# Osnova přednášky „Binární logistická regrese“

# 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 $\vartheta$ 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 $\vartheta$ (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 $\beta_0$ 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 $\beta_j$ 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 $e^{\beta_j}$ 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 $\beta_j$ 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 $\beta_0$.

### 6.1. Interval spolehlivosti pro regresní parametr

100(1-α)% asymptotický interval spolehlivosti pro regresní parametr $\beta_j$ 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 $\Delta$ (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 $\ell_S$.
Námi zkoumaný model M zahrnuje k regresorů $X_1, \ldots, X_k$. Jeho logaritmickou věrohodnostní funkci označme $\ell_M$.
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 $\alpha$ 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 $\beta_0$, 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 $\alpha$, 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 $\chi^2$ 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 $\alpha$, 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 $\ell$ 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.

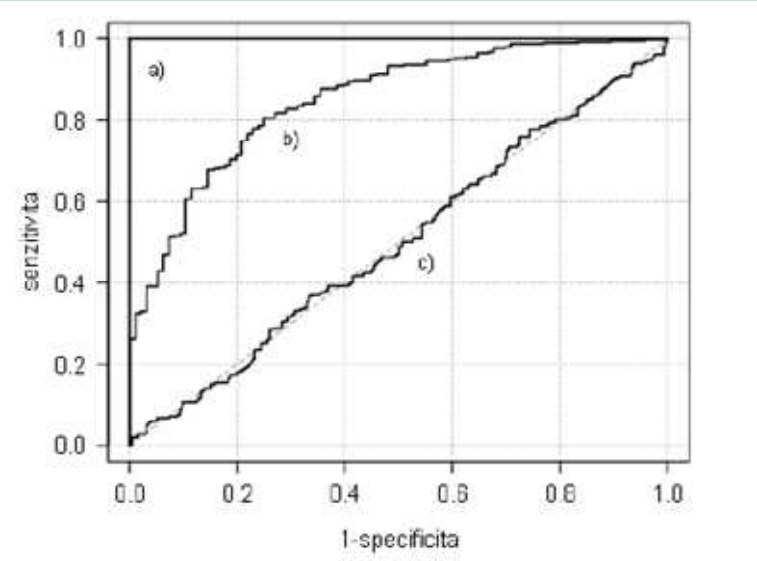
Odhad senzitivity: $\mathrm{TPF} = \frac{a}{a+c}$ (true positive fraction – relativní četnost správně klasifikovaných úspěšných objektů).

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: Model binární logistické regrese pro lékařská data

Sestavte model binární logistické regrese, který pro náhodně vybraného pacienta umožní predikovat pravděpodobnost, že se u něj vyskytne neklid po celkové anestézii. Závisle proměnnou veličinou je tedy neklid upravený (varianty 1 – vyskytl se neklid po celkové anestézii, 2 – nevyskytl se neklid po CA). Na výskyt neklidu mohou mít vliv tři kategoriální veličiny druh léku, ASA, Novalgin ano/ne a čtyři spojité veličiny věk, hmotnost, dávka, poměrová dávka.
To posoudíme pomocí testů nezávislosti a pomocí dvouvýběrových testů. Veličiny, jejichž p-hodnota bude menší než 0,25, zařadíme do modelu binární logistické regrese. Přitom u veličiny druh léku bude referenční kategorií Rapifen, u veličiny ASA kategorie II a u veličiny Novalgin ano/ne varianta ne.

**Úkol 1.:** Na hladině významnosti 0,05 testujte hypotézy, že neklid upravený a kategoriální veličiny druh léku, ASA, Novalgin ano/ne jsou nezávislé. Nezapomeňte ověřit splnění podmínek dobré aproximace.

Výsledek pro neklid upravený x druh léku:

| | Kontingenční tabulka (Nalbuphin_Rapifen.sta) Tab. : | | | |
|---|---|---|---|---|
| | Neklid upraveny | Druh léku Nalbuphin | Druh léku Rapifen | Řádk. součty |
| Četnost | neklid byl | 11 | 23 | 34 |
| Sloupc. četn. | | 19,64% | 39,66% | |
| Četnost | neklid nebyl | 45 | 35 | 80 |
| Sloupc. četn. | | 80,36% | 60,34% | |
| Četnost | Vš.skup. | 56 | 58 | 114 |

| Souhrnná tab.: Očekávané četnosti (Nalbuphin_Rapifen.sta) Pearsonův chí-kv. : 5,45188, sv=1, p=,019547 | | | |
|---|---|---|---|
| Neklid upraveny | Druh léku Nalbuphin | Druh léku Rapifen | Řádk. součty |
| neklid byl | 16,70175 | 17,29825 | 34,0000 |
| neklid nebyl | 39,29825 | 40,70175 | 80,0000 |
| Vš.skup. | 56,00000 | 58,00000 | 114,0000 |

Hypotézu o nezávislosti zamítáme na asymptotické hladině významnosti 0,05.

