 0$$

zamítáme na hladině významnosti α, pokud platí

$$W > \chi^2_{1-\alpha}(q).$$

Protože odhad $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ konverguje za předpokladu existence $E(l^*(\boldsymbol{\beta}))$ skoro jistě k $\boldsymbol{\beta}$, aproximujeme při výpočtu Waldovy statistiky W Fisherovou informační matici $\mathbf{J}_n(\boldsymbol{\beta})$ maticí $\mathbf{J}_n(\widehat{\boldsymbol{\beta}}_{\text{MLE}})$.

Poznámka 5.3. Testovat hypotézu

$$H_0 : \beta_j = 0$$

pro $j = 1, \ldots, k$ lze více způsoby:

- Pomocí Waldovy statistiky W, a to při speciální volbě

$$\mathbf{C} = \mathbf{c}_{k\times 1} = (0, \ldots, \overset{j}{1}, \ldots, 0)'.$$

- Pomocí vztahu

$$\hat{\beta}_{MLE,j} \overset{A}{\sim} N\left(\beta_j, s^\star_{jj} = \left(\mathbf{J}_n(\boldsymbol{\beta})^{-1}\right)_{jj}\right),$$

přičemž hypotézu **zamítáme**, pokud

$$\frac{|\hat{\beta}_{MLE,i}|}{\sqrt{s^\star_{jj}}} > u_{1-\frac{\alpha}{2}},$$

kde opět Fisherovou informační matici $\mathbf{J}_n(\boldsymbol{\beta})$ aproximujeme maticí $\mathbf{J}_n(\widehat{\boldsymbol{\beta}}_{\text{MLE}})$.

6. Ověřování vhodnosti modelu

6.1. Minimální, maximální model a submodely. Určení vhodné modelové rovnice je základem všech regresních modelů. Jedním z důležitých principů regresních modelů je **zásada jednoduchosti**, která znamená, že jednodušší model poměrně dobře popisující zkoumaná data dostane přednost před složitějším modelem, který data popisuje téměř dokonale.

Často musíme vzít také v úvahu současně se základním zobecněným lineárním modelem i několik z něj vyplývajících dílčích modelů, kterým se říká submodely.

Definujme nejprve důležité pojmy

Definice 6.1. **Maximální GLM**, který označíme GLM_{max}, splňuje následující podmínky
(1) Maximální model je zobecněný lineární model se stejným typem rozdělení jako zkoumaný GLM model.
(2) Maximální model a zkoumaný mají stejnou linkovací funkci.
(3) Počet parametrů maximálního modelu je roven počtu vysvětlovaných veličin n, maximálně věrohodný odhad parametru $\boldsymbol{\beta}_{max}$ je n-rozměrný vektor $\widehat{\boldsymbol{\beta}}_{max}$.

Poznámka 6.2. Z definice plyne, že vysvětlovaná veličina $\mathbf{Y}$ je maximálním modelem určena s nulovým reziduem, tj. odhadnutá hodnota

$$\widehat{\mathbf{Y}}_{max} = \widehat{\boldsymbol{\mu}}_{max} = (\hat{\mu}_{max,1}, \ldots, \hat{\mu}_{max,n})' = \mathbf{Y} = (Y_1, \ldots, Y_n)'.$$

Definice 6.3. **Minimální GLM**, který označíme GLM_{min}, splňuje následující podmínky
(1) Minimální model je zobecněný lineární model se stejným typem rozdělení jako zkoumaný GLM model.
(2) Minimální model a zkoumaný mají stejnou linkovací funkci.
(3) Počet parametrů minimálního modelu je roven 1, maximálně věrohodný odhad parametru $\boldsymbol{\beta}_{min}$ je skalár $\hat{\beta}_{min}$.

Poznámka 6.4. Pro minimální model, kde $\mathbf{X} = (1, \ldots, 1)'$, lze snadno ověřit, že

$$\hat{\mu}_{min,i} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad i = 1, \ldots, n.$$

Maximální model tedy slouží jako ukazatel „nejlepší“ regrese a minimální model naopak jako ukazatel „nejhorší“ regrese při daném rozdělení a dané linkovací funkci. Zkoumaný model se bude nacházet někde mezi těmito extrémy a ve srovnání s nimi budeme oceňovat vhodnost modelu.

Definice 6.5. Mějme zobecněný lineární model s maticí plánu $\mathbf{X}_{n\times k}$ a vektorem neznámých parametrů $\boldsymbol{\beta}$. **Submodel**, který označíme GLM_{sub}, splňuje následující podmínky
(1) Submodel je zobecněný lineární model se stejným typem rozdělení jako zkoumaný GLM model.
(2) Submodel a zkoumaný model mají stejnou linkovací funkci.
(3) Vektor neznámých parametrů $\boldsymbol{\beta}_{sub} \in \mathbb{R}^q$a matice plánu $\mathbf{Q}_{n\times q}$, pro kterou platí

$$\mathbf{Q}_{n\times q} = \mathbf{X}_{n\times k}\mathbf{T}_{k\times q}.$$

Aby GLM_{sub} byl submodelem modelu GLM, musí každý sloupec matice $\mathbf{Q}$ patřit do obalu sloupců matice $\mathbf{X}$. To bude splněno právě tehdy, bude-li $\mathbf{Q}$ typu

$$\mathbf{Q}_{n\times q} = \mathbf{X}_{n\times k}\mathbf{T}_{k\times q}.$$

Je třeba si uvědomit, že GLM_{sub} je speciálním případem modelu GLM. Platí-li tudíž pro náhodný výběr $\mathbf{Y}$ model GLM_{sub}, platí pro $\mathbf{Y}$ také model GLM.

Model GLM vybíráme tak bohatý, abychom si mohli být jisti, že popisuje dobře chování $\mathbf{Y}$. Následně bychom ovšem chtěli vědět, zda lze použít jednodušší model GLM_{sub}. Můžeme usuzovat takto, platí-li GLM_{sub}, pak rozšíření na GLM nepřinese podstatné změny a vektory $\widehat{\boldsymbol{\beta}}$ a $\widehat{\boldsymbol{\beta}}_{sub}$ by se neměly podstatně lišit. Na druhé straně, budou-li $\widehat{\boldsymbol{\beta}}$ a $\widehat{\boldsymbol{\beta}}_{sub}$ příliš odlišné, svědčí to proti možnosti redukce GLM na GLM_{sub}.

6.2. Deviace. Deviace v zobecněných lineárních modelech je obdobou rozptylu u klasických lineárních regresních modelů. Deviace je tedy kritériem vhodnosti zobecněného lineárního modelu. Jak bude patrné z definice, metoda maximální věrohodnosti totiž odpovídá

hledání minima deviace modelu.

DEFINICE 6.6. Mějme modely GLM a GLM_{max}. Nechť náhodný výběr $\mathbf{Y}$ se řídí modelem GLM_{max}. **Škálovou deviací modelu GLM** (scaled deviance) rozumíme statistiku

$$D = 2\left[l^*(\widehat{\boldsymbol{\beta}}_{max}; \mathbf{Y}) - l^*(\widehat{\boldsymbol{\beta}}; \mathbf{Y})\right],$$

kde $\widehat{\boldsymbol{\beta}}_{max}, \widehat{\boldsymbol{\beta}}$ jsou odpovídající maximálně věrohodné odhady.

LEMMA 6.7. *Nechť je dán model GLM a náhodný výběr $\mathbf{Y}$ se řídí tímto modelem. Dále nechť*

(i) existují druhé parciální derivace hustoty $f(\mathbf{y};\boldsymbol{\beta})$ podle složek $\boldsymbol{\beta}$,

(ii) platí $E\left(\frac{f''_{\beta_i\beta_j}(\mathbf{Y};\boldsymbol{\beta})}{f(\mathbf{Y};\boldsymbol{\beta})}\right) = 0 \quad (i,j = 1,\ldots,k)$

(iii) a existuje $E\, l^*(\boldsymbol{\beta};\mathbf{Y})$.

Pak asymptoticky lze statistiku

$$D_k = 2\left[l^*(\widehat{\boldsymbol{\beta}}; \mathbf{Y}) - l^*(\boldsymbol{\beta}; \mathbf{Y})\right]$$

aproximovat kvadratickou formou

$$W_k = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{J}_n(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

a rozdělení statistiky D_k lze aproximovat rozdělením $\chi^2(k)$, přičemž $\widehat{\boldsymbol{\beta}}$ je maximálně věrohodný odhad parametru $\boldsymbol{\beta}$.

VĚTA 6.8. *Mějme modely GLM a GLM_{max}. Nechť náhodný výběr $\mathbf{Y}$ se řídí modelem GLM_{max}. Dále nechť*

(i) existují druhé parciální derivace hustoty $f(\mathbf{y};\boldsymbol{\beta}_{max})$ podle složek $\boldsymbol{\beta}_{max}$,

(ii) platí $E\left(\frac{f''_{\beta_{(max,i)}\beta_{(max,j)}}(\mathbf{Y};\boldsymbol{\beta}_{max})}{f(\mathbf{Y};\boldsymbol{\beta}_{max})}\right) = 0 \quad (i,j = 1,\ldots,n)$

(iii) a existuje $E\, l^*(\boldsymbol{\beta}_{max};\mathbf{Y})$.

*Platí-li hypotéza, že **model GLM je vhodný**, pak asymptoticky lze rozdělení škálové deviace modelu GLM aproximovat rozdělením $\chi^2(n-k)$, tj.*

$$D \overset{A}{\sim} \chi^2(n-k).$$

Poznámka 6.9. Škálovou deviaci D lze užít k testování hypotézy o vhodnosti modelu GLM. Jestliže platí

$$D > \chi^2_{1-\alpha}(n-k),$$

pak považujeme model GLM za nevhodný.

VĚTA 6.10. *Mějme základní model GLM s $\boldsymbol{\beta} \in \mathbb{R}^k$ a jeho submodel GLM_{sub} s $\boldsymbol{\beta}_{sub} \in \mathbb{R}^q$, přičemž $q < k < n$. Dále nechť náhodný výběr $\mathbf{Y}$ se řídí modelem GLM a platí*

(i) existují druhé parciální derivace hustoty $f(\mathbf{y};\boldsymbol{\beta})$ podle složek $\boldsymbol{\beta}$,

(ii) platí $E\left(\frac{f''_{\beta_i\beta_j}(\mathbf{y};\boldsymbol{\beta})}{f(\mathbf{y};\boldsymbol{\beta})}\right) = 0 \quad (i,j = 1,\ldots,k)$

(iii) a existuje $E\, l^*(\boldsymbol{\beta};\mathbf{Y})$.

*Platí-li hypotéza, že **submodel $\mathbf{GLM_{sub}}$ je vhodný**, pak asymptoticky lze rozdělení statistiky $\Delta D = D_{sub} - D$ aproximovat rozdělením $\chi^2(k-q)$, tj. $\Delta D = D_{sub} - D \overset{A}{\sim} \chi^2(k-q)$.*

Poznámka 6.11. Statistiky ΔD se užívá při porovnání dvou modelů, přičemž jeden je submodelem druhého.

Definujme diagonální matici $\mathbf{C}$, na jejíž hlavní diagonále budou nuly a jedničky, tj. $c_{ii} \in \{0, 1\}$. Počet jedniček na hlavní diagonále je roven číslu $q = Tr(C) < k$. Předpokládáme, že základní model s vektorem neznámých parametrů, který tentokrát označíme $\boldsymbol{\beta}_1$, dobře popisuje chování náhodného výběru.

Uvažujme hypotézu

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_{sub} = \mathbf{C}'\boldsymbol{\beta}_1$$

a alternativní hypotézu odpovídající základnímu modelu

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1.$$

Test hypotézy H_0 vůči H_1 provedeme s využitím statistiky ΔD. Má-li statistika ΔD hodnotu

$$\Delta D > \chi^2_{1-\alpha}(k - q),$$

pak **považujeme submodel za nevhodný** a pro popis chování sledovaného náhodného výběru použijeme model základní. V opačném případě jsou oba dva modely dobré a **vybereme jednodušší model odpovídající** H_0.

Je nutné zdůraznit, že pojem „*dobře popisuje chování náhodného výběru*“ je zde použit pouze relativně z důvodu srovnání dvou modelů, kdy je třeba zjistit, zda obecnější model je výrazně lepší, či jsou oba modely přibližně srovnatelné. Neznamená ještě, když přijmeme hypotézu H_0, že jsou oba modely pro daná data (absolutně) dobré, ale například jen to, že jsou oba stejně špatné.

Proto při praktických úlohách se volí matice plánu obecnějšího modelu se všemi potencionálními vysvětlujícími veličinami a předpokládá se, že tento model popisuje data dobře. Následuje odstraňování jednotlivých vysvětlujících veličin a testování, zda vzniklý jednodušší model nepopisuje data na určité hladině významnosti stejně dobře, jako složitější model základní.

6.3. Analýza reziduí. Analýza reziduí nám poskytuje nenahraditelnou informaci o vhodnosti použitého modelu a nelze ji opomíjet, i když analýza deviace ukazuje na vhodnost modelu. Teprve zde se častokrát zjistí, že výchozí předpoklad o rozdělení náhodných chyb či tvaru linkovací funkce nebyl správný.

Analýza reziduí může odhalit body, jejichž reziduum je výrazně odlišné od ostatních pozorování, což může být způsobeno neobvyklou závislostí mezi vysvětlovanou a vysvětlujícími veličinami či prozaičtěji chybou měření, chybou přepisu hodnot do databáze apod.

Když se v grafu reziduí objeví určitá závislost reziduí na fitované hodnotě či vysvětlujících veličinách nebo třeba s rostoucí odhadnutou (fitovanou) hodnotou roste variabilita reziduí, pak je nutné celý model přehodnotit a případně jej začít vytvářet od začátku.

Uveďme nejznámější typy reziduí používaných v GLM :

(a) **Standardizovaná rezidua** (*linear*): též *Pearsonova*

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}.$$

Nevýhodou těchto reziduí je fakt, že pro nenormální rozdělení jsou značně zešikmená.

(b) **Standardizovaná transformovaná rezidua** (*transformed linear*)

$$r_i^{Tr} = \frac{g(Y_i) - Eg(Y_i)}{\sqrt{Dg(Y_i)}}.$$

Důvodem zavedení transformovací je snaha, aby transformovaná rezidua měla rozdělení, které se co nejvíce blíží normálnímu.

(1) **Anscombova rezidua** jsou založena na transformaci typu

$$g_A(\mu) = \int \frac{d\mu}{V^{\frac{1}{3}}(\mu)},$$

jejíž snahou je, aby transformovaná rezidua měla **nulovou šikmost**.

(2) **Rezidua stabilizující rozptyl** jsou založena na transformaci typu

$$g_S(\mu) = \int \frac{d\mu}{V^{\frac{1}{2}}(\mu)},$$

a cílem je, aby u transformovaných reziduí rozptyl nebyl funkcí střední hodnoty, ale konstantní.

