Osnova přednášky „Binární logistická regrese“

1. Motivace
2. Odvození modelu
3. Kódování proměnných
   - 3.1. Případ dvou kategorií
   - 3.2. Případ aspoň tří kategorií
4. Význam parametrů
5. Odhady parametrů
6. Intervaly spolehlivosti
   - 6.1. Interval spolehlivosti pro regresní parametr
   - 6.2. Interval spolehlivosti pro podíl šancí
7. Dílčí testy významnosti regresních parametrů
   - 7.1. Waldův test
   - 7.2. Test poměrem věrohodnosti
   - 7.3. Skórový test
8. Test významnosti modelu jako celku

9. Výstavba modelu
- **9.1. Principy výstavby modelu**
- **9.2. Výběr podmnožiny vysvětlujících proměnných**

10. Hodnocení modelu z různých hledisek
- **10.1. Testy dobré shody**
- **10.2. Koeficienty determinace**
- **10.3. Informační kritéria**
- **10.4. Klasifikační tabulka**
- **10.5. ROC křivka**

11. Příklad

Binární logistická regrese

1. Motivace

Tato metoda umožňuje odhad pravděpodobnosti nastoupení nějakého jevu (zapíšeme ho pomocí náhodné veličiny Y jako $\{Y = 1\}$) pomocí k známých regresorů $X_1, \ldots, X_k$, které mohou být jak spojitého, tak kategoriálního typu. Byla vytvořena v 60. letech 20. století.

Použití v praxi:

v medicíně, veličina Y popisuje přítomnost či nepřítomnost nějaké choroby;

v bankovnictví, veličina Y popisuje splácení či nesplácení úvěru;

v marketingových kampaních, veličina Y popisuje odezvu na reklamu nějakého výrobku;

v pojišťovnictví, veličina Y popisuje uplatnění či neuplatnění pojistného nároku

…

2. Odvození modelu

Uvažme závisle proměnnou náhodnou veličinu Y, která nabývá hodnoty 1 s pravděpodobností ϑ a hodnoty 0 s pravděpodobností $1-\vartheta$, tj. $Y \sim A(\vartheta)$ a její pravděpodobnostní funkce má tvar:

$\pi(y) = \begin{cases} \vartheta^y (1-\vartheta)^{1-\vartheta} \\ 0 \text{ jinak} \end{cases}$ pro $y = 0,1$. Jev {Y = 1} často interpretujeme jako úspěch, jev {Y = 0} jako neúspěch.

Předpokládejme, že máme k vysvětlujících proměnných $X_1, \ldots, X_k$.

Označme $\mathbf{X} = (1, X_1, \ldots, X_k)^T$ vektor vysvětlujících proměnných s absolutním členem a

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ vektor regresních koeficientů.

Pro predikci hodnot veličiny Y nelze použít lineární regresní model tvaru

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mathbf{X}^T \boldsymbol{\beta},$$

protože na levé straně jsou jen 0 a 1, zatímco pravá strana může nabývat jakoukoli reálnou hodnotu.

Budeme tedy modelovat nikoliv hodnoty veličiny Y, ale pravděpodobnost úspěchu ϑ (tj. střední hodnotu veličiny Y) (za předpokladu, že známe hodnoty vektoru vysvětlujících proměnných).

Pokud bychom využili model $P(Y=1/\mathbf{X}=\mathbf{x})=\mathbf{X}^T\boldsymbol{\beta}$,

mohlo by se stát, že některé predikované pravděpodobnosti úspěchu (při daném **x**) by ležely vně intervalu (0,1).

Tento problém lze částečně řešit zavedením šance: $\omega(\mathbf{x})=\frac{P(Y=1/\mathbf{X}=\mathbf{x})}{P(Y=0/\mathbf{X}=\mathbf{x})}=\frac{P(Y=1/\mathbf{X}=\mathbf{x})}{1-P(Y=1/\mathbf{X}=\mathbf{x})}$.

Šance vyjadřuje, kolikrát je při daném **x** vyšší pravděpodobnost úspěchu než neúspěchu. Nabývá hodnot z intervalu $(0,\infty)$.

Nyní je zapotřebí vzájemně jednoznačně transformovat interval $(0,\infty)$ na interval $(-\infty,\infty)$.

K tomuto účelu použijeme přirozený logaritmus šance:

$\ln\omega(\mathbf{x})=\ln\frac{P(Y=1/\mathbf{X}=\mathbf{x})}{1-P(Y=1/\mathbf{X}=\mathbf{x})}$ (jde o tzv. logitovou transformaci pravděpodobnosti úspěchu za předpokladu **X** = **x**, zkráceně logit).

Logaritmickou šanci již můžeme modelovat pomocí lineárního regresního modelu:

$$\ln \frac{P(Y=1/\mathbf{X}=\mathbf{x})}{1-P(Y=1/\mathbf{X}=\mathbf{x})} = \mathbf{X}^T\boldsymbol{\beta}.$$

Odtud můžeme vyjádřit šanci $\omega(\mathbf{x}) = e^{\mathbf{x}^T\boldsymbol{\beta}}$ a dále podmíněnou pravděpodobnost úspěchu:

$$P(Y=1/\mathbf{X}=\mathbf{x}) = \frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}} = \frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\beta}}}$$

resp. podmíněnou pravděpodobnost neúspěchu:

$$P(Y=0/\mathbf{X}=\mathbf{x}) = 1-P(Y=1/\mathbf{X}=\mathbf{x}) = 1-\frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}} = \frac{1}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}},$$

celkem

$$P(Y=y/\mathbf{X}=\mathbf{x}) = \left(\frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\beta}}}\right)^y \left(1-\frac{1}{1+e^{-\mathbf{x}^T\boldsymbol{\beta}}}\right)^{1-y} \quad \text{pro } y = 0, 1.$$

Tímto vztahem tedy modelujeme pravděpodobnost úspěchu či neúspěchu v závislosti na realizacích $x_1, \ldots, x_k$.

