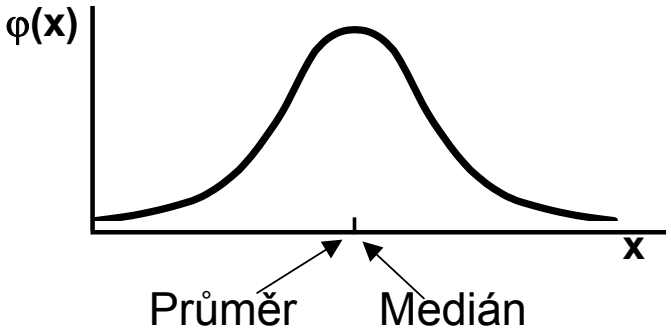
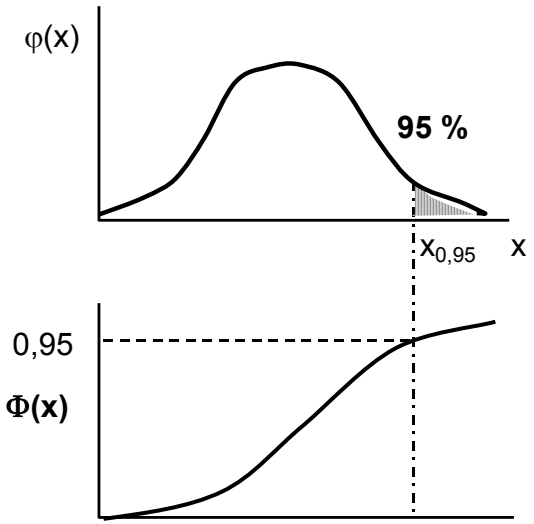
Zobecněné lineární modely
(GLM, *generalized linear models*)

MU
IBA

Parametry rozložení

- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - Středu (medián, průměr, geometrický průměr)
  - Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
  - Tvaru rozložení (skewness, kurtosis)
  - Kvantily rozložení – kolik % řady dat leží nad a pod kvantilem

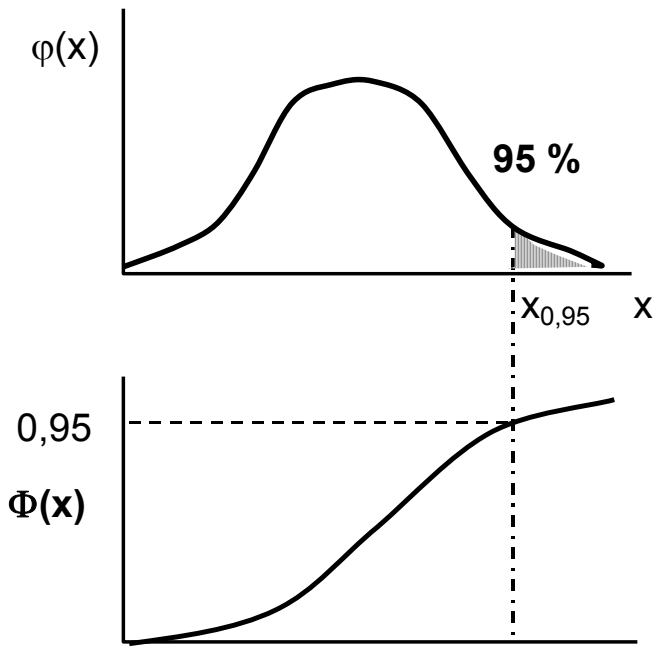




Distribuční funkce

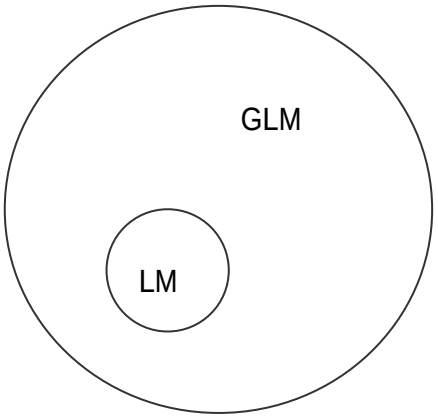
- Definice kvantilu dle distribuční funkce - Kvantil rozložení ($X_{0,95}$) je číslo, jehož hodnota distribuční funkce je rovna pravděpodobnosti, pro kterou je kvantil definován ($\Phi(x)$ … distribuční funkce), tj. pokud vezmeme nějaký bod rozložení a porovnáme jej s tímto bodem (kvantilem), máme 95% pravděpodobnost, že bude menší než hodnota kvantilu ($X_{0,95}$).