Příklady standardizovaných transformovaných reziduí

Rozdělení	Anscombova	stabilizující rozptyl
Binomické	$\frac{t\left(\frac{Y_i}{n_i}\right) - \left[t(\hat{\pi}_i) + \hat{\pi}_i^{-\frac{1}{3}}(1-\hat{\pi}_i)^{-\frac{1}{3}}(2\hat{\pi}_i - 1)/6n_i\right]}{\hat{\pi}_i^{\frac{1}{6}}(1-\hat{\pi}_i)^{\frac{1}{6}}/n_i}$	$\frac{\arcsin\sqrt{\frac{Y_i}{n_i}} - \arcsin\sqrt{\hat{\pi}_i}}{1/(2\sqrt{n_i})}$
Poissonovo	$\frac{Y_i^{\frac{2}{3}} - \left(\hat{\mu}_i^{\frac{2}{3}} - \hat{\mu}_i^{-\frac{1}{3}}/9\right)}{\frac{2}{3}\hat{\mu}_i^{\frac{1}{6}}}$	$\frac{Y_i^{\frac{1}{2}} - \hat{\mu}_i^{\frac{1}{2}}}{\frac{1}{2}}$
Gamma	$\frac{(Y_i/\alpha)^{\frac{1}{3}} - \left[\hat{\mu}_i^{\frac{1}{3}} - \hat{\mu}_i^{-\frac{2}{3}}/9\right]}{\hat{\mu}_i^{-\frac{1}{6}}/3}$	neuvažujeme

kde pro binomické rozdělení: $\hat{\pi}_i = \hat{\mu}_i/n_i$ a $t(u) = \int_0^u s^{-\frac{1}{3}}(1-s)^{-\frac{1}{3}}ds$

(c) **Deviační rezidua** (*deviance residual*)

$$r_i^D = sign(Y_i - \hat{\mu}_i)\sqrt{d_i},$$

přičemž

$$D = 2[l^*(\mathbf{b}_{max}; \mathbf{Y}) - l^*(\mathbf{b}; \mathbf{Y})] = \sum_{i=1}^{n} d_i.$$

Ještě lepší vlastnosti mají tzv. **korigovaná deviační rezidua** (bias-adjusted deviance residual)

$$r_i^{AD} = sign(Y_i - \hat{\mu}_i)\sqrt{d_i} + \rho_3(\hat{\mu}_i)/6,$$

kde

$$\rho_3(\mu) = E\left[\left(\frac{Y-\mu}{\sqrt{DY}}\right)^3\right].$$

7. Tabulky rozdělení exponenciálního typu

7.1. Tabulka rozdělení exponenciálního typu.

Příklady rozdělení exponenciálního typu						
Rozdělení	$EY_i = \mu_i$	**Kanonická link-funkce** $\theta_i = \eta_i = g(\mu_i)$	$\gamma(\theta_i)$	**Rozptylová funkce** $V(\mu_i)$	$\psi_i(\phi) = \frac{\phi}{\omega_i}$ ϕ	ω_i
$Y_i \sim N(\mu_i, \frac{\sigma^2}{\omega_i})$	μ_i	μ_i	$\frac{1}{2}\theta_i^2$	1	σ^2	ω_i
$Y_i \sim Bi(n_i, \pi_i)$	$n_i\pi_i$	$\ln \frac{\mu_i}{n_i-\mu_i}$	$n_i \ln(1+e^{\theta_i})$	$\mu_i\left(1-\frac{\mu_i}{n_i}\right)$	1	1
$n_iY_i \sim Bi(n_i, \pi_i)$	π_i	$\ln \frac{\mu_i}{1-\mu_i}$	$\ln(1+e^{\theta_i})$	$\mu_i(1-\mu_i)$	1	n_i
$Y_i \sim Po(\mu_i)$	μ_i	$\ln \mu_i$	e^{θ_i}	μ_i	1	1
$Y_i \sim Ex(\mu_i)$	μ_i	$-\frac{1}{\mu_i}$	$-\ln(-\theta_i)$	μ_i^2	1	1
$Y_i \sim G(\alpha, \beta_i)$	$\alpha\beta_i$	$-\frac{1}{\mu_i}$	$-\ln(-\theta_i)$	μ_i^2	$\frac{1}{\alpha}$	1

7.2. Tabulka různých spojovacích funkcí.

Rozdělení	**link-funkce**	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$	$\frac{\partial \eta_i}{\partial \mu_i}$
Normální, Gama	identity	μ_i	η_i	1
Exponenciální	log	$\ln \mu_i$	e^{η_i}	$\frac{1}{\mu_i}$
Poissonovo	power	μ_i^a	$e^{\frac{1}{a}\ln \eta_i}$	$a\mu_i^{a-1}$
Binomické	logit	$\ln \frac{\mu_i}{n_i-\mu_i}$	$n_i \frac{e^{\eta_i}}{1+e^{\eta_i}}$	$\frac{n_i}{\mu_i(n_i-\mu_i)}$
	probit	$\Phi^{-1}(\frac{\mu_i}{n_i})$	$n_i\Phi(\eta_i)$	$\frac{\sqrt{2\pi}}{n_i} e^{\frac{1}{2}\left(\Phi^{-1}(\frac{\mu_i}{n_i})\right)^2}$
	complement.log-log	$\ln\left(-\ln\left(1-\frac{\mu_i}{n_i}\right)\right)$	$n_i\left(1-e^{-e^{\eta_i}}\right)$	$\frac{1}{(\mu_i-n_i)\ln\left(1-\frac{\mu_i}{n_i}\right)}$
	log-log	$-\ln\left(-\ln\left(\frac{\mu_i}{n_i}\right)\right)$	$n_i e^{-e^{-\eta_i}}$	$-\frac{1}{\mu_i \ln\left(\frac{\mu_i}{n_i}\right)}$

kde Φ^{-1} je kvantilová funkce standardizovaného normálního rozdělení.

Klasický regresní model a GLM

Značení

Střední hodnoty jednotlivých pozorování $EY_i = \mu_i \quad (i = 1, \dots, n)$

Hodnoty regresorů jednotlivých pozorování $\mathbf{x}_i' = (x_{i1}, \dots, x_{ik}) \quad (i = 1, \dots, n)$

Vektor pozorovaných hodnot $\quad \mathbf{Y} = (y_1, \dots, y_n)'$

Vektory neznámých parametrů $\quad (q < k < n)$

$$\boldsymbol{\beta} = \boldsymbol{\beta}_k = (\underbrace{\beta_1, \dots, \beta_q}_{\boldsymbol{\beta}_q}, \underbrace{\beta_{q+1}, \dots, \beta_k}_{\boldsymbol{\beta}_{k-q}})'; \qquad \boldsymbol{\beta}_{max} = (\underbrace{\beta_1, \dots, \beta_q}_{\boldsymbol{\beta}_q}, \underbrace{\beta_{q+1}, \dots, \beta_k}_{\boldsymbol{\beta}_{k-q}}, \underbrace{\beta_{k+1}, \dots, \beta_n}_{\boldsymbol{\beta}_{n-k}})'$$

Matice plánu z hodnot regresorů $\quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$

Submodely $\quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{V} \end{pmatrix} \qquad h(\mathbf{X}'\mathbf{X}) = k;\ h(\mathbf{V}) = k - q;\ h(\mathbf{A}) = q$

Model

$EY_i = \mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ $Y_i \sim N(\underbrace{\mathbf{x}_i'\boldsymbol{\beta}}_{EY_i}, \underbrace{\sigma^2}_{DY_i})$	$EY_i = \mu_i \qquad \eta_i = g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} \qquad \mu_i = g^{-1}(\eta_i)$ $Y_i \sim \mathcal{L}_{exp}(\underbrace{\gamma'(\theta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})}_{EY_i}, \underbrace{\psi_i(\phi)\gamma''(\theta_i)}_{DY_i})$

Odhady

Metoda nejmenších čtverců	**Iterativni vážená metoda nejm. čtverců (max.věr.odh.)**
$\widehat{\boldsymbol{\beta}} = \mathbf{b}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$	$\widehat{\boldsymbol{\beta}}^{(s)} = \mathbf{b}_k^{(s)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}^{(s-1)}$
	$\mathbf{Z}^{(s-1)} = (Z_1^{(s-1)}, \dots, Z_n^{(s-1)})';\ Z_i^{(s-1)} = \hat{\eta}_i^{(s-1)} + (y_i - \hat{\mu}_i^{(s-1)})\frac{d\eta_i}{d\mu_i}$ $\hat{\eta}_i^{(s-1)} = \mathbf{x}_i'\mathbf{b}^{(s-1)}; \quad \hat{\mu}_i^{(s-1)} = g^{-1}(\hat{\eta}_i^{(s-1)})$ $w_i = \frac{1}{DY_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2; \quad W = diag\{w_1, \dots, w_n\}$
$\mathbf{b}_{max} = (b_1, \dots, b_n)' = (y_1, \dots, y_n)'$	

Výběrová rozdělení odhadu $\mathbf{b}_k$

$\mathbf{b}_k \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$	$\mathbf{b}_k \overset{A}{\sim} N_k(\boldsymbol{\beta}, \mathbf{J}_n^{-1})$
$\frac{1}{\sigma^2}(\mathbf{b}_k - \boldsymbol{\beta}_k)'\mathbf{X}'\mathbf{X}(\mathbf{b}_k - \boldsymbol{\beta}_k) \sim \chi^2(k)$	$(\mathbf{b}_k - \boldsymbol{\beta}_k)'\mathbf{J}_\mathbf{n}(\mathbf{b}_k - \boldsymbol{\beta}_k) \overset{A}{\sim} \chi^2(k)$ *Waldova statistika*

Výběrová rozdělení škálové deviace $D = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_k; \mathbf{y})]$

$D \sim \chi^2(n-k)$	$D \overset{A}{\sim} \chi^2(n-k)$
Normální rozdělení $D = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{SSE}{\sigma^2} = \frac{(n-k)s_k^2}{\sigma^2}$	**Poissonovo rozdělení**: $\quad D = 2\sum_{i=1}^{n}\left[y_i ln\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right]$
	Binomické rozdělení $\quad D = 2\sum_{i=1}^{n}\left[y_i ln\frac{y_i}{\hat{\mu}_i} + (n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right]$
	Gamma rozdělení $\quad D = 2\alpha\sum_{i=1}^{n}\left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - ln\frac{y_i}{\hat{\mu}_i}\right]$

Submodely

$$\boxed{H_0} : \boldsymbol{\beta} = (\beta_1, \dots, \beta_q)' \text{ proti } \boxed{H_1} : \boldsymbol{\beta} = (\beta_1, \dots, \beta_q, \beta_{q+1}, \dots, \beta_k)'$$
$$\text{nebo } \boxed{H_0} : (\beta_{q+1}, \dots, \beta_k)' = \mathbf{0} \text{ proti } \boxed{H_1} : (\beta_{q+1}, \dots, \beta_k)' \neq \mathbf{0}$$

$F \sim F(k-q, n-k)$	$\Delta D \overset{A}{\sim} \chi^2(k-q)$
$F = \frac{1}{(k-q)s_k}\mathbf{b}_{k-q}'\mathbf{V}^{-1}\mathbf{b}_{k-q}$	$\Delta D = D_{sub} - D = 2[l(\mathbf{b}_k; \mathbf{y}) - l(\mathbf{b}_q; \mathbf{y})]$

Úlohy k procvičení

Cvičení 7.1.
V souboru „`toxic.RData`“ jsou uvedeny hodnoty množství jedovaté látky, která vzniká jako vedlejší produkt při určitém chemickém procesu. Datový soubor obsahuje tyto proměnné:

`VOL`	objem vzniklé jedovaté látky (litry)
`TEMP`	teplota při chemickém procesu (°C)
`CAT`	hmotnost katalyzátoru (kg)
`METHOD`	metoda použitá při výrobě (kategoriální proměnná – A,B)

Hledejte vhodný model pro popis závislosti objemu jedovaté látky na podmínkách procesu. Testujte nejprve, zda použitá metoda má vliv na výsledný objem jedovaté látky. Pomocí stepwise procedury najděte nejvhodnější lineární model a nejvhodnější zobecněný lineární model. U obou modelů ověřte normalitu residuí.

[Metoda má vliv, vhodný model: $\texttt{VOL} = \beta_0 + \beta_1\texttt{METHODB} + \beta_2\texttt{TEMP}$, residua jsou normální]

Cvičení 7.2.
V balíku „`car`“, proměnné „`SLID`“ jsou uvedeny výsledky průzkumu z roku 1994 v kanadské provincii Ontario. Průzkum se zabýval vlivem některých faktorů na mzdu respondentů. Datový soubor obsahuje tyto proměnné:

`wages`	hodinová mzda (kanadské dolary)
`education`	počet let vzdělávání (roky)
`age`	věk (roky)
`sex`	pohlaví (1 – žena, 2 – muž)
`language`	jazyk (1 – angličtina, 2 – francouzština, 3 – ostatní)

Hledejte vhodný model pro popis závislosti platu respondenta na ostatních faktorech.

(1) Zkuste nejprve použít klasický lineární model, najděte nejvhodnější model a proveďte analýzu residuí. Jsou splněny předpoklady modelu?
(2) Stále uvažujte lineární model. Místo proměnné `wages` uvažujte `log(wages)`. Opět nalezněte nejvhodnější model. Zkuste také přidat dvojné či trojné interakce proměnných. Zlepší se kvalita modelu?
(3) Pomocí stepwise procedury najděte nejvhodnější zobecněný lineární model.

[(1) Vhodný model: $\texttt{wages} = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{education} + \beta_3\texttt{sex}$, residua nejsou normální, (2) kvalita se zlepší přidáním dvojných interakcí, (3) vhodný model: $\texttt{wages} = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{education} + \beta_3\texttt{sexMale} + \beta_4\texttt{age:sexMale} + \beta_5\texttt{education:sexMale} + \beta_6\texttt{age:education}$.]

Cvičení 7.3.
V souboru „`novorozenci.RData`“ jsou uvedeny porodní hmotnosti novorozenců a informace o jejich rodičích. Datový soubor obsahuje tyto proměnné:

`hmnov`	porodní hmotnost novorozence (g)
`vyska`	výška matky (cm)
`hmmat`	hmotnost matky (kg)
`prir`	váhový přírůstek matky během těhotenství (kg)
`pohlavi`	pohlaví dítěte (0 – dívka, 1 – chlapec)
`stav`	stav matky při porodu (1 – svobodná, 2 – vdaná, 3 – rozvedená, 4 – vdova)
`vzdmat`	vzdělání matky (1 – zákl., 2 – vyuč., 3 – středošk., 4 – vysokošk.)
`vzdot`	vzdělání otce (0 – neuved., 1 – zákl., 2 – vyuč., 3 – středošk., 4 – vysokošk.)

Hledejte vhodný model pro popis závislosti hmotnosti novorozence na jeho rodičích. Testujte nejprve, zda pohlaví má vliv na porodní hmotnost. Pomocí stepwise procedury najděte nejvhodnější model. U modelu ověřte normalitu residuí.

[Pohlaví má vliv, vhodný model: $\texttt{hmmat} = \beta_0 + \beta_1\texttt{prir} + \beta_2\texttt{pohlavi1} + \beta_3\texttt{vzdot1} + \beta_4\texttt{vzdot2} + \beta_5\texttt{vzdot3} + \beta_6\texttt{vzdot4} + \beta_7\texttt{vyska} + \beta_8\texttt{hmmat:pohlavi1}$, residua jsou normální]

KAPITOLA 8

Konkrétní GLM modely

Základní informace

(1) V následující kapitole se budeme zabývat aplikací teorie zobecněných lineárních modelů na různé typy dat. Budeme se zabývat případy, kdy pozorovaná veličina má alternativní nebo binomické rozdělení. To vede na modely typu dávka–odpověď a také na logistickou regresi. Dále budeme teoretické poznatky aplikovat na případ poissonovských dat a budeme se zabývat modelováním multinomických dat vedoucím na kontingenční tabulky. Popíšeme tedy konkrétní zobecněné lineární modely.

(2) Předpokládá se znalost základních pojmů z teorie lineárních regresních modelů a zobecněných lineárních modelů – matice plánu, metoda nejmenších čtverců, linkovací funkce, škálová deviace.

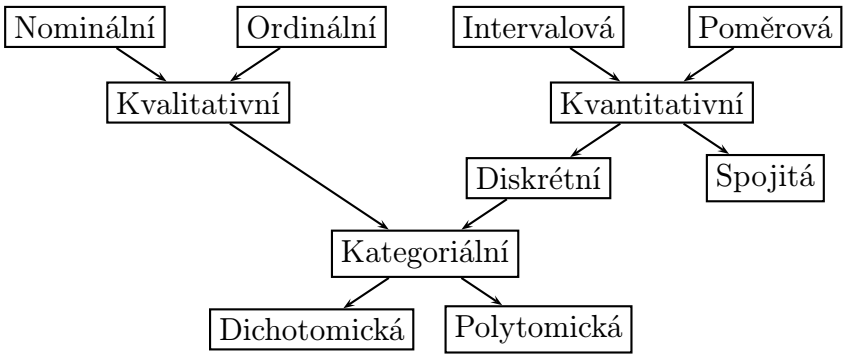
Výstupy z výukové jednotky

Studenti

- umí konstruovat a interpretovat modely typu dávka–odpověď
- umí definovat a vysvětlit model logistické regrese
- definují modely pro poissonovská data
- definují kontingenční tabulku
- modelují kontingenční tabulku jako zobecněný lineární model

1. Motivace

V minulé kapitole jsme uvedli obecnou definici zobecněného lineárního modelu a obecné konstrukce testů hypotéz o parametrech těchto modelů. V této kapitole se již budeme zabývat zobecněnými lineárními modely pro konkrétní případy podle toho, jaké rozdělení má závisle proměnná Y. Nejprve je vhodné nahlédnout do úvodní kapitoly a připomenout si základní typy proměnných (ať už závisle či nezávisle proměnných) z hlediska vztahu mezi dvěma hodnotami. Tyto typy lze názorně popsat následujícím diagramem.