Upozornění: pravděpodobnost úspěchu, šance na úspěch a logit úspěchu jsou tři různé způsoby vyjádření téhož v tom smyslu, že jsou na sebe vzájemně převoditelné. Pro interpretaci jsou vhodnější pravděpodobnosti a šance než logity.

3. Kódování kategoriálních proměnných

3.1. Případ dvou kategorií

V tomto případě kategorie vysvětlující proměnné X kódujeme nejčastěji pomocí 0 a 1. Např. proměnná X udává pohlaví pacienta. Zvolíme X = 0 pro ženu a X = 1 pro muže.

3.2. Případ aspoň tří kategorií

Vysvětlující proměnná X má $r \geq 3$ kategorií, např. X udává úroveň vzdělání osoby a má tři kategorie: ZŠ, SŠ, VŠ.

3.2.1. Kódování přeparametrizovaného modelu

Zavedeme r závislých indikátorů $Z_1, \ldots, Z_r$ tak, že každý z nich vyjadřuje vždy jednu kategorii vysvětlující proměnné X hodnotou 1 a všechny ostatní hodnotou 0.
V našem případě zavedeme tři indikátory Z_1, Z_2, Z_3 takto:

$$Z_1 = \begin{cases} 1 \text{ pro ZŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 \text{ pro SŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_3 = \begin{cases} 1 \text{ pro VŠ} \\ 0 \text{ jinak} \end{cases}.$$

Vyjádřeno tabulkou:

Úroveň faktoru	indikátory		
	Z_1	Z_2	Z_3
ZŠ	1	0	0
SŠ	0	1	0
VŠ	0	0	1

Součet v každém sloupci tabulky je 1.
Každý indikátor je možno vyjádřit jako lineární kombinaci ostatních indikátorů.
Tato vlastnost je pro mnohé statistické postupy nežádoucí, proto budeme uvažovat o jeden indikátor méně. Vynechaná úroveň vysvětlující proměnné X bude sloužit jako referenční.
Referenční úroveň volíme tak, aby to bylo výhodné z interpretačního hlediska

3.2.2. Kódování typu dummy

Zavedeme r-1 nezávislých indikátorů Z_1, …, Z_{r-1}, které jsou definovány takto:
Z_1 = 1 pro 1. kategorii vysvětlující proměnné X, Z_1 = 0 jinak,
Z_2 = 1 pro 2. kategorii vysvětlující proměnné X, Z_2 = 0 jinak,
……..
Z_{r-1} = 1 pro (r-1). kategorii vysvětlující proměnné X, Z_{r-1} = 0 jinak.
Pro r-tou kategorii vysvětlující proměnné X nabývají všechny indikátory typu dummy Z_1, …, Z_{r-1} hodnoty 0 a tím indikují její výskyt.
V našem případě máme dva indikátory:

$$Z_1 = \begin{cases} 1 \text{ pro ZŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 \text{ pro SŠ} \\ 0 \text{ jinak} \end{cases}$$

. Vynechaná úroveň VŠ je referenční.

Vyjádřeno tabulkou:

Úroveň faktoru	indikátory	
	Z_1	Z_2
ZŠ	1	0
SŠ	0	1
VŠ	0	0

Součet v každém sloupci tabulky je 1. Při interpretaci výsledků analýz s indikátory typu dummy konfrontujeme jednotlivé kategorie vysvětlující proměnné X s referenční kategorií.

3.2.3. Kódování typu effect

Zavedeme r-1 nezávislých indikátorů Z_1, …, Z_{r-1}, které jsou definovány takto:
Z_1 = 1 pro 1. kategorii vysvětlující proměnné X, Z_1 = -1 pro r-tou kategorii proměnné X, Z_1 = 0 jinak,
Z_2 = 1 pro 2. kategorii vysvětlující proměnné X, Z_2 = -1 pro r-tou kategorii proměnné X, Z_2 = 0 jinak,
……
Z_{1r-1} = 1 pro (r-1). kategorii vysvětlující proměnné X, Z_{r-1} = -1 pro r-tou kategorii proměnné X, Z_{r-1} = 0 jinak,
Pro r-tou kategorii proměnné X nabývají všechny indikátory typu effect Z_1, …, Z_{r-1} hodnoty -1 a tím indikují její výskyt.
V našem případě máme dva indikátory:

$$Z_1 = \begin{cases} 1 \text{ pro ZŠ} \\ -1 \text{ pro VŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 \text{ pro SŠ} \\ -1 \text{ pro VŠ} \\ 0 \text{ jinak} \end{cases}$$

. Vynechaná úroveň VŠ je referenční.

Vyjádřeno tabulkou:

Úroveň faktoru	indikátory	
	Z_1	Z_2
ZŠ	1	0
SŠ	0	1
VŠ	-1	-1

Součet v každém sloupci tabulky je 0. Hovoříme o sigma omezené parametrizaci.

4. Význam parametrů

Ze vztahu pro logit

$$\ln \frac{P(Y=1/\mathbf{X}=\mathbf{x})}{1-P(Y=1/\mathbf{X}=\mathbf{x})} = \mathbf{X}^T\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$ plyne:

parametr β_0 udává velikost logitu pro nulové hodnoty všech vysvětlujících proměnných.
Je-li $\beta_0 = 0$, je logaritmus šance 0, tedy šance = 1 neboli pravděpodobnost úspěchu je 0,5.
Pro $\beta_0 > 0$ je šance na úspěch větší než 0,5 a pro $\beta_0 < 0$ je šance na úspěch menší než 0,5.