- Pomocí distribuční funkce můžeme určit jaký podíl hodnot rozložení je menší než daná hodnota – využití při statistických testech



Zobecněné lineární modely (*GLM*)

- GLM (*Generalized linear models)* jsou rozšířením lineárních modelů (LM) s vetší tolerancí pro různé distribuční vlastnosti vysvětlovaných proměnných.



Přehled (nejznámějších) rozdělení z tzv. Exponenciální rodiny rozdělení

- Normální rozdělení (Normal distribution)
- Exponenciální rozdělení (Exponential distribution)
- Poissonovo rozdělení (Poisson distibution)
- Beta
- Gamma
- Alternativní rozdělení (Bernoulli d.)
- Binomické rozdělení (Binomial d.)
- Multinomické rozdělení (Multinomial d.)

- http://broiler.stat.vt.edu/~sundar/java/applets/

Lineární regresní model

- V klasickém **lineárním regresním modelu** (speciální případ zobecněného lineárního modelu) systematická část vyjadřuje lineární vztah pro střední hodnotu *E*(*Y*) a pro *xj* - prediktory neboli vysvětlující proměnné.

$$E(Y_i) = \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, i = 1,\ldots,n$$

kde ßj jsou neznámé parametry, počet nezávisle proměnných (prediktorů) je *p,* počet pozorování je *n.*

- Náhodná složka modelu je reprezentována náhodnými chybami *εi* . Rozdělení těchto náhodných veličin *εi* je **normální** (a rozptyl není funkcí střední hodnoty).
- V reálném světě má mnoho procesů jiný než lineární vztah a také variance stochastické složky může být závislá na očekávané (střední) hodnotě *E*(*Y*).

Zobecněný lineární model

- **Zobecněné lineární modely** (GLM) rozšiřují klasické lineární regresní modely ve dvou směrech:
- Předpokládané rozdělení Y pro danou nezávisle proměnnou *xj* nemusí být normální, ale může rovněž pocházet z třídy ***exponenciálních*** rozdělení, které zahrnují důležitá rozdělení jako je ***binomické, Poissonovo, exponenciální*** nebo ***gamma.***
- Normalita chyb- často nesplňený předpoklad KLM
- Závislost hodnot vysvětlované proměnné na hodnotách prediktorů je dána linkovací funkcí *g*(*μi*) (link function):

$$g(E(Y_i)) = \eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij},$$

- kde g(*μi*) je nelineární linkovací funkce, která spojuje neznámé střední hodnoty výchozího rozdělení náhodné veličiny *Y* s hodnotami nezávisle proměnných.(Na funkci g(*μi*) je kladen požadavek, aby byla monotónní a diferencovatelná)

Link function

- Volba linkovací funkce a předpokládaný typ distribuce se ovšem nemůže kombinovat náhodně. Linkovací funkce, která přísluší danému rozložení se nazývá kanonická linkovací funkce (canonical link function).
- U lineárních modelů, ve ktrerých náhodná komponenta závisí na předpokladu, že vysvětlující proměnná má normální rozdělení je kanonická linkovací funkce identita.

Linkovací funkce

- Linkovací funkce, která přísluší danému rozložení se nazývá kanonická linkovací funkce (*canonical link function*).

Typ proměnné	**"Typická" link funkce**	**Referenční distribuce**
počty (frekvence)	log	Poissonova
pravděpodobnost (relativní frekvence)	logit nebo probit	binomická
rozměry, poměry	inverze nebo log	gamma
vzácné typy měření	identita	Gaussova ("normální")

Výstavba regresního modelu:

Vhodně zvolený regresní model by měl splňovat tyto kritéria:

- obsahovat co nejmenší počet parametrů: více parametrů sice zvýší přesnost modelu, avšak při aplikace na nové data se stává nevhodným
- Parametry by měly být navzájem nezávislé: použité parametry by se neměly dát vyjádřit kombinací ostatních parametrů, což je ovšem častým problémem nelineárních modelů
- funkce by měla správně popisovat vysvětlovanou (závislou) proměnnou v extrémech závislosti i v jejím asymptotickém chování: při nízkých nebo vysokých hodnotách nezávislé proměnné některé modely poskytují nesprávné limitní hodnoty závislé proměnné

Multikolinearita

- **Multikolinearita** - Existují-li závislosti mezi jednotlivými nezávisle proměnnými modelu. Koeficienty determinace lineárních modelů (jedné nezávisle proměnné na ostatních nezávisle proměnných) jsou vysoké (větší než 0,5). Nezávisle proměnné jsou navzájem korelované.
- Odhad regresních parametrů – velký rozptyl.
- I významné nezávisle proměnné se jeví jako nevýznamné, popř. parametry mohou mít opačné znaménko…
- Obtížná interpretace parametrů beta. (Obvykle: Koeficient beta1 lze interpretovat jako střední změnu Y při jednotkové změně X1 a nezměněné hodnotě X2. Nyní však X1 a X2 vzájemně korelované, proto nelze předpokládat, že při změně X1 zůstane X2 nezměněna.)
- Příklad 1: obvod pasu a váha významně korelované
- Příklad 2: Výška platu a daně úzce korelované
- **Řešení:** méně proměnných v modelu, vyloučení korelovaných nezávislých proměnných.