V závislosti na typu proměnné Y a jejím rozdělení pravděpodobnosti budeme zkoumat již konkrétní zobecněné lineární modely.

Poznámka 1.1. Většina textu v této kapitole byla převzata z [**5**]. Pro podrobnější studium tohoto tématu proto odkazujeme na tento zdroj.

2. Modely pro alternativní a binomická data

2.1. Úvod. Předpokládejme, že sledovaná náhodná veličina U_i $(i = 1, \dots, N)$ nabývá pouze dvou hodnot 0 a 1, tj. má alternativní rozdělení:

$$U_i \sim A(\pi_i) \sim f_U(u) = P(U_i = u) = \begin{cases} \pi_i & u = 1 \\ 1 - \pi_i & u = 0 \\ 0 & \text{jinak} \end{cases} = \begin{cases} \pi_i^u (1 - \pi_i)^{1-u} & u = 0, 1 \\ 0 & \text{jinak} \end{cases}.$$

Předpokládejme, že náhodná veličina U_i závisí na m veličinách $x_{i1}, \dots, x_{ik}$, tzv. **kovariáty**. Data můžeme mít zadána různým způsobem:

- **jednotlivá pozorování** U_i:

hodnoty kovariát	pozorované binární veličiny
$x_{i1}, \dots, x_{ik}$	U_i

- **skupinově**, tj. pro každou kombinaci kovariát známe **absolutní četnosti** úspěchů Y_j a celkový počet pokusů n_j, tedy máme k dispozici binomická data

$$Y_j = \sum_{i=1}^{n_j} U_i \sim Bi(n_j, \pi_j) \sim f_Y(y) = P(Y_j = y) = \begin{cases} \binom{n_j}{y} \pi_j^y (1 - \pi_j)^{n_j - y} & y = 0, 1, \dots, n_j \\ 0 & \text{jinak} \end{cases}$$

kde

$$j = 1, \dots, n; \qquad N = n_1 + \cdots + n_n$$

a data můžeme zapsat formou tabulky

hodnota kovariát	počet úspěchů	počet pokusů
$x_{j1}, \dots, x_{jk}$	Y_j	n_j

- **skupinově**, tj. pro každou kombinaci kovariát máme **relativní četnost** úspěchů $Z_j = \frac{Y_j}{n_j}$ a celkový počet pokusů n_j

$$Z_j = \frac{Y_j}{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} U_i \sim f_Z(y) = P(Z_j = y) = \begin{cases} \binom{n_j}{n_j y} \pi_j^{n_j y} (1 - \pi_j)^{n_j - n_j y} & y = 0, \frac{1}{n_j}, \dots, 1 \\ 0 & \text{jinak} \end{cases}$$

kde

$$j = 1, \dots, n; \qquad N = n_1 + \cdots + n_n$$

Data lze zapsat do tabulky

kovariáty	relativní úspěšnost	počet pokusů
$x_{j1}, \dots, x_{jk}$	$Z_j = \frac{Y_j}{n_j}$	n_j

- pro **nominální** či **ordinální kovariáty** můžeme data psát do tzv. **kontingenčních tabulek.** Uvažujme jednoduchý příklad:

kovariáty		$U = 0$	$U = 1$
$x_1 = 1$	$x_2 = 1$	n_{110}	n_{111}
	$x_2 = 2$	n_{120}	n_{121}
$x_1 = 2$	$x_2 = 1$	n_{210}	n_{211}
	$x_2 = 2$	n_{220}	n_{221}

V dalším se soustřeďme na **relativní četnosti úspěchů**

$$Z_i = \frac{Y_i}{n_i}.$$

Hlavním úkolem statistické analýzy je pak nalézt vztah mezi Z_i, (tj. i Y_i) a $x_{i1}, \ldots, x_{ik}$, tj. funkci

$$\pi_i = \pi(\mathbf{x}_i) = \pi(x_{i1}, \ldots, x_{ik}).$$

Protože chceme použít GLM modely, modelujeme pravděpodobnosti π_i pomocí linkovacích funkcí

$$g(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}.$$

Nejjednodušším modelem je **lineární model**

$$\pi_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

Avšak tento model má řadu nevýhod, především je třeba zajistit, aby $\mathbf{x}_i'\boldsymbol{\beta}$ nabývala hodnot mezi 0 a 1, tedy je třeba přidat nějaké dodatečné podmínky. Proto, abychom tuto podmínku dodrželi, využijeme nějakou **distribuční funkci**

$$F(t) = \int_{-\infty}^{t} f(s)ds \qquad f(s) \geq 0 \qquad \int_{-\infty}^{\infty} f(s)ds = 1$$

s odpovídající **hustotou** $f(s)$, která se v tomto případě nazývá **toleranční funkce** (toleranční distribuce). Nyní si ukážeme několik modelů, které využívají různé toleranční distribuce.

2.2. Modely dávka – odpověď. Typickým příkladem těchto modelů je vztah mezi dávkou toxické látky a odezvy (kladná-přežití, záporná-smrt) jedince na tuto dávku. Odezvy bývají obvykle udávány jako procenta kladné odezvy (quantal responses).

Symetrické modely

Definice 2.1. Jestliže uvažujeme toleranční distribuci jako **rovnoměrně spojitou** na nějakém intervalu (a, b), tj

$$f_0(s) \sim Rs(a, b) \qquad f(s) = \begin{cases} \frac{1}{b-a} & s \in (a, b) \\ 0 & \text{jinak} \end{cases}$$

pak

$$\pi_0(x) = F_0(x) = \int_a^x f(s)ds = \frac{x-a}{b-a} \quad \text{pro} \quad x \in (a, b)$$

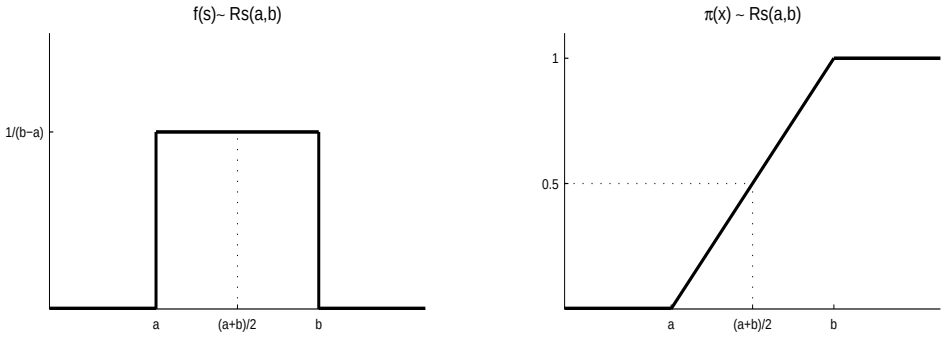
a tento model je **lineárním modelem**

$$\pi_0(x) = \frac{x-a}{b-a} = \beta_0 + \beta_1 x \qquad \text{tj.} \qquad \beta_0 = -\frac{a}{b-a} \quad \beta_1 = \frac{1}{b-a} > 0$$

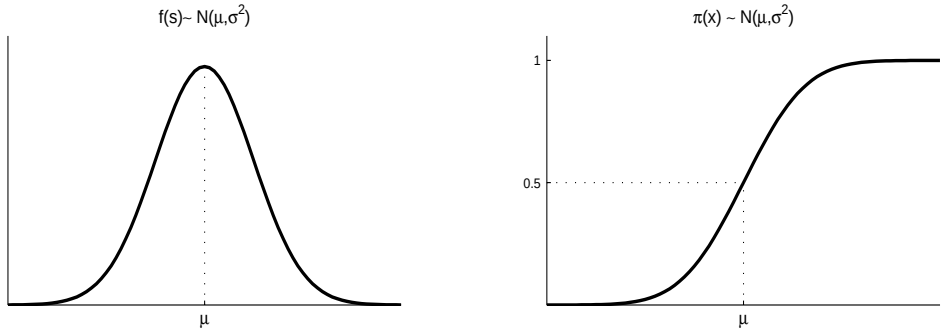
s identickou linkovací funkcí

$$g_0(\pi) = \pi.$$

V praxi tento model však nemá přílišné uplatnění.



Obrázek 1: Rovnoměrné rozdělení na (a, b).



Obrázek 2: Normální rozdělení $N(\mu, \sigma^2)$.

Další možností je vzít **normální hustotu** jako **toleranční funkci**. Připomeňme, že střední hodnota, medián i modus je roven parametru μ a rozptyl parametru $\sigma^2 > 0$.

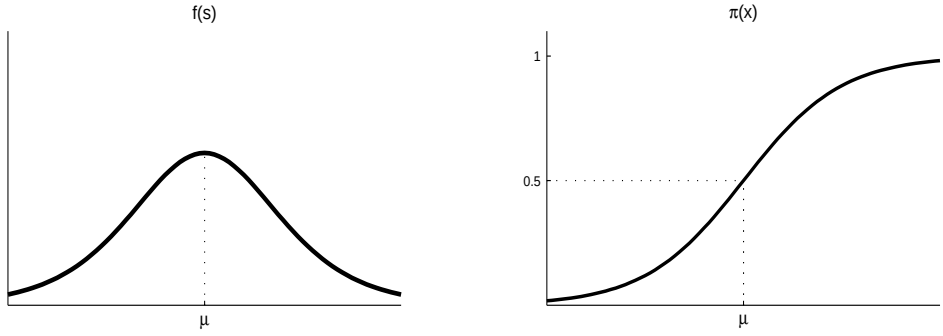
DEFINICE 2.2. Jestliže toleranční funkcí je **normální hustota**, mluvíme o **probitovém modelu**:

$$\pi_1(x) = F_1(x) = \int_{-\infty}^{x} f_1(s)ds = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2} ds = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

kde Φ je distribuční funkce standardizovaného normálního rozdělení. Pak tzv. **probitovou linkovací funkcí** je kvantilová funkce normálního rozdělení

$$g_1(\pi) = \Phi^{-1}(\pi) = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \qquad \text{tj.} \qquad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

Hodnota mediánu $x = \mu$ se nazývá **mediánová smrtící dávka** (*median lethal dose* - LD50) a odpovídá dávce, při které polovina jedinců má kladnou a polovina zápornou odezvu. Probitový model má široké uplatnění v biologických a sociálních vědách.



Obrázek 3: Logistické rozdělení.

DEFINICE 2.3. **Logistický model** je model, kde toleranční funkce je hustota **logistického rozdělení** (se střední hodnotou, mediánem i modusem μ a rozptylem $\frac{\pi^2}{3}\sigma^2 = 3.2899\sigma^2 = (1.8138\sigma)^2$)

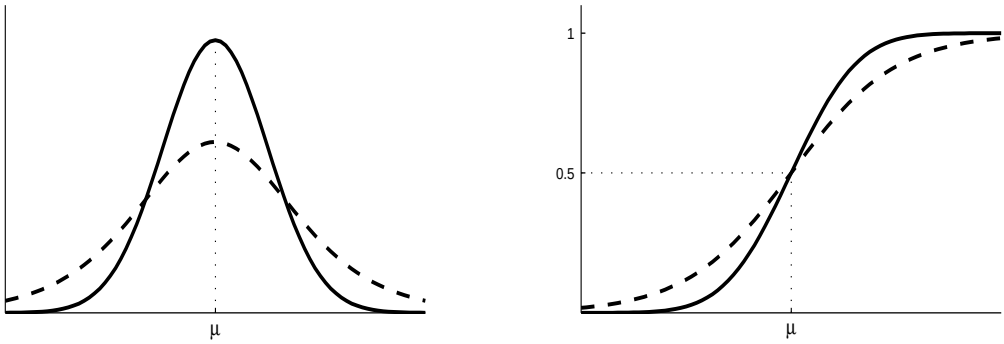
$$f_2(s) = \frac{1}{\sigma}\frac{\exp\left(\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{s-\mu}{\sigma}\right)\right]^2} = \frac{1}{\sigma}\frac{\exp\left(-\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(-\frac{s-\mu}{\sigma}\right)\right]^2},$$

takže

$$\pi_2(x) = F_2(x) = \int_{-\infty}^{x} \frac{1}{\sigma}\frac{\exp\left(\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{s-\mu}{\sigma}\right)\right]^2}\, ds = \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)} = \frac{1}{1+\exp\left(-\frac{x-\mu}{\sigma}\right)}$$

s tzv. **logit linkovací funkcí**

$$g_2(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \qquad \text{tj.} \qquad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$



Obrázek 4: Srovnání probitového a logistického (- - -) modelu při stejných parametrech μ a σ. Názorně je vidět, že logistický model má větší rozptyl a těžší konce.

Asymetrické (extremální) modely

DEFINICE 2.4. Pokud za toleranční funkci zvolíme **Log-Weibullovo rozdělení** (*extreme-minimal-value distribution*) ve tvaru

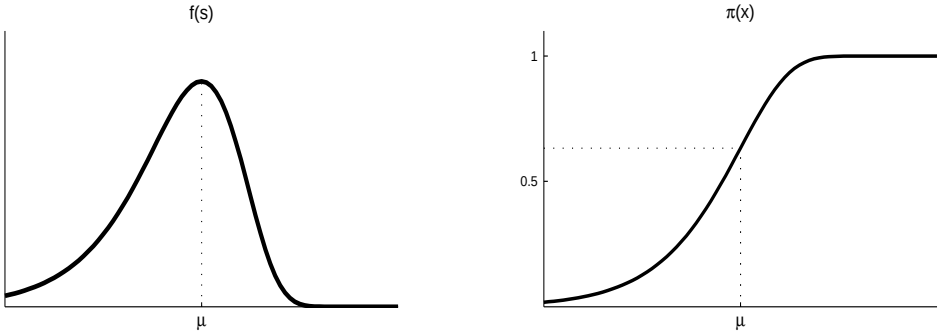
$$f_3(s) = \tfrac{1}{\sigma}\exp\left(\tfrac{s-\mu}{\sigma}\right)\exp\left[-\exp\left(\tfrac{s-\mu}{\sigma}\right)\right],$$

pak

$$\pi_3(x) = F_3(x) = \int_{-\infty}^{x} \tfrac{1}{\sigma}\exp\left(\tfrac{s-\mu}{\sigma}\right)\exp\left[-\exp\left(\tfrac{s-\mu}{\sigma}\right)\right]\,ds = 1-\exp\left[-\exp\left(\tfrac{x-\mu}{\sigma}\right)\right]$$

s tzv. **komplementární log-log linkovací funkcí**

$$g_3(\pi) = \log[-\log(1-\pi)] = \tfrac{x-\mu}{\sigma} = \beta_0+\beta_1 x \quad \text{tj.} \quad \beta_0 = -\tfrac{\mu}{\sigma} \quad \beta_1 = \tfrac{1}{\sigma} > 0 \quad \text{tedy } \beta_1 > 0.$$



Obrázek 5: Log-Weibullovo rozdělení.