Pro interpretaci parametrů β_j zavedeme podíl šancí:

$$OR(x_j) = \frac{\omega(x_1,\ldots,x_j+1,\ldots,x_k)}{\omega(x_1,\ldots,x_j,\ldots,x_k)} = \ldots = e^{\beta_j}.$$

Jednotková změna j-té vysvětlující proměnné znamená v průměru e^{β_j} násobnou změnu šance na úspěch, zůstanou-li všechny ostatní vysvětlující proměnné stejné.
U kategoriálních proměnných závisí interpretace parametrů na způsobu, jakým kódujeme kategorie.
Parametry u indikátorových proměnných vyjadřují po odlogaritmování příslušné násobky šance na úspěch v referenční kategorii.

5. Odhady parametrů

Pro odhad parametrů v logistickém regresním modelu musíme mít k dispozici n > k nezávislých pozorování $y_1, \ldots, y_n$ závisle proměnné veličiny a příslušných regresorů $x_{i1}, \ldots, x_{ik}$, i = 1, …, n. Tato pozorování získáme na n objektech.

V logistickém regresním modelu nelze kvůli charakteru závisle proměnné veličiny Y použít metodu nejmenších čtverců. Odhady parametrů hledáme metodou maximální věrohodnosti. Zavedeme logaritmickou věrohodnostní funkci $\ell(\boldsymbol{\beta}; \mathbf{x}_i)$. Řešením systému věrohodnostních rovnic $\frac{\partial \ell(\boldsymbol{\beta}; \mathbf{x}_i)}{\partial \beta_j} = 0$, j = 0, 1, …, k získáme odhady $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Toto řešení nelze obecně nalézt v algebraickém tvaru, proto se hledá numericky.
Pro každý objekt pak můžeme vypočítat odhad pravděpodobnosti úspěchu či neúspěchu:

$\hat{P}(Y_i = y_i / \mathbf{X}_i = \mathbf{x}_i) = \frac{\left(e^{-\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}\right)^{1-y_i}}{1 + e^{-\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}}$. Tomuto odhadu se říká skóre a značí se $\hat{\vartheta}_i$.

6. Intervaly spolehlivosti

Směrodatná chyba odhadu regresního parametru β_j se zančí $se(\hat{\beta}_j)$. Hodnoty $se(\hat{\beta}_j)$ větší než 2 indikují numerické problémy, např. multikolonearitu mezi vysvětlujícími proměnnými nebo u kategoriálních proměnných nulové zastoupení objektů v některé kategorii. Toto upozornění se však nevztahuje na odhad směrodatné chyby odhadu β_0.

6.1. Interval spolehlivosti pro regresní parametr

100(1-α)% asymptotický interval spolehlivosti pro regresní parametr β_j má meze:

$$\left(\hat{\beta}_j - se(\hat{\beta}_j)u_{1-\alpha/2}, \hat{\beta}_j + se(\hat{\beta}_j)u_{1-\alpha/2}\right), \; j = 0, 1, \ldots, k$$

6.2. Interval spolehlivosti pro podíl šancí

Jestliže se j-tá vysvětlující proměnná zvětší o Δ (a ostatní vysvětlující proměnné se nezmění), pak podíl šancí je $e^{\Delta\beta_j}$ a 100(1-α)% asymptotický interval spolehlivosti pro podíl šancí je

$$\left(e^{\hat{\beta}_j - se(\hat{\beta}_j)u_{1-\alpha/2}}, e^{\hat{\beta}_j + se(\hat{\beta}_j)u_{1-\alpha/2}}\right), \; j = 0, 1, \ldots, k\,.$$

7. Dílčí testy významnosti regresních parametrů

Na hladině významnosti α testujeme hypotézu $H_0 : \beta_j = 0$ proti $H_1 : \beta_j \neq 0$, j = 1, 2, …, k.

Nulová hypotéza tvrdí, že j-tá vysvětlující proměnná je v modelu zbytečná.

7.1. Waldův test

Testová statistika $T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ se za platnosti H_0 asymptoticky řídí rozložením N(0,1)

(tedy statistika $T_0^2 = \left[\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right]^2$ se za platnosti H_0 asymptoticky řídí rozložením $\chi^2(1)$).

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $|T_0| \geq u_{1-\alpha/2}$ resp. $T_0^2 \geq \chi^2_{1-\alpha}(1)$.

7.2. Test poměrem věrohodnosti

Zavedeme několik nových pojmů.
Saturovaný model S má odlišný parametr pro každé pozorování (má tedy n parametrů a sám o sobě není použitelný). Každý model je jeho podmodelem. Jeho logaritmickou věrohodnostní funkci označme ℓ_S.
Námi zkoumaný model M zahrnuje k regresorů $X_1, \ldots, X_k$. Jeho logaritmickou věrohodnostní funkci označme ℓ_M.
Přiléhavost modelu M k datům lze posoudit pomocí deviance $D_M = 2(\ell_S - \ell_M)$. Přiléhavější model než saturovaný model však neexistuje, proto $\ell_S = 0$ a tudíž $D_M = -2\ell_M$. Čím je model méně přiléhavý, tím je jeho deviance vyšší.
Model M_j zahrnuje k-1 regresorů $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$. Jeho devianci označme D_{M_j}.
Pro test hypotézy, že j-tá vysvětlující proměnná je v modelu zbytečná, použijeme testovou statistiku $T_0 = D_{M_j} - D_M$, která se za platnosti H_0 asymptoticky řídí rozložením $\chi^2(1)$. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $T_0 \geq \chi^2_{1-\alpha}(1)$.

7.3. Skórový test

Označme L(**β**) věrohodnostní funkci.

Testová statistika $T_0 = \dfrac{\left.\dfrac{\partial L(\beta)}{\partial \beta}\right|_{\beta=0}}{-E\left(\left.\dfrac{\partial^2 L(\beta)}{\partial^2 \beta^2}\right|_{\beta=0}\right)}$ se za platnosti H_0 asymptoticky řídí rozložením $\chi^2(1)$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $T_0 \geq \chi^2_{1-\alpha}(1)$.