Výsledek pro neklid upravený x ASA:

| | Kontingenční tabulka (Nalbuphin_Rapifen.sta)<br>Tab. : | | | |
|---|---|---|---|---|
| | Neklid upraveny | ASA<br>I | ASA<br>II | Řádk.<br>součty |
| Četnost | neklid byl | 31 | 3 | 34 |
| Sloupc. četn. | | 31,00% | 21,43% | |
| Četnost | neklid nebyl | 69 | 11 | 80 |
| Sloupc. četn. | | 69,00% | 78,57% | |
| Četnost | Vš.skup. | 100 | 14 | 114 |

| Souhrnná tab.: Očekávané četnosti (Nalbuphin_Rapifen.sta)<br>Pearsonův chí-kv. : ,537548, sv=1, p=,463451 | | | |
|---|---|---|---|
| Neklid upraveny | ASA<br>I | ASA<br>II | Řádk.<br>součty |
| neklid byl | 29,8246 | 4,17544 | 34,0000 |
| neklid nebyl | 70,1754 | 9,82456 | 80,0000 |
| Vš.skup. | 100,0000 | 14,00000 | 114,0000 |

Hypotézu o nezávislosti nezamítáme na asymptotické hladině významnosti 0,05.

Výsledek pro neklid upravený x Novalgin ano/ne:

| | Kontingenční tabulka (Nalbuphin_Rapifen.sta)<br>Tab. : | | | |
|---|---|---|---|---|
| | Neklid upraveny | Novalgin ano/ne<br>0 | Novalgin ano/ne<br>1 | Řádk.<br>součty |
| Četnost | neklid byl | 9 | 25 | 34 |
| Sloupc. četn. | | 28,13% | 30,49% | |
| Četnost | neklid nebyl | 23 | 57 | 80 |
| Sloupc. četn. | | 71,88% | 69,51% | |
| Četnost | Vš.skup. | 32 | 82 | 114 |

| Souhrnná tab.: Očekávané četnosti (Nalbuphin_Rapifen.sta)<br>Pearsonův chí-kv. : ,061398, sv=1, p=,804300 | | | |
|---|---|---|---|
| Neklid upraveny | Novalgin ano/ne<br>0 | Novalgin ano/ne<br>1 | Řádk.<br>součty |
| neklid byl | 9,54386 | 24,45614 | 34,0000 |
| neklid nebyl | 22,45614 | 57,54386 | 80,0000 |
| Vš.skup. | 32,00000 | 82,00000 | 114,0000 |

Hypotézu o nezávislosti nezamítáme na asymptotické hladině významnosti 0,05.

**Úkol 2.:** Před provedením dvouvýběrových testů ověřte normalitu proměnných věk, hmotnost, dávka, poměrová dávka ve skupinách pacientů, u nichž se vyskytl resp. nevyskytl neklid.

Výsledky pro pacienty, u nichž se vyskytl neklid:

| Proměnná | Testy normality (Nalbuphin_Rapifen.sta)<br>Zhrnout podmínku: v15=1<br>N | max D | Lilliefors p | W | p |
|---|---|---|---|---|---|
| Věk | 34 | 0,233684 | p < ,01 | 0,832491 | 0,000116 |
| Hmotnost | 34 | 0,311151 | p < ,01 | 0,715116 | 0,000001 |
| Dávka | 34 | 0,333163 | p < ,01 | 0,724877 | 0,000001 |
| Dávka mg/kg | 34 | 0,372417 | p < ,01 | 0,697266 | 0,000000 |

Výsledky pro pacienty, u nichž se nevyskytl neklid:

| Proměnná | Testy normality (Nalbuphin_Rapifen.sta)<br>Zhrnout podmínku: v15=2<br>N | max D | Lilliefors p | W | p |
|---|---|---|---|---|---|
| Věk | 80 | 0,202938 | p < ,01 | 0,815354 | 0,000000 |
| Hmotnost | 80 | 0,173416 | p < ,01 | 0,746045 | 0,000000 |
| Dávka | 80 | 0,224512 | p < ,01 | 0,855445 | 0,000000 |
| Dávka mg/kg | 80 | 0,266870 | p < ,01 | 0,763710 | 0,000000 |

Ve všech případech Lilieforsův i Shapirův – Wilkův test zamítá hypotézu o normalitě. Dále tedy použijeme neparametrické testy.

**Úkol 3.:** Na hladině významnosti 0,05 testujte dvouýběrovým Wilcoxonovým testem, že rozložení proměnných věk, hmotnost, dávka, poměrová dávka ve skupinách pacientů, u nichž se vyskytl resp. nevyskytl neklid, je stejné.