Pro výše uvedené rozdělení můžeme vyjádřit jeho číselné charakteristiky:

střední hodnota$=\mu-\gamma\sigma \doteq \mu-0.57721\sigma$ kde $\gamma \doteq 0.57721$ je tzv.

medián$=\mu+\sigma\log(\log 2) \doteq \mu-0.36651\sigma$ *Euler-Mascheroniho*

modus$=\mu$ konstanta.

rozptyl$=\tfrac{\pi^2}{6}\sigma^2 = 1.6449\sigma^2 = (1.2825\sigma)^2,$

DEFINICE 2.5. Pokud jako toleranční funkci zvolíme **zobecněné Gumbelovo rozdělení** (*extreme-maximal-value distribution*) ve tvaru

$$f_4(s) = \tfrac{1}{\sigma}\exp\left(-\tfrac{s-\mu}{\sigma}\right)\exp\left[-\exp\left(-\tfrac{s-\mu}{\sigma}\right)\right],$$

dostaneme

$$\pi_4(x) = F_4(x) = \int_{-\infty}^{x} \tfrac{1}{\sigma}\exp\left[-\tfrac{s-\mu}{\sigma}-\exp\left(-\tfrac{s-\mu}{\sigma}\right)\right]\,ds = \exp\left[-\exp\left(-\tfrac{x-\mu}{\sigma}\right)\right]$$

s tzv. **log-log linkovací funkcí**

$$g_3(\pi) = -\log[-\log(\pi)] = \tfrac{x-\mu}{\sigma} = \beta_0+\beta_1 x \quad \text{tj.} \quad \beta_0 = -\tfrac{\mu}{\sigma} \quad \beta_1 = \tfrac{1}{\sigma} > 0.$$

Pro výše uvedené rozdělení opět vyjádřeme jeho číselné charakteristiky:

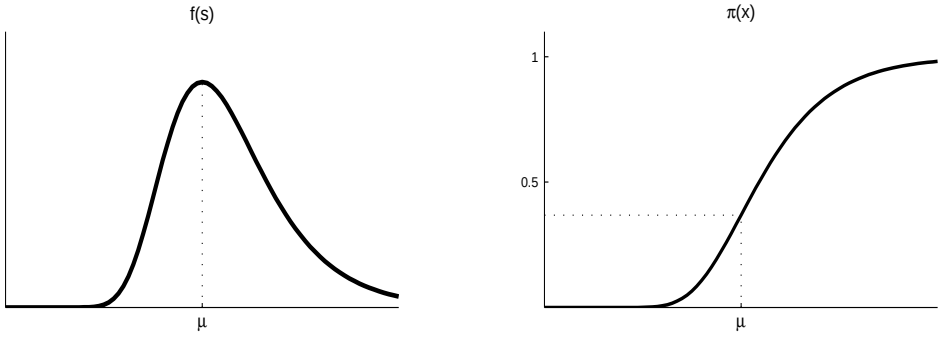
kde $\gamma \doteq 0.57721$ je tzv.

střední hodnota $=\mu+\gamma\sigma \doteq \mu+0.57721\sigma$ *Euler-Mascheroniho*

medián $=\mu-\sigma\log(\log 2) \doteq \mu+0.36651\sigma$ konstanta.

modus $=\mu$

rozptyl $=\tfrac{\pi^2}{6}\sigma^2 = 1.6449\sigma^2 = (1.2825\sigma)^2,$



Obrázek 6: Zobecněné Gumbelovo rozdělení.

Poznámka 2.6. Pokud náhodná veličina U má rozdělení **rovnoměrně spojité** na intervalu $(0,1)$, tj.

$$U \sim Rs(a,b),$$

pak

$X = \mu + \sigma \log\left(\frac{U}{1-U}\right)$ má logistické rozdělení s hustotou $f_2(x)$,

$X = \mu + \sigma \log(-\log(1-U))$ Log-Weibullovo rozdělení s hustotou $f_3(x)$,

$X = \mu - \sigma \log(-\log(U))$ Gumbelovo rozdělení s hustotou $f_4(x)$.

Těchto vztahů se využívá při generování pseudonáhodných čísel příslušných rozdělení.

2.3. Logistická regrese. Protože nejčastěji se používá **logit** linkovací funkce

$$g_2(\pi) = \log\left(\frac{\pi}{1-\pi}\right),$$

budeme se proto věnovat logistické regresi podrobněji.

Předpokládejme, že závisle proměnná Y je binární proměnná, která nabývá hodnoty jedna, pokud sledovaný jev nastal, v opačném případě je rovna nule.

Protože jde o regresní model, bude nás zajímat vztah pravděpodobností úspěchu či neúspěchu k hodnotám regresorů (kovariát) $\mathbf{x} = (x_1, \ldots, x_k)'$, budeme tedy zkoumat pravděpodobnost

$$P(Y = 1|x_1, \ldots, x_k) = \pi(\mathbf{x}) = \frac{\exp\{\eta(\mathbf{x})\}}{1 + \exp\{\eta(\mathbf{x})\}} = \frac{1}{1 + \exp\{-\eta(\mathbf{x})\}}$$

a

$$P(Y = 0|x_1, \ldots, x_k) = 1 - \pi(\mathbf{x}) = \frac{1}{1 + \exp\{\eta(\mathbf{x})\}} = \frac{\exp\{-\eta(\mathbf{x})\}}{1 + \exp\{-\eta(\mathbf{x})\}}$$

Předpokládejme, že lineární prediktor je roven

$$\eta(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}$$

a ukážeme, že má smysl uvádět absolutní člen samostatně.
Všimněme se nejprve, že podíl

$$\frac{odds(1)}{odds(0)} = \frac{P(Y = 1|x_1, \ldots, x_k)}{P(Y = 0|x_1, \ldots, x_k)} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})$$

má bezprostřední interpretaci. Porovnává pravděpodobnost jedničky (tj. výskyt sledovaného jevu při daných hodnotách kovariát) a nuly (nevýskyt sledovaného jevu při daných hodnotách kovariát). Anglickému označení **odds** odpovídá české označení **šance**. Hodnota šance není shora ohraničená, zdola však nulou. Pokud zlogaritmujeme šanci, dostaneme **logit**, který nabývá hodnot od mínus do plus nekonečna.

Nyní budeme předpokládat, že máme jedinou kovariátu x , která je také binární, takže nabývá dvou různých hodnot, které můžeme bez újmy na obecnosti označit jako 0 a 1. V

tom případě jde o kategoriální proměnnou, nebo–li x je umělá proměnná k dvouhodnotovému faktoru.

Za těchto podminek je šance pro $x = 0$ rovna

$$odds(0) = exp(\beta_0),$$

takže parametr β_0 je roven logitu pravděpodobnosti výskytu sledovaného jevu v bodě $x = 0$. Pro $x = 1$ dostaneme

$$odds(1) = exp(\beta_0 + \beta_1).$$

Poměr šancí (nebo také křížový poměr, anglicky *odds ratio*) pro binární x je pak roven

$$OR = \frac{odds(1)}{odds(0)} = \exp(\beta_1),$$

takže parametr β_1 je roven logaritmu poměru šancí. Odtud tedy dostáváme, že pokud pravděpodobnost sledovaného jevu nezávisí na hodnotě proměnné x, je poměr šancí roven jedné, takže platí

$$\beta_1 = 0.$$

I v případě, že vysvětlující proměnná je spojitá, má zajímavou interpretaci především parametr β_1, neboť

$$\frac{x+1}{x} = \frac{\beta_0 + \beta_1(x+1)}{\beta_0 + \beta_1 x} = \exp(\beta_1),$$

takže parametr β_1 vypovídá o změně vztažené k jednotkovému přírůstku nezávisle proměnné x, tentokrát je to změna logaritmu poměru šancí.

Příklad 2.7. V souboru „`beetle.RData`“ jsou uvedeny údaje o úmrtnosti *Potemníka skladištního* (*Tribolium confusum*) v reakci na sirouhlík CS_2. Datový soubor obsahuje tyto proměnné

`dose`	množství sirouhlíku (mg/l)
`population`	počet kusů ve zkoumaném vzorku
`killed`	počet mrtvých kusů ve zkoumaném vzorku

Řešení. Pro modelování závislosti použijeme logistický model, probitový model a model s komplementární log-log linkovací funkcí. Výsledky jsou znázorněny na Obr. 7.

3. Modely pro poissonovská data

Celočíselná data lze modelovat pomocí diskrétních rozdělení. V předchozí sekci jsme se zabývali alternativními a binomickými daty. Nyní soustřeďme pozornost na poissonovská data.

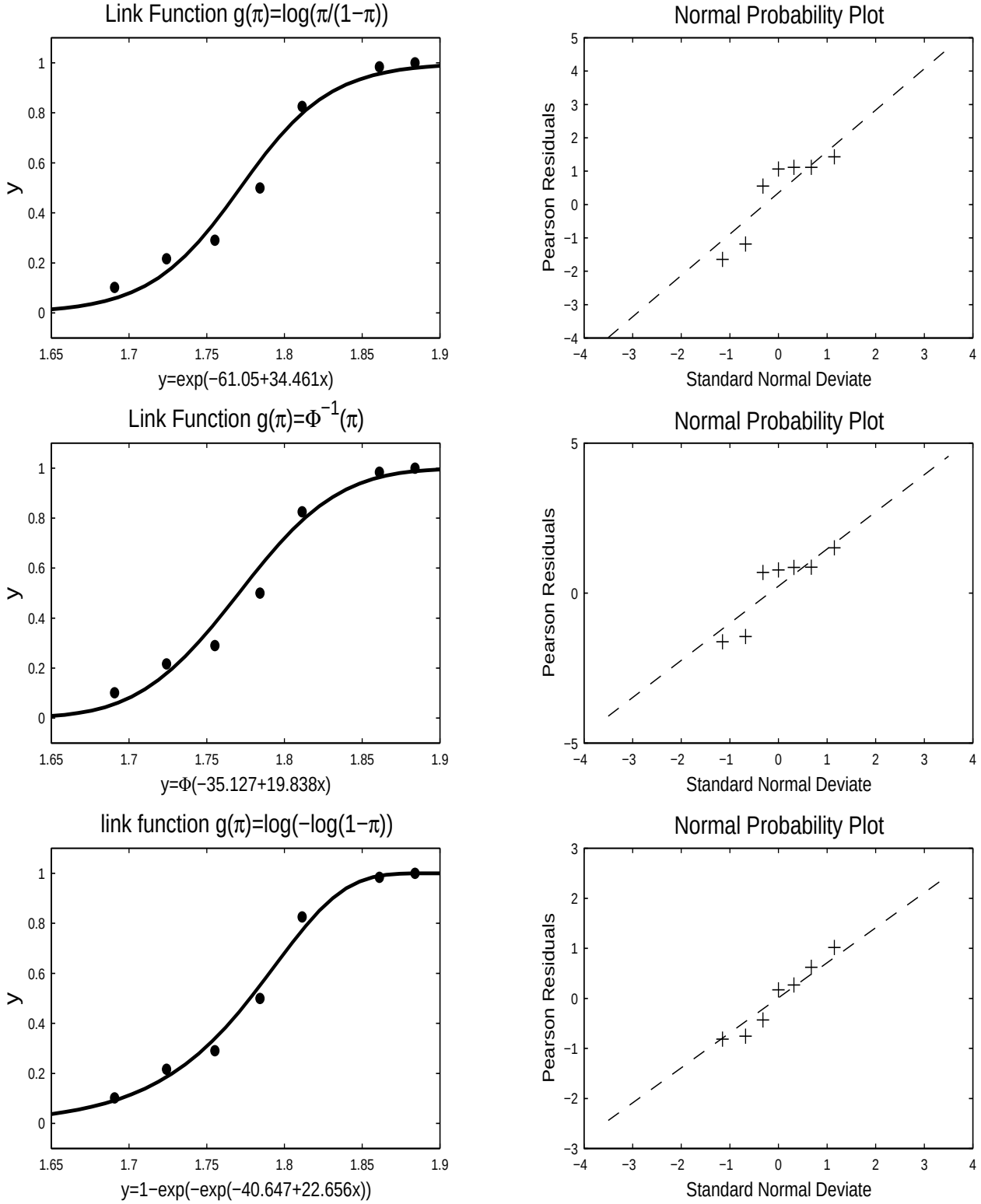
Předpokládejme, že náhodný výběr rozsahu n je z Poissonova rozdělení, tj.

$$Y_i \sim Po(\lambda_i) \sim f_Y(y) = P(Y_i = y) = \begin{cases} \frac{\lambda^y e^{-\lambda_i}}{y!} = & \lambda_i > 0; \quad y = 0, 1, 2, \ldots \\ 0 & \text{jinak} \end{cases}$$

přičemž

$$EY_i = DY_i = \lambda_i.$$

Poznámka 3.1. Dále se tímto rozdělením řídí náhodná veličina, kterou je **počet výskytu sledovaného jevu v určitém časovém intervalu délky** t (nebo počet výskytu sledovaného jevu na ploše velikosti t apod.).
Jestliže jsou splněny následující podmínky



Obrázek 7: Modely pro úmrtnost *Potemníka skladištního.*

a) jev může nastat v kterémkoliv časovém okamžiku,
b) počet výskytů jevu během časového intervalu závisí jen na jeho délce a ne na jeho počátku ani na tom, kolikrát jev nastoupil před jeho počátkem,
c) pravděpodobnost, že jev nastoupí více než jednou v intervalu délky t, konverguje k nule rychleji než t,
d) λ je **střední hodnota počtu výskytů jevu za časovou jednotku**

pak uvedená náhodná veličina má rozdělení $Po(\lambda)$.

Náhodnou veličinou, která má Poissonovo rozdělení, je tedy např.

- počet vadných výrobků ve velké sérii, jestliže pravděpodobnost vyrobení vadného výrobku je velmi malá
- počet těžkých dopravních úrazů za den v určitém městě
- počet zákazníků v prodejně během nějakého časového intervalu
- počet částic v jednotce plochy nebo objemu, např. počet částic v zorném poli mikroskopu
- počet telefonních volání v časovém intervalu t
- počet létavic pozorovaných během intervalu délky t

Předpokládejme opět, že náhodná veličina Y_i závisí na k veličinách $x_{i1}, \ldots, x_{ik}$, (tzv. **kovariáty**) a úkolem bude najít vztah mezi nimi, tj. hledáme funkci

$$\lambda_i = \lambda(\mathbf{x}_i) = \lambda(x_{i1}, \ldots, x_{ik}).$$

Protože chceme použít GLM modely, modelujeme pravděpodobnosti λ_i pomocí linkovacích funkcí

$$g(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}.$$

DEFINICE 3.2. Pokud v modelu uvažujeme identickou linkovací funkci, tj. platí

$$\lambda_i = \mathbf{x}_i'\boldsymbol{\beta},$$

mluvíme o **lineárním modelu**.

Avšak tento model má řadu nevýhod, především je třeba zajistit, aby $\mathbf{x}_i'\boldsymbol{\beta}$ nabývala pouze kladných hodnot, nejčastěji se proto volí následující dvě možnosti

DEFINICE 3.3. Pokud v modelu předpokládáme vztah

$$\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}),$$

tj. uvažujeme linkovací funkci

$$g_1(\lambda_i) = \log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta},$$

hovoříme o **log-lineárním modelu**.