8. Test významnosti modelu jako celku

Na hladině významnosti α testujeme hypotézu $H_0 : \beta_1 = \ldots = \beta_k = 0$ proti H_1: aspoň jeden parametr $\beta_j \neq 0$. Nulová hypotéza tvrdí, že dostačující je tzv. nulový model obsahující pouze parametr β_0, tj. model $P(Y = 1/\mathbf{X} = \mathbf{x}) = \frac{1}{1 + e^{-\beta_0}}$.

Označme D_0 devianci nulového modelu a D_M devianci našeho modelu s k regresory. Testová statistika $T_0 = D_0 - D_M$ se za platnosti H_0 asymptoticky řídí rozložením $\chi^2(k)$. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $T_0 \geq \chi^2_{1-\alpha}(k)$.

9. Výstavba modelu binární logistické regrese

Při výstavbě modelu rozhodujeme, které z vysvětlujících proměnných X_1, …, X_k jsou důležité pro vysvětlení proměnné Y.

Nejprve je vhodné se zabývat významností modelu jako celku, tj. otestovat, zda je vůbec možné z daných k proměnných vytvořit model, který bude lepší než model konstanty. Použijeme test poměrem věrohodnosti, jehož testová statistika je rozdílem deviancí nulového modelu a modelu s k regresory. Není-li na dané hladině významnosti nulová hypotéza zamítnuta, nemá smysl se modelem zabývat.

9.1. Principy výstavby modelu

Jde o vytvoření takového modelu, který bude obsahovat co nejmenší množství proměnných (resp. jejich kombinací) a přitom bude ještě dostatečně dobře vysvětlovat zkoumaná data.

Na začátku celého procesu se doporučuje prozkoumat vztah mezi vysvětlovanou veličinou a každou vysvětlující veličinou zvlášť.
U spojitých proměnných použijeme dvouvýběrový test.
U kategoriálních proměnných provedeme test nezávislosti v kontingenční tabulce.
Je-li některá četnost v kontingenční tabulce nulová, budou konečné výstupy modelu s takovou proměnnou obsahovat nesmyslné hodnoty. Poměr šancí totiž bude kvůli dosazení nuly do vzorce buď nula nebo nekonečno. Této situaci zabráníme, když logicky sloučíme varianty této proměnné nebo – je-li to možné – variantu s nulovou četností vyloučíme.

9.2. Výběr podmnožiny vysvětlujících proměnných

a) Ruční postup

1. krok: Provedeme jednorozměrnou analýzu pro všechny vysvětlující proměnné a do modelu zahrneme ty, pro které test významnosti poskytne p-hodnotu menší než 0,25. (Ignoruje se původní k-rozměrná struktura dat.)

2. krok: Postupně vynecháváme proměnné, které jsou vysoce nevýznamné.

3. krok: Do modelu zahrneme ty interakce mezi proměnnými, které mají věcný smysl a jejichž p-hodnota je nejvýše 0,05. Může se stát, že při zahrnutí interakce do modelu se jedna ze vstupních proměnných stane nevýznamnou. Je lépe ji v modelu ponechat.

b) Automatizovaný postup

Používají se krokové (stepwise) metody – dopředná nebo zpětná.

10. Hodnocení vytvořeného modelu z různých hledisek

10.1. Testy dobré shody

Nulová hypotéza tvrdí, že naměřené a predikované hodnoty se neliší. Jsou založeny na porovnání naměřených hodnot y_i a odhadnutých skóre $\hat{\vartheta}_i$, i = 1, 2, …, n.

Pearsonův χ^2 test

Testová statistika má tvar $\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\vartheta}_i\right)^2}{\hat{\vartheta}_i\left(1-\hat{\vartheta}_i\right)}$ a za platnosti nulové hypotézy se asymptoticky řídí rozložením $\chi^2(n-k-1)$. Hypotézu o shodě dat s modelem tedy zamítáme na asymptotické hladině významnosti α, když $\chi^2 \geq \chi^2_{1-\alpha}(n-k-1)$.

Hosmerův – Lemeshowův test

Tento test je preferován pro spojité nezávisle proměnné a vyžaduje dostatečně rozsáhlý datový soubor. Datový soubor je uspořádán vzestupně podle skórů $\hat{\vartheta}_i$ a je rozdělen do G (zpravidla G = 10, musí být splněna podmínka G > k + 1) přibližně stejně velkých skupin. V každé z těchto skupin se zjišťuje skutečný a očekávaný počet objektů, pro něž Y = 1 resp. Y = 0.

Označme n_g rozsah g-té skupiny, $o_{1g} = \sum_{i=1}^{n_g} y_i$ resp. $o_{0g} = \sum_{i=1}^{n_g} (1 - y_i)$ skutečný počet objektů v g-té skupině, pro něž Y = 1 resp. Y = 0. Analogicky $e_{1g} = \sum_{i=1}^{n_g} \hat{\vartheta}_i$ resp. $e_{0g} = \sum_{i=1}^{n_g} (1 - \hat{\vartheta}_i)$ je očekávaný počet objektů v g-té skupině, pro něž Y = 1 resp. Y = 0.

Testová statistika $T_0 = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{(o_{kg} - e_{kg})^2}{e_{kg}}$ se za platnosti nulové hypotézy asymptoticky řídí rozložením $\chi^2(G-2)$. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $T_0 \geq \chi^2_{1-\alpha}(G-2)$.

Pro korektní použití H-L testu je nutné, aby všechny teoretické četnosti byly větší než 1 a většina z nich musí být větší než 5.