Výstavba modelu

	LM	GLM
Odhad parametrů modelu	Metoda nejmenších čtverců	Metoda maximální věrohodnosti
Signifikantní prediktory	T-test F-test	Test poměrem věrohodností Waldův test
Hodnocení vhodnosti modelu	Koeficient determinace	deviace AIC

Věrohodnost (*Likelihood*)

- **Věrohodnostní funkce *(likelihood function)*** – je funkcí parametrů modelu. Je to pravděpodobnost, s jakou lze získat naměřená data, v případě, že parametry modelu jsou dané.

- **Metoda maximální věrohodnosti (*maximal likelihood method*)** - vede k takovým hodnotám odhadu parametrů, které maximalizují pravděpodobnost získání naší pozorované množiny dat.

- Jsou hledány takové parametry, pro které je věrohodnostní funkce maximální.

- Pro snadnější výpočet se věrohodnostní funkce logaritmuje – logaritmická věrohodnostní funkce (*log-likelihood function*)

Věrohodnostní funkce

- Pro danou distribuci určenou *f*(*yi* ; ***ß,F***) a pozorování ***y*** = (*y*1, *y*2, . . ., *yn*)', je věrohodnostní funkce (log-likelihood function) pro ***ß*** a *F*, vyjádřená jako funkce střední hodnoty *μ* = (*μ1,…, μn*) závisle proměnné {*Y*1, *Y*2, . . . , *Yn*} a má tvar:

$$l(\boldsymbol{\mu};\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i;\boldsymbol{\beta},\phi)$$

- U klasických lineárních regresních modelů se jako kritérium pro odhad neznámých ***ß*** = (*ß*0, *ß*1,..., *ßp*)' parametrů používá metoda nejmenších čtverců. Pokud jsou pozorování nezávislé a normálně rozložené s konstantním variancí *s2*, je odhad parametru ***ß*** a s2 pomocí metody nejmenších čtverců ekvivalentní k odhadu maximální věrohodnosti.

Maximální věrohodnost

- Odhad maximální věrohodnosti parametru *ß* můžeme dosáhnout např. pomocí iterativního algoritmu re-weighted least squares (IRLS) (podrobný popis McCullagh and Nelder, 1989) nebo Newton-Raphsonova metoda (podrobný popis Harrell,2001).

Významnost proměnných

- Máme –li odhadnuté regresní parametry, musíme určit statistickou významnost všech nebo jednotlivých vysvětlujích proměnných zahrnutých do modelu. To znamená, že zkoumáme zda daná proměnná (nebo skupina *k*-danných proměnných) po přidání do modelu přinese statisticky významné zpřesnění předpovězených hodnot *Y*.
- Položíme tedy vektor regresních koeficientů *k* testovaných kovariát rovný nule.
- tedy:

$$H_0 : \beta_j = 0, j = 1....m$$

- V našem případě jsou kovariáty vysvětlující proměnné (prediktory), které zahrnujeme do modelu a u nichž zatím neznáme jejich příspěvek do modelu.
- K otestování významnosti regresních koeficientů se používá např. test poměrem věrohodnosti (likelihood ratio test), Waldův test...

Test poměrem věrohodnosti

- používáme na porovnání globální hypotézy, že žádné efekty nejsou statisticky významné proti plnému modelu odhadnutému MMV.
- Používá statistiku *LR,*.

$$LR = -2\ln\left(\frac{l\ za\ hypotézy\ H_0}{l\ s\ MMV\ odhadem}\right)$$

- Položíme tedy proti sobě alternativu modelu bez testovaných kovariát proti úplnému modelu (se všemi kovariátami) s regresními parametry odhadnutými metodou maximální věrohodnosti.

- Při velkém *n* má statistika *LR* přibližně χ2 rozdělení *(chí-kvadrát*) se stupni volnosti rovnými počtu odhadovaných parametrů. LR test se využívá při testování hypotézy, že všechny regresní koeficienty jsou rovny nule , ***ß*** = (*ß*0, *ß*1,..., *ßp*) = 0.