DEFINICE 3.4. Pokud v modelu předpokládáme vztah

$$\lambda_i = (\mathbf{x}_i'\boldsymbol{\beta})^2,$$

tj. uvažujeme linkovací funkci

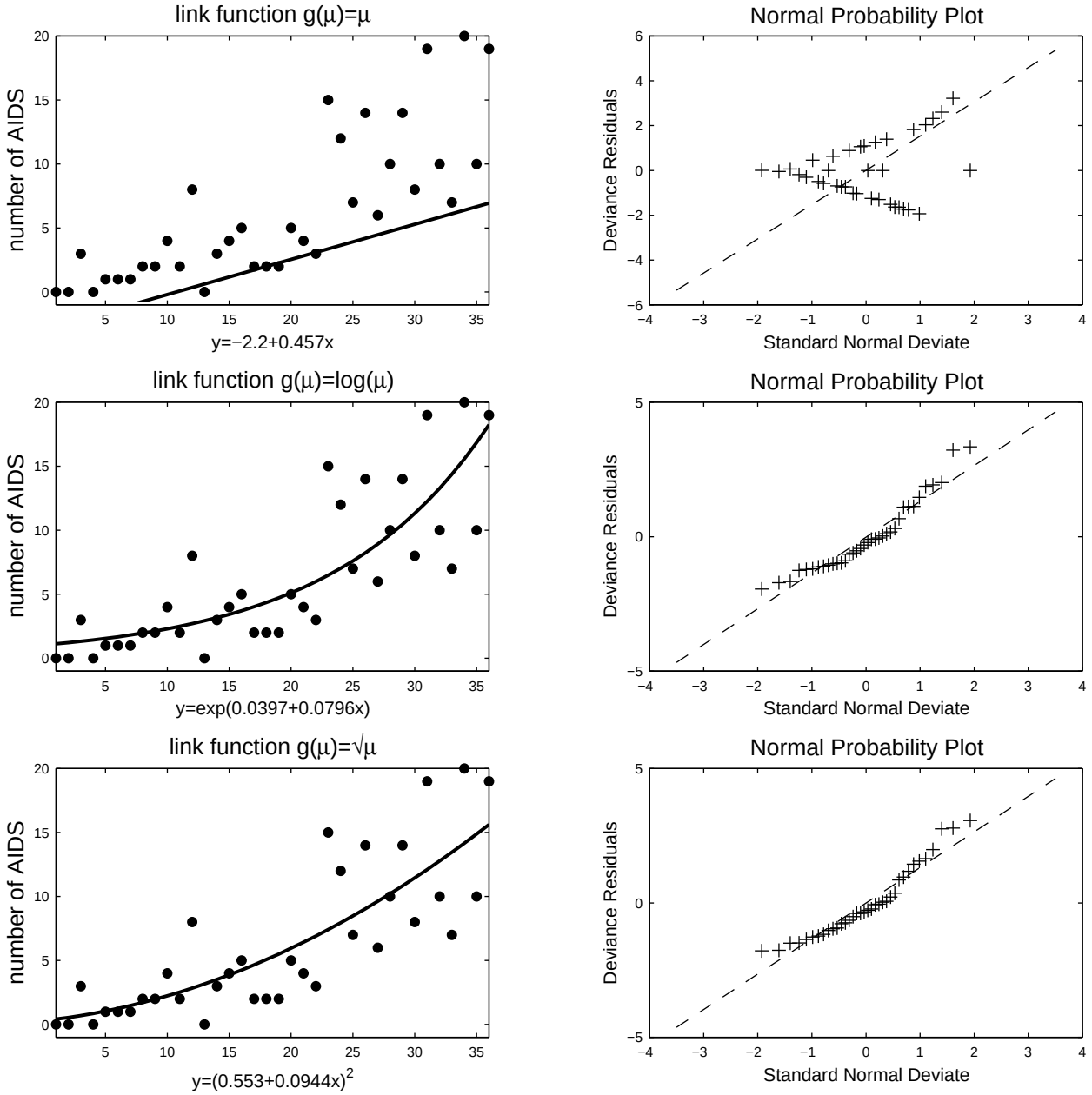
$$g_2(\lambda_i) = \sqrt{\lambda_i} = \mathbf{x}_i'\boldsymbol{\beta},$$

hovoříme o **odmocninovém modelu** (*square-root-linear model*).

Příklad 3.5. V souboru „`aids.RData`“ jsou uvedeny údaje o počtech nových případů AIDS ve Velké Británii za období prosinec 1982 až listopad 1985. Datový soubor obsahuje tyto proměnné

`month`	měsíc
`year`	rok
`number`	počet nových případů AIDS

Řešení. Pro modelování závislosti použijeme lineární model, log-lineární model a odmocninový model. Výsledky jsou znázorněny na Obr. 8.



Obrázek 8: Modely pro výskyt nových onemocnění AIDS ve Velké Británii.

3.1. Modelování binomických dat pomocí poissonovského modelu. V praxi často nastává situace, kdy sice máme data s binomickým rozdělením, avšak pravděpodobnost „úspěchu“ v jednotlivých pozorováních je velmi malá vzhledem k počtu těchto pozorování. V takovém případě již nelze použít binomického modelu, ale je třeba jej nahradit modelem poissonovským.

Připomeňme že pomocí Poissonova rozdělení $Po(\lambda)$ lze dobře aproximovat binomické rozdělení $Bi(n, \pi)$ za podmínek

$$n \to \infty \quad \& \quad \pi \to 0 \quad \& \quad n\pi \to \lambda < \infty,$$

obvykle se doporučuje $n > 30$ a $\pi < 0,1$.

Chceme-li tedy aproximovat binomické rozdělení $Bi(n_i, \pi_i)$ pomocí Poissonova rozdělení $Po(\lambda_i = n_i\pi_i)$ a přitom použijeme logaritmickou linkovací funkci, platí

$$\lambda_i = n_i\pi_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \Rightarrow \log(\lambda_i) = \underbrace{\log(n_i)}_{\text{tzv. „offset“}} + \log(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta}.$$

Z výše uvedeného vztahu je tedy vidět, že při definici zobecněného lineárního modelu pomocí poissonovské třídy rozdělení je k ostatním prediktorům ještě třeba přidat logaritmus proměnné udávající velikosti populace (pomocí příkazu `offset`).

4. Problematika příliš velkého nebo příliš malého rozptylu

V praktickém modelování často narážíme na problémy s příliš velkou variabilitou dat (*overdispersion*) nebo příliš malou variabilitou dat (*underdispersion*). Existuje řada možných vysvětlení, proč k tomu dochází. Tak například v biologických studiích může být overdispersion důsledkem agregovaného výskytu organismů. Nebo je tento jev důsledkem závislosti v datech, které standardní model nepředpokládá. Příliš malý či velký rozptyl může vzniknout také nezařazením některé důležité vysvětlující proměnné.

Popišme podrobněji tento jev. Předpokládáme, že náhodný výběr $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$ z rozdělení exponenciálního typu se řídí GLM modelem, tj. má sdruženou pravděpodobnostní funkci nebo sdruženou hustotu tvaru

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \theta_i) = \exp\left\{ \sum_{i=1}^{n} \frac{y_i\theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right\}.$$

Předpokládejme, že pro hustotu exponenciálního typu platí

$$\psi_i(\phi) = \frac{\phi}{\omega_i} > 0,$$

kde $\omega_i > 0$ jsou známé **apriorní váhy** a $\phi > 0$ je neznámý tzv. **faktor měřítka** (scale factor) nebo bývá též nazýván **rušivý parametr**.

Při testování vhodnosti modelu hraje důležitou roli tzv. (škálová) deviace, kterou můžeme vyjádřit

$$\begin{aligned} D &= 2\left[l^*(\widehat{\boldsymbol{\beta}}_{max}; \mathbf{Y}) - l^*(\widehat{\boldsymbol{\beta}}; \mathbf{Y}) \right] \\ &= \frac{1}{\phi} 2 \sum_{i=1}^{n} \omega_i \left[Y_i(\hat{\theta}_{i,max} - \hat{\theta}_i) - \gamma(\hat{\theta}_{i,max}) + \gamma(\hat{\theta}_i) \right] \\ &= \frac{1}{\phi} D^* \end{aligned}$$

a D^* nazveme **neškálovou deviací** (unscaled deviance). Protože platí

$$D = \frac{1}{\phi} D^* \overset{A}{\sim} \chi^2(n-k) \quad \Rightarrow \quad ED = \frac{1}{\phi} ED^* \approx n - k,$$

neboť střední hodnota χ^2 rozdělení je rovna počtu stupňů volnosti, pak

$$\hat{\phi}_{D^*} = \frac{D^*}{n-k}.$$

Další často používanou mírou vhodnosti modelu je tzv. zobecněná Pearsonova statistika

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \overset{A}{\sim} \chi^2(n-k)$$

a proto dalším momentovým odhadem založeným na této statistice je

$$\hat{\phi}_{X^2} = \frac{X^2}{n-k}.$$

Přehled rušivých parametrů pro některá rozdělení exponenciálního typu je dán v následující tabulce

Rozdělení	ϕ
Normální rozdělení	σ^2
Poissonovo rozdělení	1
Binomické rozdělení	1
Gamma rozdělení	$1/\alpha$

Problém s příliš velkou či malou variabilitou se týká těch rozdělení, u kterých má být scale parametr roven jedné, tj. binomického a Poissonova rozdělení. Pokud pro reálná data dojde k tomu, že například pro binomické či Poissonovo rozdělení je rozptyl větší než střední hodnota, pak jde o *overdispersion*. Pokud je například u dat, pro která jsme předpokládali Poissonovo rozdělení, rozptyl naopak menší než střední hodnota, pak jde o *underdispersion*. V těchto případech není hodnota disperzního (scale) parametru ϕ (jakožto poměru $\frac{DY}{EY}$) rovna 1. Ve výpisu výsledků modelu nás na tuto situaci upozorní výrazně větší (menší) hodnota reziduální (tedy nevysvětlené) deviace ve srovnání s reziduálním počtem stupňů volnosti, což je střední hodnota χ^2 rozdělení.

V prostředí R je k řešení tohoto problému k dispozici modifikovaná volba pro třídu exponenciálního rozdělení. V případě binomického rozdělení máme možnost volby

`family=quasibinomial`

a pro Poissonovo rozdělení

`family=quasipoisson`.

Nejde o nový typ exponenciálního rozdělení, ale o změnu ve výpočtu druhého momentu, pro jehož odhad se použije jednoduchý momentový odhad disperzního parametru ϕ. Výsledná korekce rozptylu je pak důležitá při testování hypotéz, neboť zohledňuje vyšší/nižší variabilitu v datech a zabraňuje tak nadbytku/nedostatku falešně pozitivních výsledků testů hypotéz o parametrech modelu.

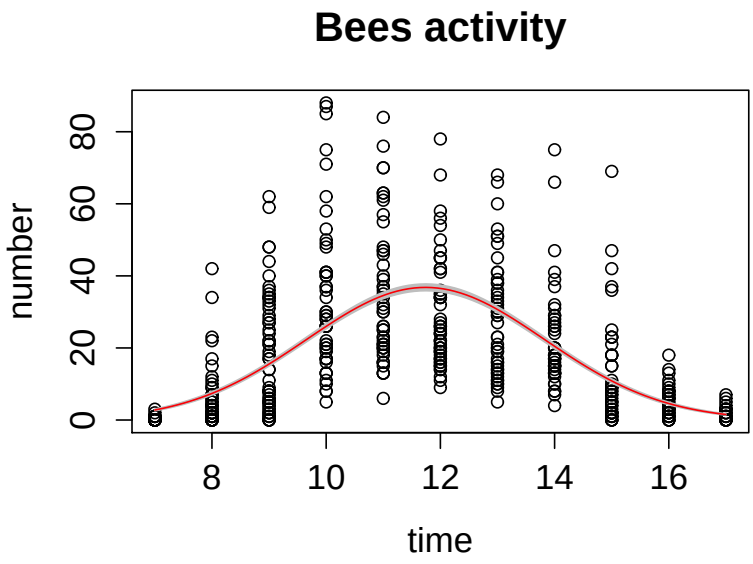
Příklad 4.1. V souboru „`bees.RData`“ jsou uvedeny údaje o aktivitě včel v závislosti na čase. Jednou z důležitých charakteristik při zkoumání včelí aktivity je počet včel, které opustí úl kvůli práci ve vnějším prostředí. Studie se zabývala měřením této veličiny během několika slunečných dní v závislosti na čase během dne. Datový soubor obsahuje tyto proměnné

`number` počet včel, které opustily úl

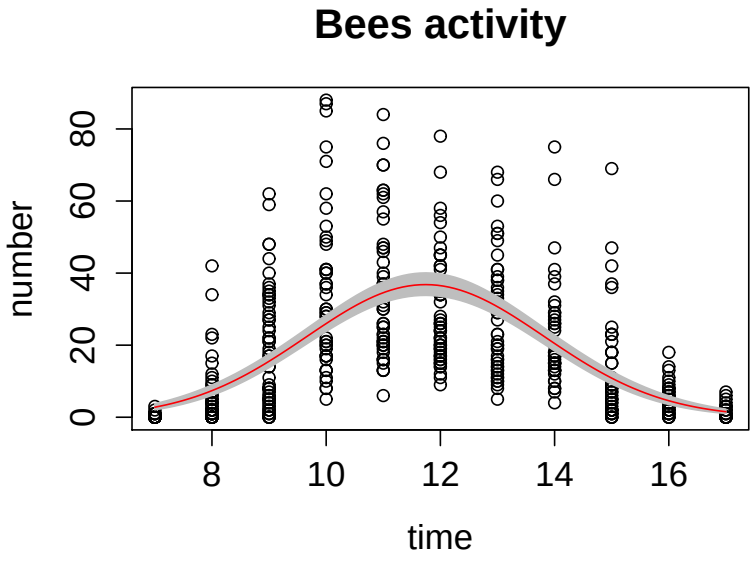
`time` čas, kdy byl tento údaj zaznamenán

Modelujte závislost počtu včel, které opustí úl, na čase během dne.

Řešení. Budeme předpokládat, že závisle proměnná `number` má Poissonovo rozdělení a pro modelování závislosti použijeme poissonovský model. Jako linkovací funkci zvolíme kanonickou, tj. logaritmus. Do modelu vstupuje jediná vysvětlující proměnná `time` a přidáme také její druhou mocninu. Po výpočtu všech potřebných parametrů je vidět, že hodnota reziduální deviace (4 879,3) je nepoměrně vyšší než počet stupňů volnosti (501), což je střední hodnota. Je tedy zřejmé, že došlo k „overdispersion“ a v jazyce *R* je třeba volit `family=quasipoisson`. Použití této volby neovlivňuje odhady koeficientů, ale mění jejich odhady variability, což se projeví např. v intervalu spolehlivosti. To je vidět i z grafického srovnání obou výsledků, viz Obr. 9 a Obr. 10.



Obrázek 9: Odhad regresní funkce **bez** vyrovnání se s problematikou velkého rozptylu.



Obrázek 10: Odhad regresní funkce **s** vyrovnáním se s problematikou velkého rozptylu.

5. Modely pro multinomická data

5.1. Kontingenční tabulky. Mějme náhodný výběr $\mathbf{Y} = \mathbf{Y}_n = (Y_1, \ldots, Y_n)'$ rozsahu n, pro který $n = J \cdot K$, kde $J, K \in \mathbb{N}^+$ jsou kladná přirozená čísla, tj. náhodný výběr lze rozepsat takto

$$\mathbf{Y} = \mathbf{Y}_n = (Y_1, \ldots, Y_n)' = (Y_{11}, \ldots, Y_{1K}, \ldots, Y_{J1}, \ldots, Y_{JK})'.$$

Předpokládejme, že náhodný výběr $\mathbf{Y}$ je z Poissonova rozdělení, tj.

$$Y_{jk} \sim Po(\lambda_{jk}) \qquad j = 1, \ldots, J;\ k = 1, \ldots, K$$

s tzv. **celkovou** dodatečnou podmínkou

$$N = \sum_{j=1}^{J}\sum_{k=1}^{K} y_{jk} \qquad N \in \mathbb{N}^{+},$$

kde y_{jk} jsou realizace náhodných veličin Y_{jk}.
Pak sdružená (nepodmíněná) pravděpodobnostní funkce náhodného vektoru $\mathbf{Y}$ je rovna

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= P(Y_{11} = y_{11}, \ldots, Y_{1K} = y_{1K}, \ldots, Y_{J1} = y_{J1}, \ldots, Y_{JK} = y_{JK}) \\ &= \begin{cases} \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\lambda_{jk}^{y_{jk}} e^{-\lambda_{jk}}}{y_{jk}!} & y_{jk} = 0, 1, 2, \ldots;\ j = 1, \ldots, J;\ k = 1, \ldots, K, \\ 0 & \text{jinak.} \end{cases} \end{aligned}$$

Součet nezávislých náhodných veličin s Poissonovým rozdělením má opět Poissonovo rozdělení, tj.

$$Z_{\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk} \sim Po\left(\lambda_{\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K}\lambda_{jk}\right) \sim p_Z(z) = \begin{cases} \frac{\lambda_{\cdot\cdot}^{z} e^{-\lambda_{\cdot\cdot}}}{z!} & z = 0, 1, 2, \ldots, \\ 0 & \text{jinak.} \end{cases}.$$

Nás ovšem zajímá rozdělení náhodného vektoru $\mathbf{Y}$ za podmínky $Z_{\cdot\cdot} = N$, (tj. že součet jeho složek je roven pevně danému kladnému přírozenému číslu N) s pravděpodobnostní funkcí $p_{\mathbf{Y}|Z_{\cdot\cdot}=N}$, kterou lze snadno vypočítat ze vztahu

$$p_{\mathbf{Y}|Z_{\cdot\cdot}=N}(\mathbf{y}) = \begin{cases} \frac{p_{\mathbf{Y}}(\mathbf{y})}{p_Z(N)} & p_{Z_{\cdot\cdot}}(N) \neq 0, \\ 0 & jinak, \end{cases}$$

kterou s využitím vztahů

$$\begin{aligned} e^{-\lambda_{\cdot\cdot}} &= \prod_{j=1}^{J}\prod_{k=1}^{K} e^{-\lambda_{jk}} \\ \lambda_{\cdot\cdot}^{N} &= \lambda_{\cdot\cdot}^{\sum_{j=1}^{J}\sum_{k=1}^{K} y_{jk}} = \prod_{j=1}^{J}\prod_{k=1}^{K}\lambda_{\cdot\cdot}^{y_{jk}} \end{aligned}$$

lze upravit takto

$$p_{\mathbf{Y}|Z_{\cdot\cdot}=N}(\mathbf{y}) = \begin{cases} \frac{\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\lambda_{jk}^{y_{jk}} e^{-\lambda_{jk}}}{y_{jk}!}}{\frac{\lambda_{\cdot\cdot}^{N} e^{-\lambda_{\cdot\cdot}}}{N!}} = & \text{pro} \quad y_{jk} = 0, 1, \ldots, N;\ j = 1, \ldots, J; \\ & \qquad\qquad k = 1, \ldots, K, \\ = N! \prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\left(\frac{\lambda_{jk}}{\lambda_{\cdot\cdot}}\right)^{y_{jk}}}{y_{jk}!} & \sum_{j=1}^{J}\sum_{k=1}^{K} y_{jk} = N \\ 0 & \text{jinak.} \end{cases}$$

a položíme-li

$$\frac{\lambda_{jk}}{\lambda_{\cdot\cdot}} = \pi_{jk} \qquad \text{pak} \qquad \sum_{j=1}^{J}\sum_{k=1}^{K}\pi_{jk} = 1.$$

Z předchozích úvah vidíme, že platí následující věta.