10.2. Koeficienty determinace

Tyto koeficienty porovnávají nulový model s deviancí D_0 a náš model s devianci D_M.

McFaddenův koeficient determinace: $R_{MF}^2 = 1 - \frac{D_M}{D_0}$.

Tento koeficient se získá prostým dosazením deviance místo příslušných součtů čtverců do vztahu pro koeficient determinace u lineární regrese.

Coxové – Snellův koeficient determinace: $R_{CS}^2 = 1 - e^{(D_M - D_0)/n}$

Nevýhodou tohoto koeficientu je, že nemůže překročit hodnotu $1 - e^{-D_0/n}$, je tedy vždy menší než 1, což ztěžuje interpretaci.

Nagelkerkův koeficient determinace: $R_N^2 = \frac{1 - e^{(D_M - D_0)/n}}{1 - e^{-D_0/n}}$

Nagelkerkův koeficient vznikne z Coxové – Snellova koeficientu vydělením maximální možnou hodnotou $1 - e^{-D_0/n}$.

Čím je posuzovaný model M více vzdálen od nulového modelu, tím jsou koeficienty determinace vyšší.

10.3. Informační kritéria

Informační kritéria slouží k porovnání modelů (vytvořených na týchž datech) s různým počtem regresorů. S rostoucím počtem regresorů roste i hodnota logaritmické věrohodnostní funkce (a tím i „důvěryhodnost“ modelu), na druhé straně však velký počet regresorů nemusí být vždy vhodný, např. z ekonomického hlediska.

Informační kritéria jsou navržena tak, aby penalizovala velký počet regresorů. Za lepší je považován model, který poskytuje nižší hodnotu informačního kritéria.

Označme ℓ hodnotu logaritmické věrohodnostní funkce nějakého modelu, který má k regresorů a byl vytvořen na základě datového souboru rozsahu n.

Akaikeovo informační kritérium: $\mathrm{AIC} = -2\ell + 2\mathrm{k}$

Bayesovo informační kritérium: $\mathrm{BIC} = -2\ell + \mathrm{k}\ln\mathrm{n}$

10.4. Klasifikační tabulka

Do klasifikační tabulky zaznamenáváme počty správně a nesprávně zařazených objektů:

predikce	skutečnost		celkem
	Y = 1	Y = 0	
Y = 1	a	b	a+b
Y = 0	c	d	c+d
celkem	a+c	b+d	n

Na hlavní diagonále je tedy počet objektů, které model správně predikoval. Relativní četnost správně predikovaných objektů je $\frac{a+d}{n}$.

Abychom mohli sestavit tuto klasifikační tabulku, musíme stanovit tzv. dělicí bod C pro odhadnutou pravděpodobnost úspěchu $\hat{\vartheta}_i$, i = 1, …, n. Můžeme volit jakoukoli hodnotu z intervalu (0, 1), zpravidla však predikujeme Y = 1 pro $\hat{\vartheta}_i \geq 0{,}5$ a Y = 0 pro $\hat{\vartheta}_i < 0{,}5$, tedy C = 0,5.

10.5. ROC křivka

Pomocí klasifikační tabulky lze odhadnout senzitivitu a specificitu daného klasifikačního procesu (v našem případě logistického regresního modelu).

Senzitivita … pravděpodobnost, že objekt, u něhož nastal úspěch, byl správně zařazen mezi úspěšné objekty.

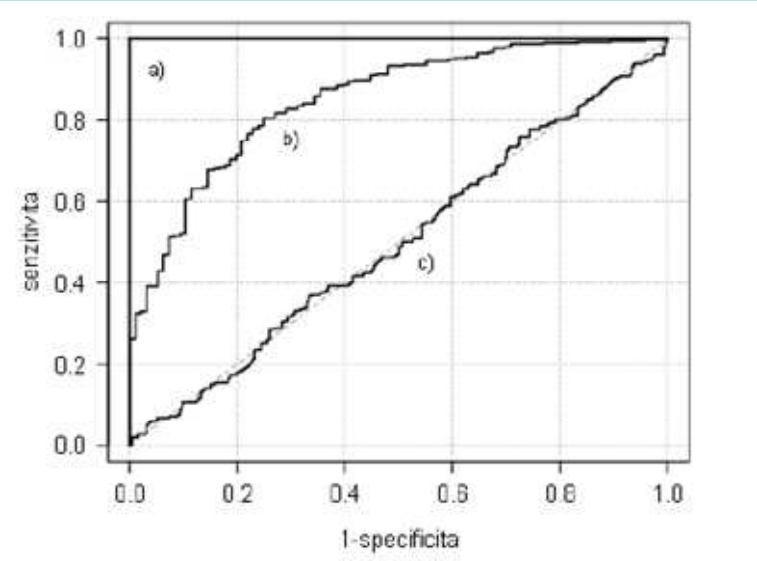
Odhad senzitivity: $\mathrm{TPF} = \frac{a}{a+c}$ (true positive fraction – relativní četnost správně klasifikovaných úspěšných objektů).

Specificita … pravděpodobnost, že objekt, u něhož nastal neúspěch, byl správně zařazen mezi neúspěšné objekty.

Odhad specificity: $\mathrm{TNF} = \frac{d}{b+d}$ (true negative fraction – relativní četnost správně klasifikovaných neúspěšných objektů).

Pro různé hodnoty dělicího bodu C dosáhneme různé hodnoty odhadů senzitivity a specificity. Graficky se jejich vztah reprezentuje právě pomocí ROC křivky. Na vodorovnou osu vynášíme hodnoty 1 – TNF (tj. 1 – odhad specificity) a na svislou osu hodnoty TPF (tj. odhady senzitivity).
Pro ideální model má ROC křivka tvar lomené čáry procházející body [0;0], [0;1] a [1;1].
Pro náhodný model ROC křivka kopíruje úsečku spojující body [0;0] a [1;1].
ROC křivka pro reálný model by měla být pokud možno co nejblíže ke křivce pro ideální model.