- Nulovou hypotézu zamítáme na hladině významnosti α jestliže *LR* > χ21-α (*p*), kde *p* je počet odhadovaných parametrů.

Waldův test

- Waldův test se používá k otestování statistické významnosti daného prediktoru nebo skupiny prediktorů
- Waldova statistika *W*, která je zevšeobecněním *t*- nebo *z*-statistiky je funkce rozdílu MMV odhadu a hypotetické hodnoty regresního parametru testované kovariáty, normalizovaného odhadem standardní odchylky MMV odhadu.
- Tato statistika má při dostatečně velkém *n* přibližně χ^2 rozdělení *(chí-kvadrát*) se stupni volnosti rovnými počtu odhadovaných parametrů.
- Nulovou hypotézu zamítáme na hladině významnosti α jestliže $W > \chi^2 1-\alpha\ (p)$, kde *p* je počet odhadovaných parametrů.

Ověřování vhodnosti modelu

- Podobně jako reziduální součet čtverců v linárních regresních modelech se i v ZLM testuje hypotéza o vhodnosti modelu (goodness-of-fit). Určení vhodné modelové rovnice je základem všech regresních modelů.

- Jedním z důležitých principů regresních modelů je zásada jednoduchosti- jednodušší model, který dobře popisuje naše data je vhodnější než složitější model popisující data téměř dokonale

škálová deviace D (scaled deviation)

K otestování vhodnosti modelu slouží škálová deviace D:

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2[l(\mathbf{y};\mathbf{y}) - l(\hat{\boldsymbol{\mu}};\mathbf{y})]$$

Kde $l(\mathbf{y};\mathbf{y})$ je maximálně věrohodný odhad, ve kterém jsou fitované hodnoty rovny pozorovaným hodnotám a $l(\hat{\boldsymbol{\mu}};\mathbf{y})$ je věrohodnostní funkce odhadnutých parametrů ***ß***.

Deviace je velmi užitečná při srovnání dvou modelů z nichž jeden je podmodelem (submodelen) druhého.

MMV odpovídá hledání minima deviace modelu.
Je-li $D > \chi^2_{1-\alpha}$ (n-m), kde m (n) je počet odhadovaných parametrů submodelu (celkového modelu), pak je model nevhodný.

Akaikovo informační kritérium
(Akaike information criterion, AIC)

- AIC=-2(maximum logaritmované věrohodnosti – počet parametrů modelu)

- Čím je hodnota AIC menší, tím je model lepší.
- AIC penalizuje modely s velkým počtem parametrů

Analýza reziduí

- Analýza reziduí je důležitou součástí ověřování vhodnosti modelu. Můžeme tak zjistit, zda výchozí předpoklad o rozdělení náhodných chyb či tvaru linkovací funkce byl správný.

- Pomocí reziduí zjistíme body, jejichž reziduum je velmi odlišné od ostatních pozorování. Pokud se v grafu objeví závislost reziduí na prediktorech nebo variabilita reziduí roste v závislosti na veličinách modelu, musíme celý model znovu přehodnotit, popř. jej vytvořit od začátku.

- Typy reziduí: Pearsonova (standartizovaná) rezidua, Ascombova rezidua, deviační rezidua, rezidua stabilizující rozptyl etc

Nejznámější typy reziduí v GLM

- **Pearsonova** (standardizovaná) **rezidua** (linear)
  -nevýhodou je, že pro nenormální rozdělení jsou zešikmená

- **Standardizovaná transformovaná rezidua** (transformed linear)

  -rezidua se transformují aby se jejich rozložení blížilo normálnímu

  1) ***Anscombova rezidua*** – snaha je, aby transformovaná rezidua měla nulovou šikmost

  2) ***Rezidua stabilizující rozptyl*** - cílem je, aby u transformovaných reziduí nebyl rozptyl funkcí střední hodnoty, ale konstantní

Literatura

- Harrel F. E., Jr. (2001): Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis. Springer, Springer Series in Statistics, New York.
- McCullagh, P., Nelder, J.A. (1989): Generalized Linear Models (2nd edition), Chapman & Hall
- McCullagh C. E., Searle S. R. (2001): Generalized, Linear, and Mixed Models, John Wiley & Sons.
- Lemeshow, Stanley & Hosmer, David W., Jr.Logistic regression, p. 1-11. *In* Encyclopaedia of Biostatistics, 1st ed. [Online.] Wiley, London. http://www.wiley.co.uk./eob/sample4.pdf. [13 January 2004, last date accessed]