VĚTA 5.1. *Rozdělení náhodného vektoru* $\mathbf{Y}$ *za podmínky* $Z_{..} = N$ *je* ***multinomické*** *s pravděpodobnostní funkcí*

$$p_{\mathbf{Y}|Z_{..}=N}(\mathbf{y}) = \begin{cases} N! \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{\pi_{jk}^{y_{jk}}}{y_{jk}!} & pro \quad y_{jk} = 0, 1, \ldots, N; \quad j = 1, \ldots, J; \\ & \qquad\qquad\qquad\quad k = 1, \ldots, K, \\ & \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} = N \qquad \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{jk} = 1 \\ 0 & jinak \end{cases},$$

tj.

$$\mathbf{Y}|Z_{..} = N \quad \sim \quad Mn(N, \pi_{11}, \ldots, \pi_{1K}, \ldots, \pi_{J1}, \ldots, \pi_{JK}) \,,$$

přičemž

$$\begin{aligned} EY_{jk} &= N\pi_{jk} \\ DY_{jk} &= N\pi_{jk}(1 - \pi_{jk}) \\ C(Y_{jk}, Y_{j'k'}) &= -N\pi_{jk}\pi_{j'k'} \end{aligned}$$

Poznámka 5.2. Multinomické rozdělení popisuje situaci, kdy máme $n = J \times K$ neslučitelných jevů, které označme

$$(AB)_{11}, \ldots, (AB)_{JK}.$$

Jednotlivé jevy mohou nastat v každém z N nezávislých pokusů s pravděpodobnostmi

$$\pi_{11}, \ldots, \pi_{JK} \qquad \text{přičemž} \qquad \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{jk} = 1.$$

Multinomické rozdělení je zobecněním binomického rozdělení a je patrně nejdůležitějším diskrétním mnohorozměrným rozdělením. Svým významem by se dalo přirovnat k mnohorozměrnému normálnímu rozdělení, jemuž se podobá především díky dvěma vlastnostem: podmíněná i marginální rozdělení jsou opět multinomická. Realizace náhodných veličin i teoretické pravděpodobnosti lze uspořádat do tzv. **kontingenční tabulky**:

Kontingenční tabulka četností

faktor	faktor B				
A	B_1	B_2	$\cdots$	B_K	$\sum$
A_1	y_{11}	y_{12}	$\cdots$	y_{1K}	$N_{1.}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2K}	$N_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	y_{J1}	y_{J2}	$\cdots$	y_{JK}	$N_{J.}$
$\sum$	$N_{.1}$	$N_{.2}$	$\cdots$	$N_{.K}$	$N = N_{..}$

Kontingenční tabulka pravděpodobností

faktor	faktor B				
A	B_1	B_2	$\cdots$	B_K	$\sum$
A_1	π_{11}	π_{12}	$\cdots$	π_{1K}	$\pi_{1.}$
A_2	π_{21}	π_{22}	$\cdots$	π_{2K}	$\pi_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	π_{J1}	π_{J2}	$\cdots$	π_{JK}	$\pi_{J.}$
$\sum$	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.K}$	$\pi_{..} = 1$

Čísla $N_{j.}$ a $N_{.k}$ se nazývají **marginální četnosti** a $\pi_{j.}$ a $\boxed{\pi_{.k}}$ jsou **marginální pravděpodobnosti**. Tabulky popisujeme slovně tak, že říkáme, že N jednotek bylo klasifikováno podle znaku A do J tříd a podle znaku B do K tříd. V praxi kontingenční tabulka vzniká tak, že na daných objektech sledujeme dva znaky (faktory). Vybereme-li náhodně N objektů, můžeme výsledky shrnout do kontingenční tabulky typu $J \times K$

Nejčastěji se v kontingenčích tabulkách testuje hypotéza, že

faktory A a B jsou nezávislé

tj.

faktor A	faktor B			$\sum$
	$\cdots$	B_k	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_j	$\cdots$	$\pi_{j.}\pi_{.k}$	$\cdots$	$\pi_{j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum$	$\cdots$	$\pi_{.k}$	$\cdots$	1

$$\pi_{jk} = \pi_{j.}\pi_{.k}, \text{ takže potom } EY_{jk} = N\pi_{j.}\pi_{.k}\ , \text{ přičemž } \sum_{j=1}^{J}\pi_{j.} = \sum_{k=1}^{K}\pi_{.k} = 1.$$

Poznámka 5.3. Poznamenejme, že existuje samozřejmě více testů v kontingenčních tabulkách, např. test homogenity. Případně lze tyto testy rozšířit na vícerozměrné kontingenční tabulky. Pro více informací o modelech pro tyto případy odkazujeme čtenáře na [**5**].

5.2. Log-lineární modely. Zkusme se na kontingenční tabulky dívat pohledem zobecněných lineárních modelů. V předchozím případě hypotéza nezávislosti vede k **multiplikativnímu modelu**, který logaritmováním lze převést na model **lineární** a odtud pramení všeobecně zažité pojmenování *log-lineární modely*.

Nyní pro předchozí model hledejme odpovídající GLM model:
Pro model s **celkovou dodatečnou podmínkou** lze hypotézu o **nezávislosti dvou faktorů** definovat takto

$$EY_{jk} = N\pi_{j.}\pi_{.k}\ , \qquad \text{přičemž} \qquad \sum_{j=1}^{J}\pi_{j.} = 1 \qquad \text{a} \qquad \sum_{k=1}^{K}\pi_{.k} = 1.$$

V GLM s **log-lineární linkovací funkcí** máme $\eta_{jk} = \log EY_{jk} = \mathbf{x}'_{jk}\boldsymbol{\beta}$, tedy

$$\eta_{jk} = \log EY_{jk} = \log(N\pi_{j.}\pi_{.k}) = \underbrace{\mu}_{=\log N} + \underbrace{\alpha_j}_{=\log \pi_{j.}} + \underbrace{\beta_k}_{=\log \pi_{.k}}\ .$$

Pokud bychom nepředpokládali nezávislost faktorů A a B, dostaneme **maximální model**

$$\eta_{jk} = \log EY_{jk} = \log(N\pi_{jk}) = \underbrace{\mu}_{=\log N} + \underbrace{\alpha_j + \beta_k + (\alpha\beta)_{jk}}_{=\log \pi_{jk}}$$

Vidíme, že základní i maximální modely jsou přeparametrizovány. Proto se musí upravit, například tak, že položíme

$$\alpha_1 = \beta_1 = 0 \quad \text{resp. navíc} \quad (\alpha\beta)_{11} = 0 \qquad\qquad \text{tzv. metoda horního rohu}$$

nebo

$$\sum_{j=1}^{J}\alpha_j = 1 \quad \text{a} \quad \sum_{k=1}^{K}\beta_k = 0 \quad \text{resp. navíc} \quad \sum_{j=1}^{J}\sum_{k=1}^{K}(\alpha\beta)_{jk} = 0.$$

Všimněme si **počtů parametrů** pro jednotlivé úrovně

μ	obecná střední hodnota	1
α	hlavní efekt	$J-1$
β	hlavní efekt	$K-1$
$\alpha\beta$	interakce prvního řádu	$(J-1)(K-1)$
	celkem	$n = JK$

Vidíme, že hypotéza nezávislosti dvou faktorů v kontingenčních tabulkách je ekvivalentní s hypotézou neexistence interakcí v analýze rozptylu (deviace), tj.

$$H_0: \ (\alpha\beta)_{jk} = 0 \qquad j = 1, \ldots, J; \ \ k = 1, \ldots, K.$$

V log-lineárních modelech jsou obvykle výrazy vyšších řádů definovány jako odchylky od výrazů nižšího řádu. Tak například v základním modelu výraz α_j reprezentuje rozdíl efektu řádku j od obecné střední hodnoty μ. Takže model je **hiearchický** v tom smyslu, že výrazy vyšších řádů nejsou obsaženy ve výrazech nižších řádů.

Shrňme předchozí výsledky:

$$GLM_{max} \quad H_0 : EY_{jk} = N\pi_{jk} \Rightarrow \eta_{jk} = \log EY_{jk} = \underbrace{\mu}_{\log N} + \underbrace{\alpha_j + \beta_k + (\alpha\beta)_{jk}}_{\log \pi_{jk}}$$

$$GLM \qquad H_0 : EY_{jk} = N_{j.}\pi_{j.}\pi_{.k} \Rightarrow \eta_{jk} = \log EY_{jk} = \log(N\pi_{j.}\pi_{.k}) = \underbrace{\mu}_{=\log N} + \underbrace{\alpha_j}_{=\log \pi_{j.}} + \underbrace{\beta_k}_{=\log \pi_{.k}}.$$

V této kapitole jsme se pokusili o stručný popis modelování kontingenčních tabulek v souvislosti se zobecněnými lineárními modely. Pro podrobnější analýzu závislosti náhodných veličin pomocí kontingenčních tabulek odkazujeme čtenáře na další kapitolu.

Příklad 5.4. V následující kontingenční tabulce jsou obsaženy údaje studie 400 pacientů o počtech různých typů onemocnění rakovinou kůže (*Malignant Melanoma*) v závislosti na části těla, kde se vyskytují.

	Část těla		
Typ rakoviny	končetiny	hlava a krk	trup
Hutchinson's melanotic freckle	10	22	2
neurčitý	28	11	17
Nodular	73	19	33
Superficial spreading melanoma	115	16	54

Na hladině významnosti $\alpha = 0,05$ testujte hypotézu, zda typ rakoviny kůže závisí na části těla, kde se vyskytuje.

Řešení. Nejprve definujeme oba log-lineární modely, tj. model `m1`, který předpokládá nezávislost obou faktorů a model `m2`, který počítá i s interakcemi. Model `m1` je tedy submodelem modelu `m2`. K testování využijeme analýzu deviace, Pearsonův test. Jeho p-hodnota vychází $2,05 \times 10^{-9}$ a proto zamítáme hypotézu o nezávislosti typu rakoviny kůže na části těla, kde se vyskytuje. Výsledky obou modelů lze také znázornit pomocí mozaikového grafu. Graf pro model `m1` je znázorněn na Obr. 11, graf pro model `m2` je vykreslen na Obr. 12.

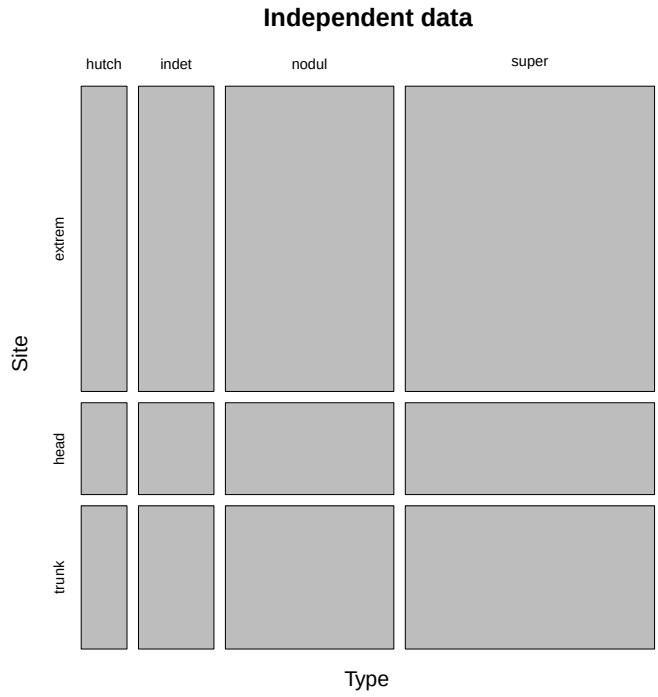
Úlohy k procvičení

Cvičení 5.1. V souboru „`heart.RData`“ jsou uvedena data o přítomnosti infarktu myokardu v závislosti na věku pacienta. Datový soubor obsahuje tyto proměnné:

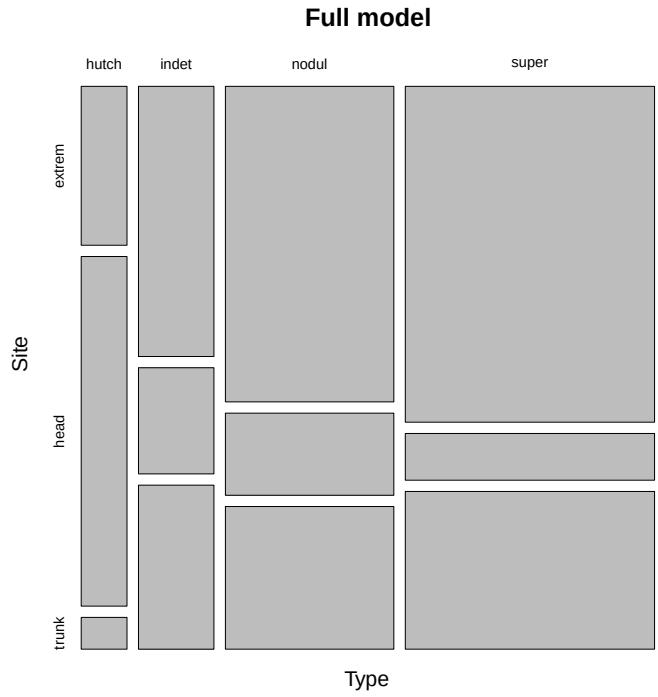
`age` věk pacienta (roky)

`chd` indikátor infarktu (1 – nastal, 0 – nenastal)

Pro modelování závislosti použijte logistický model, probitový model a model s komplementární log-log linkovací funkcí. Výsledky vykreslete do obrázku.



Obrázek 11: Mozaikový graf pro model, který předpokládá nezávislost.



Obrázek 12: Mozaikový graf pro model s interakcemi.

Cvičení 5.2. V souboru „`nemocnice.RData`“ jsou uvedeny údaje o zotavení pacientů v závislosti na závažnosti onemocnění a nemocnici, ve které se léčili. Datový soubor obsahuje tyto proměnné:

`Infection_Severity`	vážnost onemocnění
`Treatment_Outcome`	indikátor uzdravení (1 – zdravý, 0 – smrt)
`Hospital`	typ nemocnice (1, 2, 3)

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku.