Ukázka ROC křivek pro ideální model (a), reálný model (b), náhodný model (c)



(Obrázek je převzat z bakalářské práce Filipa Zlámala Logistická regrese v R)

Velikost A plochy AUC pod ROC křivkou je nejběžnější kvantitativní index popisující ROC křivku. Predikční schopnost modelu hodnotíme pomocí tabulky:

A	hodnocení
0,9 - 1	výborně
0,8 – 0,9	velmi dobře
0,7 – 0,8	dobře
0,6 - 0,7	dostatečně
0,5 – 0,6	nedostatečně

11. Příklad

V souboru 50 rodin byly zjišťovány tyto údaje:

- zda v posledních dvou letech rodina navštívila jistou rekreační oblast (veličina ID, nabývá hodnoty 1 pro odpověď „ano“, hodnoty 2 pro odpověď „ne“)
- roční příjem v tisících dolarů (veličina X_1)
- postoj k cestování (veličina X_2, devítibodová škála, 1 = naprosto odmítavý, 9 = veskrze kladný)
- význam přičítaný rodinné dovolené (veličina X_3, devítibodová škála, 1 = nejnižší, 9 = nejvyšší)
- počet členů rodiny (veličina X_4)
- věk nejstaršího člena rodiny (veličina X_5).

Sestavte model binární logistické regrese, který pro náhodně vybranou rodinu umožní predikovat pravděpodobnost, že navštíví danou rekreační oblast. Závisle proměnnou veličinou je tedy ID. Na návštěvu rekreační oblasti může mít vliv pět spojitých veličin X_1, …, X_5.
To posoudíme pomocí dvouvýběrových testů. Veličiny, jejichž p-hodnota bude menší než 0,25, zařadíme do modelu binární logistické regrese.

Úkol 1.: Před provedením dvouvýběrových testů ověřte normalitu proměnných X_1, …, X_5 ve skupinách rodin, které navštívily resp. nenavštívily danou rekreační oblast.

Výsledky pro rodiny, které oblast nenavštívily:

Proměnná	Testy normality (dovolena.sta) Zhrnout podmínku: ID=0		
	N	W	p
X1: roční příjem v tisících dolarů	29	0,940188	0,101411
X2: postoj k cestování (škála 9 bodů)	29	0,964071	0,412187
X3: význam rodinné dovolené (škála 9 bodů)	29	0,964432	0,420319
X4: počet členů rodiny	29	0,917696	0,026668
X5: věk nejstaršího člena	29	0,944508	0,131598

Výsledky pro rodiny, které oblast navštívily:

Proměnná	Testy normality (dovolena.sta) Zhrnout podmínku: ID=1		
	N	W	p
X1: roční příjem v tisících dolarů	21	0,935874	0,180430
X2: postoj k cestování (škála 9 bodů)	21	0,930271	0,139382
X3: význam rodinné dovolené (škála 9 bodů)	21	0,934717	0,171087
X4: počet členů rodiny	21	0,928224	0,126815
X5: věk nejstaršího člena	21	0,967589	0,679311

Na hladině významnosti 0,05 zamítáme hypotézu o normalitě u veličiny X_4 ve skupině rodin, které danou rekreační oblast nenavštěvují.

Úkol 2.: Na hladině významnosti 0,05 testujte dvouvýběrovým t-testem, že rozložení proměnných X_1, …, X_5 v obou skupinách rodin je stejné.

Výsledky dvouvýběrového t-testu:

Proměnná	t-testy; grupováno: ID (dovolena.sta) Skup. 1: návštěva ne Skup. 2: návštěva ano Průměr návštěva ne	Průměr návštěva ano	t	sv	p	Poč.plat návštěva ne	Poč.plat. návštěva ano	Sm.odch. návštěva ne	Sm.odch. návštěva ano	F-poměr Rozptyly	p Rozptyly
X1	42,84483	59,76190	-7,40751	48	0,000000	29	21	7,013894	9,142783	1,699176	0,193069
X2	4,24138	5,14286	-1,89805	48	0,063712	29	21	1,661651	1,651839	1,011916	0,995884
X3	4,27586	5,76190	-3,15623	48	0,002760	29	21	1,623412	1,670472	1,058816	0,872933
X4	3,72414	4,33333	-1,73042	48	0,089980	29	21	1,130630	1,354006	1,434168	0,372786
X5	46,93103	53,61905	-3,01289	48	0,004122	29	21	7,568407	7,990471	1,114643	0,776989

Na hladině významnosti 0,05 zamítáme hypotézu o shodě středních hodnot pouze pro proměnné X_1, X_3 a X_5. Do modelu logistické regrese však zahrneme všechny proměnné, protože i u proměnných X_2 a X_4 jsou odpovídající p-hodnoty menší než 0,25.

Úkol 3.: Pomocí testu poměrem věrohodnosti s hladinou významnosti 0,05 zjistěte, zda má smysl uvažovat model binární logistické regrese s pěti nezávisle proměnnými veličinami X_1, …, X_5. Navíc porovnejte devianci nulového modelu s deviancí modelu s uvedenými pěti nezávisle proměnnými veličinami.

	Testování glonální nulové hypotézy: BETA=0 (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, žeID1 = návštěva ano (Vzorek pro analýzu)		
	Chí-kvadrát	SV	p
Poměr věrohodnos	47,143991	5	0,000000
Skóre	30,885726	5	0,000010
Wald.	8,840059	5	0,115616

Zajímá nás test poměrem věrohodnosti. Protože jeho p-hodnota je blízká 0, hypotézu o nevýznamnosti modelu zamítáme na hladině významnosti 0,05.