Cvičení 5.3. V souboru „`cancer.RData`“ jsou uvedeny údaje o počtu onemocnění rakovinou kůže u žen v závislosti na věku a oblasti v USA, ve které pacientky žily. Datový soubor obsahuje tyto proměnné:

`Cases`	počet onemocnění
`Town`	město (0 – Minneapolis (Minnesota), 1 – Dallas (Texas))
`Age`	věková skupina pacientky
`Population`	celkový počet žen dané věkové skupiny v příslušném městě

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku. Porovnejte pravděpodobnost vzniku onemocnění u 60-ti leté pacientky žijící v Minneapolisu s pravděpodobností pro stejně starou pacientku žijící v Dallasu.

[Minneapolis: 0.00117, Dallas: 0.00276.]

Cvičení 5.4. V souboru „`car_income.RData`“ jsou uvedeny údaje o koupi nového auta během posledních 12-ti měsíců v závislosti na příjmu domácnosti a stáří původního auta. Datový soubor obsahuje tyto proměnné:

`purchase`	indikátor nákupu nového auta (1 – ano, 0 – ne)
`income`	roční příjem domácnosti (v tis. dolarů)
`age`	stáří původního auta (roky)

Nejprve vykreslete závislosti proměnné `purchase` na ostatních. Pro modelování závislosti nalezněte vhodný logistický model. Jsou všechny proměnné statisticky významné? Znovu modelujte s použitím proměnné `age` jako `factor`. Opět sledujte statistickou významnost `age`. Vyzkoušejte tuto proměnnou zakomponovat do modelu jako `factor` s méně úrovněmi. Výsledky vykreslete do obrázku.

Cvičení 5.5. V souboru „`druhy.RData`“ jsou k dispozici data, která se týkají dlouhodobého zemědělského experimentu. Bylo sledováno 90 pozemků (pastvin) o rozloze 25 m × 25 m, lišících se v biomase, pH půdy a druhové bohatosti (počet rostlinných druhů na celém pozemku). Je dobře známo, že s rostoucí biomasou dochází k poklesu druhové bohatosti. Ale zůstává otázka, zda rychlost poklesu nesouvisí s úrovní pH v půdě. Proto byly jednotlivé pozemky klasifikovány podle hodnoty pH v půdě do tří úrovní (nízká, střední a vysoká úroveň) a do experimentu bylo vybráno vždy po 30 pozemcích pro každou úroveň. Spojitá veličina `Biomass` je dlouhodobým průměrem naměřených červnových hodnot biomasy. Datový soubor obsahuje tyto proměnné:

`pH`	úroveň pH v půdě (`low` – nízká, `mid` – střední, `high` – vysoká)
`Biomass`	množství biomasy
`species`	počet rostlinných druhů

Nejprve vykreslete závislosti proměnné `species` na ostatních. Pro modelování závislosti nalezněte vhodný poissonovský model. Vyzkoušejte postupně logaritmickou, identickou a

odmocninovou linkovací funkci. Jsou všechny proměnné statisticky významné? Pokud ne, zkuste modely zjednodušit a pomocí analýzy deviace rozhodněte, zda takové zjednodušení je možné. Získané výsledné modely vykreslete do obrázku. Pomocí všech modelů odhadněte počet rostlinných druhů na pozemku s hodnotou biomasy 9 a střední úrovní pH v půdě.

[Odhady počtu druhů pro log link: 8,895, identity link: 4,513, sqrt link: 7,414.]

Cvičení 5.6. V souboru „`sharks.RData`“ jsou k dispozici data, která popisují počty napadení žraloky na Floridě v letech 1946 až 1999. Známe také velikost populace. Datový soubor obsahuje tyto proměnné:

`Year`	rok
`Population`	velikost populace
`Attacks`	počet napadení žraloky
`Fatalities`	počet úmrtí způsobených žraloky

Nejprve vykreslete bodový graf počtu napadení na 1 milión obyvatel v závislosti na čase. Pro modelování použijte binomický i poissonovský model s kanonickou linkovací funkcí. Pro matici plánu uvažujte kubický polynom v proměnné `Year`. Predikce obou modelů i s intervalem spolehlivosti pro regresní funkci vykreslete do obrázku. Zkoumejte také, jestli nenastal problém příliš velkého nebo příliš malého rozptylu. Pokud ano, předefinujte model a výsledky znovu vykreslete do obrázku. Pomocí výsledného modelu odhadněte, kolik útoků (na 1 milión obyvatel) způsobí žraloci na Floridě v roce 2013 a také v jakém intervalu se tato hodnota s 95% pravděpodobností bude pohybovat.

[Nastal problém příliš velkého rozptylu. Odhad: 33,96 útoků na 1 milión obyvatel, interval spolehlivosti: $[3,207; 359,55]$.]

Cvičení 5.7. V následující kontingenční tabulce jsou obsaženy údaje o počtech různých typů onemocnění horních cest dýchacích (*Respiratory Tract Infections*) v závislosti na čase.

	Časové období				
Diagnóza	1-3/96	4-6/96	7-9/96	10-12/96	1-3/97
Acute bronchitis	113	58	40	108	100
Acute sinusitis	99	37	23	50	32
URI	410	228	125	366	304
Pneumonia	60	43	30	56	45

Na hladině významnosti $\alpha = 0,05$ testujte hypotézu, zda onemocnění horních cest dýchacích závisí na čase.

[závisí]

KAPITOLA 9

Analýza závislosti dvou veličin

Základní informace

(1) V následující kapitole se budeme zabývat otázkou, zda dvě náhodné veličiny jsou stochasticky nezávislé. Testování hypotézy o nezávislosti se provádí různými způsoby podle toho, jakého typu jsou dané náhodné veličiny – zda jsou nominální, ordinální, intervalové či poměrové. Jednotlivé případy budou podrobněji rozebrány a uvedeny příslušné testy. Bude také věnován prostor zjišťování intenzity případné závislosti sledovaných dvou veličin. K tomuto účelu budou zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1 (resp. od -1 do 1).

(2) Předpokládá se znalost základních pojmů z teorie pravděpodobnosti a matematické statistiky – náhodná veličina, číselné charakteristiky náhodné veličiny, stochastická nezávislost náhodných veličin, testování hypotéz.

Výstupy z výukové jednotky

Studenti

- umí testovat nezávislost nominálních veličin
- umí určit Cramérův koeficient pro měření síly závislosti
- testují nezávislost ve čtyřpolních tabulkách
- umí testovat nezávislost ordinálních veličin
- testují nezávislost intervalových či poměrových veličin
- umí porovnat koeficient korelace s danou konstantou
- umí porovnat dva korelační koeficienty

1. Motivace

Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny jsou stochasticky nezávislé. Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá nebo zda počet dnů absence a věk pracovníka jsou nezávislé. Testování hypotézy o nezávislosti se provádí různými způsoby podle toho, jakého typu jsou dané náhodné veličiny – zda jsou nominální, ordinální, intervalové či poměrové.
Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1 (resp. od -1 do 1). Čím je takový koeficient bližší 1 (resp. -1), tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

Poznámka 1.1. Většina textu v této kapitole byla převzata z [**3**]. Pro podrobnější studium tohoto tématu proto odkazujeme na tento zdroj.

2. Testování nezávislosti nominálních veličin

V této sekci se budeme zabývat stochastickou nezávislostí náhodných veličin nominálního typu. Připomeňme si nejprve, co tento pojem znamená. Nominální proměnná je taková, o jejíž dvou hodnotách můžeme pouze říci, zda jsou stejné či různé (škola, fakulta, obor, krevní

skupiny: A, B, O, A/B), tj. obsahová interpretace je možná jenom u relace rovnosti. Hodnotami mohou být texty (písmena), případně i číselné kódy.

NÁVOD 2.1 (Popis testu). Nechť X, Y jsou dvě nominální náhodné veličiny. Nechť X nabývá variant $x_{[1]}, \ldots, x_{[r]}$ a Y nabývá variant $y_{[1]}, \ldots, y_{[s]}$. Pořídíme dvourozměrný náhodný výběr rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y). Zjištěné absolutní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})$ uspořádáme do kontingenční tabulky:

	y	$y_{[1]} \ldots y_{[s]}$	$n_{j.}$
x	n_{jk}		
$x_{[1]}$		$n_{11} \ldots n_{1s}$	$n_{1.}$
$\vdots$		$\ldots \ldots \ldots$	$\ldots$
$x_{[r]}$		$n_{r1} \ldots n_{rs}$	$n_{r.}$
$n_{.k}$		$n_{.1} \ldots n_{.s}$	n

Testujeme hypotézu $H_0 : X, Y$ jsou stochasticky nezávislé náhodné veličiny proti $H_1 : X, Y$ nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$$K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}.$$

Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$. Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

DEFINICE 2.2. Výraz $\frac{n_{j.}n_{.k}}{n}$ se nazývá **teoretická četnost**.

Poznámka 2.3 (Podmínka dobré aproximace)**.** Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti aspoň v 80% případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20% neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

Dále se budeme zabývat intenzitou případné závislosti sledovaných veličin. K tomuto účelu byl zkonstruován Cramérův[1] koeficient.

DEFINICE 2.4. **Cramérův koeficient** je tvaru

$$V = \sqrt{\frac{K}{n(m-1)}},$$

kde $m = \min\{r, s\}$. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y. Čím blíže je 0, tím je tato závislost volnější.

Příklad 2.5. V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

[1] Carl Harald Cramér (1893 – 1985). Švédský matematik.

Typ školy	Sociální skupina				$n_{j.}$
	I	II	III	IV	
univerzitní	50	30	10	50	140
technický	30	50	20	10	110
ekonomický	10	20	30	50	110
$n_{.k}$	90	100	60	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení.

$\frac{n_{1.}n_{.1}}{n} = \frac{140\cdot 90}{360} = 35$, $\frac{n_{1.}n_{.2}}{n} = \frac{140\cdot 100}{360} = 38,9$, $\frac{n_{1.}n_{.3}}{n} = \frac{140\cdot 60}{360} = 23,3$,
$\frac{n_{1.}n_{.4}}{n} = \frac{140\cdot 110}{360} = 42,8$, $\frac{n_{2.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27,5$, $\frac{n_{2.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30,6$,
$\frac{n_{2.}n_{.3}}{n} = \frac{110\cdot 60}{360} = 18,3$, $\frac{n_{2.}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33,6$, $\frac{n_{3.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27,5$,
$\frac{n_{3.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30,6$, $\frac{n_{3.}n_{.3}}{n}\frac{110\cdot 60}{360} = 18,3$, $\frac{n_{3.}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33,6$
$K = \frac{(50-35)^2}{35} + \frac{(30-38,9)^2}{38,9} + \ldots + \frac{(50-33,6)^2}{33,6} = 76,84$,

$r = 3, s = 4, \chi^2_{0,95}(6) = 12,6$. Protože $K \geq 12,6$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05.

Cramérův koeficient:

$$V = \sqrt{\frac{76,4}{360\cdot 2}} = 0,3267.$$

2.1. Čtyřpolní tabulky. Speciálním případem kontingenčních tabulek, kdy $r = s = 2$ jsou čtyřpolní tabulky. Zavádí se pro ně jiné značení.

Definice 2.6. Nechť $r = s = 2$. Pak hovoříme o **čtyřpolní kontingenční tabulce** a používáme označení: $n_{11} = a, n_{12} = b, n_{21} = c, n_{22} = d$.

x	y		$n_{j.}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	$a+b$
$x_{[2]}$	c	d	$c+d$
$n_{.k}$	$a+c$	$b+d$	n

Poznámka 2.7. Pro tuto tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test. (Je popsán např. v knize [**11**].)

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá **podíl šancí (odds ratio)**. Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	$a+b$
neúspěch	c	d	$c+d$
$n_{.k}$	$a+c$	$b+d$	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za prvních okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$.

DEFINICE 2.8. **Podíl šancí (odds ratio)** ve čtyřpolní tabulce je definován jako $OR = \frac{ad}{bc}$.

VĚTA 2.9. *Pomocí* $100(1-\alpha)\%$ *asymptotického intervalu spolehlivosti pro podíl šancí lze na asymptotické hladině významnosti* α *testovat hypotézu o nezávislosti nominálních veličin* X *a* Y*. Asymptotický* $100(1-\alpha)\%$ *interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má meze:*

$$\ln OR \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}.$$

Jestliže po odlogaritmování nezahrne interval spolehlivosti 1, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α*.*

Příklad 2.10. U 135 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Řešení.
$OR = \frac{ad}{bc} = \frac{17\cdot 58}{11\cdot 39} = 2,298,\ \ln OR = 0,832,$

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0,439,\ u_{0,975} = 1,96$$

$\ln dm = 0,832 - 0,439\cdot 1,96 = -0,028,\ \ln hm = 0,832 + 0,439\cdot 1,96 = 1,692 \ \Rightarrow\ dm = e^{-0,28} = 0,972,\ hm = e^{1,692} = 5,433$
Protože interval (0,972; 5,433) obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

3. Testování nezávislosti ordinálních veličin

V dalším se budeme věnovat vzájemnému vztahu náhodných veličin ordinálního typu. Ordinální (pořadová) náhodná veličina je taková, u jejíž dvou hodnot můžeme navíc určit pořadí (úroveň spokojenosti, vzdělání), tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání. Jako hodnoty lze použít text, datum, číslo.

NÁVOD 3.1 (Popis testu). Nechť X, Y jsou dvě ordinální náhodné veličiny. Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ z rozložení, jímž se řídí náhodný vektor (X, Y). Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny $Y_i, i = 1, \ldots, n$. Testujeme hypotézu H_0 : X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1 : X, Y jsou pořadově závislé náhodné veličiny (resp. proti levostranné alternativě H_1: mezi X a Y existuje nepřímá pořadová závislost resp. proti pravostranné alternativě H_1: mezi X a Y existuje přímá pořadová závislost).

Testová statistika se nazývá Spearmanův koeficient pořadové korelace a má tvar:

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2.$$

H_0 zamítáme na hladině významnosti α

(1) ve prospěch oboustranné alternativy, když $|r_S| \geq r_{S,1-\alpha}(n)$
(2) ve prospěch levostranné alternativy, když $r_S \leq -r_{S,1-2\alpha}(n)$
(3) ve prospěch pravostranné alternativy, když $r_S \geq r_{S,1-2\alpha}(n)$

$r_{S,1-\alpha}(n)$ je kritická hodnota, kterou pro $\alpha = 0,05$ nebo 0,01 a $n \leq 30$ najdeme v tabulkách. Pro $n > 30$ H_0 zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když

$$|r_s| \geq \frac{u_{1-\alpha/2}}{\sqrt{n-1}}$$

(analogicky pro jednostranné alternativy).

Poznámka 3.2. Spearmanův koeficient r_S současně měří sílu pořadové závislosti náhodných veličin X, Y. Nabývá hodnot z intervalu $\langle -1, 1 \rangle$. Čím je jeho hodnota bližší -1 (resp. 1), tím je silnější nepřímá (resp. přímá) pořadová závislost veličin X, Y. Čím je jeho hodnota bližší 0, tím je slabší pořadová závislost veličin X, Y.

Příklad 3.3. Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient r_S a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení.

$$\begin{aligned} r_s &= 1 - \frac{6}{7(7^2-1)} \left[(4-4)^2 + (1-2)^2 + (6-5)^2 \right. \\ &\quad \left. + (3-1)^2 + (2-3)^2 + (7-7)^2 \right] \\ &= 0,857 \end{aligned}$$

Kritická hodnota: $r_{S,0,95}(7) = 0,745$. Protože $0,857 \geq 0,745$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

4. Testování nezávislosti intervalových či poměrových veličin

V poslední části této kapitoly se budeme zabývat nezávislostí náhodných veličin intervalového nebo poměrového typu. Připomeňme, že intervalová (rozdílová) proměnná je taková,

pro jejíž dvě hodnoty můžeme navíc (k možnostem ordinální proměnné) vypočítat, o kolik je jedna hodnota větší (resp. menší) než druhá (měsíční příjem domácnosti, počet dětí v rodině). Hodnotami jsou tedy čísla. Poměrová (podílová) proměnná je ta, pro jejíž dvě hodnoty můžeme navíc (k možnostem intervalové proměnné) vypočítat, kolikrát je jedna hodnota větší (resp. menší) než druhá, tzn. jedná se pouze o kladné hodnoty (počet členů domácnosti).

4.1. Pearsonův koeficient korelace. V teorii pravděpodobnosti byl zaveden Pearsonův koeficient korelace náhodných veličin X, Y (které jsou aspoň intervalového charakteru) vztahem

$$R(X,Y) = \begin{cases} \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} & \text{pro } \sqrt{D(X)},\ \sqrt{D(Y)} > 0, \\ 0 & \text{jinak.} \end{cases}$$

Připomeneme jeho vlastnosti:

(1) $R(X,X) = 1$
(2) $R(X,Y) = R(Y,X)$
(3) $R(a+bX, c+dY) = \operatorname{sgn}(bd) R(X,Y)$
(4) $-1 \le R(X,Y) \le 1$ a rovnosti je dosaženo tehdy a jen tehdy, když existují reálné konstanty a, b, kde $b \neq 0$ tak, že $P(Y = a + bX) = 1$, přičemž $R(X,Y) = 1$ pro $b > 0$ a $R(X,Y) = -1$ pro $b < 0$.

Z těchto vlastností plyne, že $R(X,Y)$ je vhodnou mírou těsnosti lineárního vztahu náhodných veličin X, Y.

Definice 4.1. $R(X,Y)$ většinou nemůžeme počítat přímo, protože to vyžaduje znalost simultánního rozložení náhodného vektoru (X,Y). V praxi jsme většinou odkázáni na náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$. Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

(1) výběrové průměry

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i,\ M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

(2) výběrové rozptyly

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2,\ S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2,$$

(3) výběrovou kovarianci

$$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$$

S jejich pomocí zavedeme **výběrový koeficient korelace**

$$R_{12} = \frac{S_{12}}{S_1 S_2} \qquad \text{pro } S_1 S_2 > 0.$$

Poznámka 4.2. Vlastnosti koeficientu korelace uvedené v 4.1 se přenášejí i na výběrový koeficient korelace.

4.2. Koeficient korelace dvourozměrného normálního rozdělení. Připomeňme si základní vlastnosti dvourozměrného normálního rozdělení. Ty jsou popsány v následujících tvrzeních.

VĚTA 4.3. *Nechť náhodný vektor (X, Y) má dvourozměrné normální rozložení s hustotou*

$$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + (\frac{y-\mu_2}{\sigma_2})^2\right]},$$

přičemž $\mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = D(X), \sigma_2^2 = D(Y), \rho = R(X,Y)$.
Pak marginální hustoty jsou:

$$\varphi_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \varphi_2(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{\frac{(y-\mu_2)^2}{2\sigma_2^2}}.$$

VĚTA 4.4. *Je-li $\rho = 0$, pak pro $\forall (x,y) \in \mathbb{R}^2 : \varphi(x,y) = \varphi_1(x)\varphi_2(y)$, tedy náhodné veličiny X, Y jsou stochasticky nezávislé. Jinými slovy: stochastická nezávislost složek X, Y normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti.*

DEFINICE 4.5. Je-li $\rho \neq 0$, jsou náhodné veličiny X, Y stochasticky závislé. Je-li $\rho > 0$, říkáme, že jsou **kladně korelované**, je-li $\rho < 0$, říkáme, že jsou **záporně korelované**.

V dalším textu budeme předpokládat, že náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ pochází z **dvourozměrného normálního rozdělení** s parametry $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$. Využitím tohoto předpokladu dostáváme návod, jak testovat nezávislost dvou náhodných veličin pomocí výběrového koeficientu korelace.

VĚTA 4.6. *Testujeme $H_0 : \rho = 0$ proti oboustranné alternativě $H_1 : \rho \neq 0$ (resp. proti levostranné alternativě $H_1 : \rho < 0$ resp. proti pravostranné alternativě $H_1 : \rho > 0$). Testová statistika má tvar:*

$$T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}.$$

Platí-li nulová hypotéza, pak $T \sim t(n-2)$.
Kritický obor pro test H_0 proti oboustranné alternativě:

$$W = (-\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty),$$

proti levostranné alternativě:

$$W = (-\infty, -t_{1-\alpha}(n-2)\rangle$$

a proti pravostranné alternativě:

$$W = \langle t_{1-\alpha}(n-2), \infty).$$

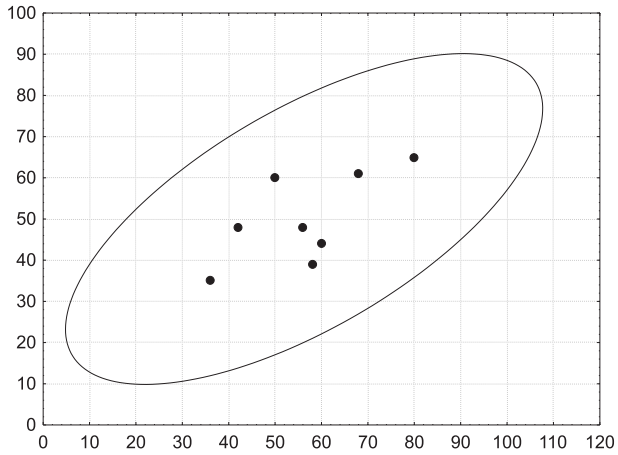
H_0 zamítáme na hladině významnosti α, když $T \in W$.

Příklad 4.7. Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

Řešení. Nejprve se musíme přesvědčit, že uvedené výsledky lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení. Lze tak učinit orientačně pomocí dvourozměrného tečkového diagramu. Tečky by měly vytvořit elipsovitý obrazec.



Obrázek 1: Dvourozměrný tečkový diagram

Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti.

Testujeme $H_0 : \rho = 0$ proti pravostranné alternativě $H_1 : \rho > 0$.

Výpočtem zjistíme: $R_{12} = 0,6668, T = 2,1917$. V tabulkách najdeme $t_{0,95}(6) = 1,9432$. Kritický obor: $W = \langle 1,9432; \infty)$. Protože $T \in W$, hypotézu o neexistenci kladné korelace výsledků z 1. a 2. testu zamítáme na hladině významnosti 0,05.

4.3. Porovnání koeficientu korelace s danou konstantou. Nyní nás bude zajímat, jak testovat hypotézu, že se korelační koeficient rovná libovolné konstantě. Tento test se provádí např. tehdy, když experimentátor porovnává vlastnosti svých dat s vlastnostmi uváděnými v literatuře.

Věta 4.8. *Nechť c je reálná konstanta. Testujeme $H_0 : \rho = c$ proti $H_1 : \rho \neq c$. Test je založen na statistice*

$$U = \left(Z - \frac{1}{2} \ln \frac{1+c}{1-c} - \frac{c}{2(n-1)} \right) \sqrt{n-3},$$

která má za platnosti H_0 pro $n \geq 10$ asymptoticky rozložení $N(0,1)$, přičemž

$$Z = \frac{1}{2} \ln \frac{1+R_{12}}{1-R_{12}}$$

je tzv. Fisherova Z-transformace. Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Příklad 4.9. U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu $H_0 : \rho = 0,9$ proti $H_1 : \rho \neq 0,9$.

Řešení.

$Z = \frac{1}{2}\ln\frac{1+0{,}85}{1-0{,}85} = 1{,}2562,$

$U = \left(1{,}2562 - \frac{1}{2}\ln\frac{1+0{,}9}{1-0{,}9} - \frac{0{,}9}{2(600-1)}\right)\sqrt{600-3} = -5{,}2976,$

$u_{0{,}975} = 1{,}96,\ W = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$.

Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

4.4. Porovnání dvou koeficientů korelace. Dále uvedeme test pro situaci, kdy máme k dispozici dva nezávislé náhodné výběry z dvourozměrných normálních rozdělení a chceme zjistit, zda se jejich korelační koeficienty statisticky liší.

VĚTA 4.10. *Nechť jsou dány dva nezávislé náhodné výběry o rozsazích n a n^* z dvourozměrných normálních rozložení s korelačními koeficienty ρ a ρ^*. Testujeme $H_0 : \rho = \rho^*$ proti $H_1 : \rho \neq \rho^*$. Označme R_{12} výběrový koeficient korelace 1. výběru a R_{12}^* výběrový koeficient korelace 2. výběru. Položme*

$$Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} \quad \text{a} \quad Z^* = \frac{1}{2}\ln\frac{1+R_{12}^*}{1-R_{12}^*}.$$

Platí-li H_0, pak testová statistika

$$U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}$$

má asymptoticky rozložení $N(0,1)$. Kritický obor pro test H_0 proti oboustranné alternativě tedy je

$$W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty).$$

H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Příklad 4.11. Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový koeficient korelace mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že se koeficienty korelace v obou skupinách neliší.

Řešení.

$Z = \frac{1}{2}\ln\frac{1+0{,}65}{1-0{,}65} = 0{,}7753,$

$Z^* = \frac{1}{2}\ln\frac{1+0{,}37}{1-0{,}37} = 0{,}3884,$

$U = \frac{0{,}7753-0{,}3884}{\sqrt{\frac{1}{100-3}+\frac{1}{142-3}}} = 2{,}9242,$

$u_{0{,}975} = 1{,}96, W = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$.

Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

4.5. Interval spolehlivosti pro koeficient korelace. V praxi bývá velice užitečný také interval spolehlivosti pro koeficient korelace, který nám poskytuje názornou představu o závislosti dvou normálně rozdělených náhodných veličin.

VĚTA 4.12. *Jestliže dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení, jehož koeficient korelace se příliš neliší od nuly ($|\rho| < 0,5$) a rozsah výběru je dostatečně velký ($n \geq 100$), lze odvodit, že $100(1-\alpha)\%$ interval spolehlivosti pro ρ má meze*

$$R_{12} \pm u_{1-\alpha/2} \frac{1 - R_{12}^2}{\sqrt{n-3}}.$$

Nejsou-li uvedené podmínky splněny, pak nelze tento vzorec použít, protože rozložení výběrového korelačního koeficientu je příliš zešikmené. V takovém případě využijeme následujícího tvrzení.

VĚTA 4.13. *Náhodná veličina*

$$Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}}$$

má i při malém rozsahu výběru přibližně normální rozložení se střední hodnotou

$$E(Z) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

(2. sčítanec lze při větším n zanedbat) a rozptylem $D(Z) = \frac{1}{n-3}$.
Standardizací veličiny Z dostaneme veličinu

$$U = \frac{Z - E(Z)}{\sqrt{D(Z)}},$$

která má asymptoticky rozložení $N(0,1)$.
Tudíž $100(1-\alpha)\%$ asymptotický interval spolehlivosti pro $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ bude mít meze $Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$.
Interval spolehlivosti pro ρ pak dostaneme zpětnou transformací.

Poznámka 4.14. Jelikož $Z = \operatorname{arctgh} R_{12}$, dostáváme $R_{12} = \operatorname{tgh} Z$ a meze intervalu spolehlivosti pro ρ můžeme psát ve tvaru

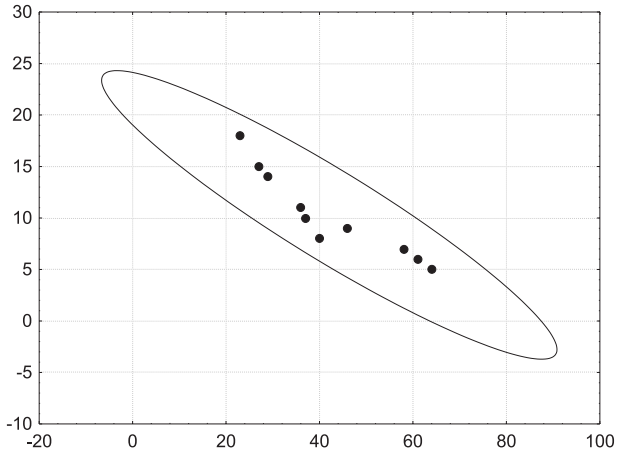
$$\operatorname{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right), \text{ přičemž } \operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Příklad 4.15. Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtěte výběrový koeficient korelace a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný koeficient korelace ρ.

Řešení. Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu, viz. Obr. 2. Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.



Obrázek 2: Dvourozměrný tečkový diagram

Testujeme $H_0 : \rho = 0$ proti $H_1 : \rho \neq 0$. Vypočítáme $R_{12} = -0,9325$, tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost. Testová statistika: $T = -7,3053$, kvantil $t_{0,975}(8) = 2,306$, kritický obor $W = (-\infty, -2,306\rangle \cup \langle 2,306, \infty)$. Jelikož $T \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y. Vypočítáme

$$Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}} = \frac{1}{2} \ln \frac{1 - 0,9325}{1 + 0,9325} = -1,6772.$$

Meze 95% asymptotického intervalu spolehlivosti pro ρ jsou tgh $(-1,6772 \pm \frac{1,96}{\sqrt{7}})$, tedy $-0,9842 < \rho < -0,7336$ s pravděpodobností přibližně 0,95.

Úlohy k procvičení

Cvičení 4.1 (Testování nezávislosti nominálních veličin)**.** Na hladině významnosti 0,05 testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví a vypočtěte Cramérův koeficient, jsou-li k dispozici následující údaje:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

[hypotézu o nezávislosti pohlaví a pedagogické hodnosti nezamítáme, Cramérův koeficient: 0,187]

Cvičení 4.2 (Testování nezávislosti ordinálních veličin)**.** 12 různých softwarových firem nabízí programy pro vedení účetnictví. Programy byly posouzeny odbornou komisí a komisí složenou z profesionálních účetních. Výsledky v 1. a 2. komisi: (6,4), (7,5), (1,2), (8,10), (4,6), (2.5,1), (9,7), (12,11), (10,8), (2.5,3), (5,12), (11,9). Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu o nezávislosti pořadí v obou komisích.

[$r_s = 0,715$, nulovou hypotézu zamítáme]

Cvičení 4.3 (Testování nezávislosti intervalových a poměrových veličin)**.** V dílně pracuje 15 dělníků, u nichž byl zjištěn počet směn odpracovaných za měsíc (veličina X) a počet zhotovených výrobků (veličina Y). Orientačně ověřte dvourozměrnou normalitu dat, vypočtěte výběrový koeficient korelace mezi X a Y, sestrojte pro něj 99% asymptotický interval spolehlivosti a na hladině 0,01 testujte hypotézu o nezávislosti X a Y.

x	20	21	18	17	20	18	19	21	20	14	16	19	21	15	15
y	92	93	83	80	91	85	82	98	90	60	73	86	96	64	81

[$r_{12} = \frac{s_{12}}{s_1 s_2} = 0,927$, hypotézu o nezávislosti veličin X a Y zamítáme, IS pro ρ : $(0,7131; 0,983)$]

Cvičení 4.4. Nechť $(X_1, Y_1),\ldots,(X_{16}, Y_{16})$ je náhodný výběr z dvourozměrného normálního rozložení. Výběrový koeficient korelace R_{XY} nabyl hodnoty $-0,87$. Jestliže provedeme transformaci $U_i = 1 + 3X_i$, $V_i = -3 - Y_i$, $i = 1,\ldots,16$, jakou hodnotu nabude výběrový koeficient korelace R_{UV}?

[$R_{UV} = 0,87$]

Cvičení 4.5. 400 náhodně vybraných pracovníků potravinářského podniku bylo dotázáno na příčiny nespokojenosti na pracovišti. Výsledky jsou uvedeny v tabulce:

kategorie pracovníků	hlavní příčina nespokojenosti				
	pracovní prostředí	špatné vztahy	organizace práce	výdělek	jiné
dělníci	80	50	75	40	55
THP	10	10	25	30	25

Na hladině významnosti 0,05 testujte hypotézu, že hlavní příčina nespokojenosti nezávisí na kategorii, do níž je pracovník zařazen. Vypočtěte Cramérův koeficient.

[hypotézu o nezávislosti zamítáme, Cramérův koeficient je $V = 0,25$]

Rejstřík