Deviance nulového modelu je 68,0292, deviance modelu s pěti nezávisle proměnnými je 20,8852. Jak ukázal test poměrem věrohodnosti, pokles deviance je významný na hladině významnosti 0,05.

Úkol 4.: Odhadněte parametry modelu a podle výsledku Waldova testu ponechte v modelu ty proměnné, pro něž jsou p-hodnoty menší než 0,25. Interpretujte podíly šancí v tomto novém modelu a proveďte test poměrem věrohodností.

Odhady parametrů:

Efekt	ID - Odhady parametrů (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, že ID = návštěva ano							
	Úroveň Efekt	Sloupec	Odhad	Standard chyba	Wald. Stat.	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1	-34,7469	12,50700	7,718363	-59,2602	-10,2336	0,005466
X1		2	0,3203	0,11219	8,148696	0,1004	0,5401	0,004309
X2		3	0,8072	0,44522	3,287311	-0,0654	1,6799	0,069817
X3		4	0,7156	0,38627	3,432061	-0,0415	1,4727	0,063942
X4		5	-0,0885	0,51445	0,029594	-1,0968	0,9198	0,863416
X5		6	0,2404	0,10208	5,545822	0,0403	0,4405	0,018525
Měřítko			1,0000	0,00000		1,0000	1,0000	

Ve sloupci Odhad vidíme odhady regresních parametrů, dále směrodatné chyby těchto odhadů, hodnoty Waldových statistik pro test nevýznamnosti regresních parametrů, meze 95% intervalů spolehlivosti pro regresní parametry a p-hodnoty pro test nevýznamnosti regresních parametrů. (Řádek Měřítko nehraje roli.) Pouze proměnná X_4 je vysoce nevýznamná, ostatní proměnné v modelu ponecháme.

Dostaneme novou tabulku odhadů parametrů:

Efekt	Úroveň Efekt	Sloupec	Odhad	Standard chyba	Wald. Stat.	Dolní LS 95,0%	Horní LS 95,0%	p
ID1 - Odhady parametrů (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, že ID1 = návštěva ano								
Abs.člen		1	-35,0954	12,43292	7,968074	-59,4634	-10,7273	0,004761
X1		2	0,3185	0,11222	8,057523	0,0986	0,5385	0,004532
X2		3	0,8193	0,44140	3,445366	-0,0458	1,6844	0,063429
X3		4	0,7208	0,38601	3,486900	-0,0358	1,4774	0,061856
X5		5	0,2405	0,10140	5,624647	0,0417	0,4392	0,017710
Měřítko			1,0000	0,00000		1,0000	1,0000	

Pravděpodobnost, že rodina patří do skupiny rodin, které navštěvují danou rekreační oblast, je vyjádřena rovnicí

$$P(ID = 1 / X_1 = x_1 \wedge X_2 = x_2 \wedge X_3 = x_3 \wedge X_5 = x_5) = \frac{1}{1 + e^{35,0954 - 0,3185 \cdot x_1 - 0,8193 \cdot x_2 - 0,7208 x_3 - 0,2405 x_5}}$$

Model zařadí rodinu do skupiny s ID = 1, pokud odhadnutá pravděpodobnost je aspoň 0,5.

Poměry šancí:

Efekt	ID1 - Poměry šancí (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, že ID1 = návštěva ano Úroveň Efekt	Sloupec	Šance Poměr	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1				
X1		2	1,375118	1,103621	1,713406	0,004532
X2		3	2,268945	0,955219	5,389457	0,063429
X3		4	2,056078	0,964872	4,381364	0,061856
X5		5	1,271861	1,042626	1,551497	0,017710
Měřítko			1,000000			

Zde jsou uvedeny odhady podílů šancí na návštěvu dané rekreační oblasti, 95% intervaly spolehlivosti pro teoretické podíly šancí a p-hodnoty pro testy hypotéz, že teoretické podíly šancí jsou 1, tj. pro test hypotézy, že jednotlivé regresory neovlivňují návštěvu oblasti. Povšimněte si, že p-hodnoty pro testy těchto hypotéz jsou totožné s p-hodnotami pro testy nevýznamnosti regresních parametrů.

Např. podíl šancí 1,3751 uvedený u X_1 znamená, že při vzrůstu proměnné X_1 o jednotku (tj. průměrný roční příjem se zvýší o 1000 dolarů) vzroste šance na zařazení do skupiny rodin navštěvujících danou rekreační oblast v průměru 1,3751 krát.

Test poměrem věrohodnosti:

	Testování glonální nulové hypotézy: BETA=0 (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, žeID1 = návštěva ano (Vzorek pro analýzu)		
	Chí-kvadrát	SV	p
Poměr věrohodnos	47,114384	4	0,000000
Skóre	30,871760	4	0,000003
Wald.	8,738087	4	0,067990

Úkol 5.: Proveďte Hosmerův-Lemeshowovův test a Pearsonův test dobré shody.

Výsledky Hosmerova-Lemeshowova testu:

	ID1 - Kvalita proložení: Hosmer-Lemeshow Test (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Hosmer Lemeshow = 3,8441, p hodn. = 0,870904										
Odezva	Skupi1a	Skupi2a	Skupi3a	Skupi4a	Skupi5a	Skupi6a	Skupi7a	Skupi8a	Skupi9a	Skupi10	Row Tot.
0: Pozorov.	5,00	5,00	5,00	5,00	4,00	2,00	3,00	0,00	0,00	0,00	29,0
Očekáv.	4,99	4,97	4,92	4,80	4,35	3,15	1,57	0,24	0,01	0,00	
1: Pozorov.	0,00	0,00	0,00	0,00	1,00	3,00	2,00	5,00	5,00	5,00	21,0
Očekáv.	0,01	0,03	0,08	0,20	0,65	1,85	3,43	4,76	4,99	5,00	
Vš. skup.	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	50,0

H-S test poskytl p-hodnotu 0,8709, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že model souhlasí s daty.

Výsledky Pearsonova testu:

	ID1 - Statistiky kvality modelu (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, žeID1 = návštěva ano (Vzorek pro analýzu)		
	SV	Stat.	Stat/sv
Odchylka	45	20,914816	0,464774
Deviance v měřít	45	20,914816	0,464774
Pearsonovo Chi2	45	19,602100	0,435602
Scaled P. Chi2	45	19,602100	0,435602
AIC		30,914816	
BIC		40,474931	
Cox-Snell R2		0,610265	
Nagelkerke R2		0,820812	
Log-věrohodnost		-10,457408	

Testová statistika Pearsonova testu nabyla hodnoty 19,602. Kritický obor $W = \langle \chi^2_{0,95}(45), \infty) = \langle 61{,}6562, \infty)$, tedy nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Úkol 6.: Určetete Nagelkerkův koeficient a AIC. AIC modelu se čtyřmi nezávisle proměnnými porovnejte s AIC modelu s pěti nezávisle proměnnými.
Nagelkerkův koeficient nabývá hodnoty 0,8208, což svědčí o tom, že náš model je značně vzdálen od nulového modelu.
AIC pro 4 nezávisle proměnné = 30,9148, AIC pro 5 nezávisle proměnných = 32,8852. Je tedy lepší model se 4 proměnnými.

Úkol 7.: Sestavte klasifikační tabulku.

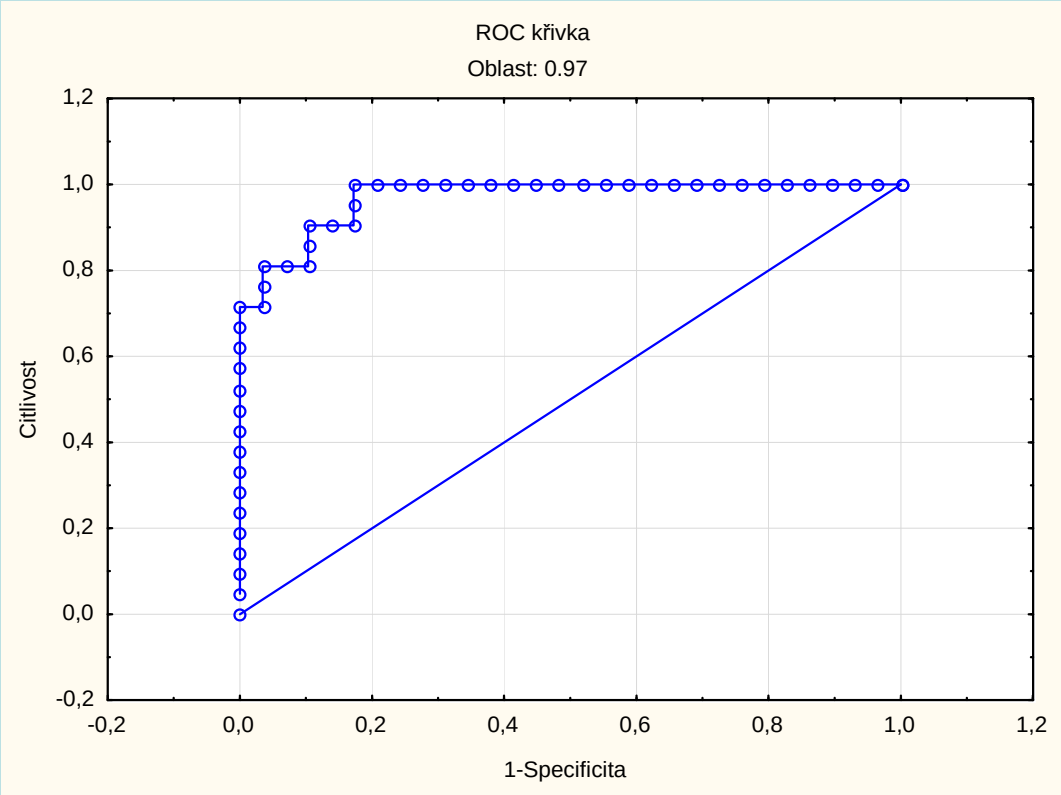
Klasifikační tabulka:

	Klasifikace případů (dovolena.sta) Odds ratio: 52,000000 Log odds ratio: 3,951244		
	Předpovězená: návštěva ano	Předpovězená: návštěva ne	Procento správných
Pozorované: návštěva ano	18	3	85,7142857
Pozorované: návštěva ne	3	26	89,6551724

Z 21 rodin, které sledovanou rekreační oblast navštívily, model správně klasifikoval 18 rodin, tj. 85,7 %.
Z 29 rodin, které sledovanou rekreační oblast nenavštívily, model správně klasifikoval 26 rodin, tj. 89,7 %.

Celková úspěšnost správné klasifikace je tedy $\frac{18+26}{50} = 88\,\%$.

Úkol 8.: Sestrojte ROC křivku.



Vidíme, že ROC křivka se blíží ideálnímu tvaru a plocha pod ní je AUC = 0,97, tedy predikční schopnost modelu je výborná.

Shrnutí:

Vytvořený model logistické regrese je statisticky významný na hladině významnosti 0,05 a není v rozporu s danými daty.

Nagelkerkova koeficientu determinace nabývá hodnoty 0,82.

Úspěšnost správné klasifikace je 88 %.

Plocha AUC pod ROC křivkou je 0,97.

Lze soudit, že pravděpodobnost návštěvy dané rekreační oblasti lze uspokojivě vysvětlit působením čtyř sledovaných proměnných (roční příjem v tisících dolarů, postoj k cestování, význam přičítaný rodinné dovolené a věk nejstaršího člena rodiny).