Výsledky dvouvýběrového Wilcoxonova testu:

| | Mann-Whitneyův U Test (w/ oprava na spojitost) (Nalbuphin_Rapifen.sta)<br>Dle proměn. Neklid upraveny<br>Označené testy jsou významné na hladině p <,05000 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Proměnná | Sčt poř. neklid byl | Sčt poř. neklid nebyl | U | Z | p-hodn. | Z upravené | p-hodn. | N platn. neklid byl | N platn. neklid nebyl | 2*1str. přesné p |
| Věk | 1710,000 | 4845,000 | 1115,000 | -1,51438 | 0,129929 | -1,53785 | 0,124087 | 34 | 80 | 0,130388 |
| Hmotnost | 1801,000 | 4754,000 | 1206,000 | -0,95075 | 0,341733 | -0,95292 | 0,340629 | 34 | 80 | 0,343429 |
| Dávka | 1558,000 | 4997,000 | 963,000 | -2,45584 | 0,014056 | -2,47110 | 0,013470 | 34 | 80 | 0,013560 |
| Dávka mg/kg | 1464,500 | 5090,500 | 869,500 | -3,03496 | 0,002406 | -3,03682 | 0,002391 | 34 | 80 | 0,002130 |

Na hladině významnosti 0,05 se prokázal rozdíl u proměnných dávka a poměrová dávka.
U proměnné věk je p-hodnota menší než 0,25, proto ji do modelu zařadíme.

**Úkol 4.:** Pomocí testu poměrem věrohodnosti s hladinou významnosti 0,05 zjistěte, zda má smysl uvažovat model binární logistické regrese se čtyřmi nezávisle proměnnými veličinami (druh léku, věk, dávka, poměrová dávka).
Navíc porovnejte devianci nulového modelu s deviancí modelu s uvedenými čtyřmi nezávisle proměnnými veličinami.
**Návod:** Statistiky – Pokročilé lineární/nelineární modely – Zobecněné lineární/nelineární modely – Logitový model – OK – Proměnné – Závislé Neklid upravený, Kategor. nezáv. (faktory) Druh léku, Spojité nezáv. prom. Věk, Dávka, Poměrová dávka – OK – na záložce Detaily zaškrtneme Parametrizace Ref. – OK – Kval. proložení.
Ve stromové struktuře v levé části Pracovního sešitu zvolíme Testování globální nulové hypotézy:

| | Testování glonální nulové hypotézy: BETA=0 (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | |
|---|---|---|---|
| | Chí-kvadrát | SV | p |
| Poměr věrohodnos | 8,857147 | 4 | 0,064771 |
| Skóre | 7,956908 | 4 | 0,093170 |
| Wald. | 7,085899 | 4 | 0,131418 |

Zajímá nás test poměrem věrohodnosti. Protože jeho p-hodnota je 0,0648, hypotézu o nevýznamnosti modelu nezamítáme na hladině významnosti 0,05.

Devianci modelu se čtyřmi uvažovanými regresory získáme tak, že ve stromové struktuře v levé části Pracovního sešitu vybereme Statistiky kvality modelu:

| | Neklid upraveny - Statistiky kvality modelu (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | |
|---|---|---|---|
| | SV | Stat. | Stat/sv |
| Odchylka | 109 | 130,079322 | 1,193388 |
| Deviance v měřít | 109 | 130,079322 | 1,193388 |
| Pearsonovo Chi2 | 109 | 112,152652 | 1,028923 |
| Scaled P. Chi2 | 109 | 112,152652 | 1,028923 |
| AIC | | 140,079322 | |
| BIC | | 153,760315 | |
| Cox-Snell R2 | | 0,074753 | |
| Nagelkerke R2 | | 0,106123 | |
| Log-věrohodnost | | -65,039661 | |

Devianci nulového modelu získáme tak, že ve vstupní tabulce pro logistickou regresi zadáme pouze závisle proměnnou veličinu Neklid upravený a žádné nezávisle proměnné veličiny. Zvolíme Kvalita proložení a vybereme Statistiky kvality modelu.

| | Neklid upraveny - Statistiky kvality modelu (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | |
|---|---|---|---|
| | SV | Stat. | Stat/sv |
| Odchylka | 113 | 138,936469 | 1,229526 |
| Deviance v měřít | 113 | 138,936469 | 1,229526 |
| Pearsonovo Chi2 | 113 | 114,000000 | 1,008850 |
| Scaled P. Chi2 | 113 | 114,000000 | 1,008850 |
| AIC | | 140,936469 | |
| BIC | | 143,672667 | |
| Log-věrohodnost | | -69,468235 | |
| 8 | 0 | 0,000000 | 0,000000 |
| 9 | 0 | 0,000000 | 0,000000 |

Zjistíme, že deviance nulového modelu je 138,9365.
Deviance modelu se čtyřmi uvažovanými regresory je 130,0793, došlo tedy k nepříliš výraznému poklesu deviance. Test poměrem věrohodnosti poskytl p-hodnotu 0,0648, což je větší než hladina významnosti 0,05. Vidíme, že tato cesta nepovede k vybudování kvalitního modelu.

**Úkol 5.:** Vytvořte model se všemi sedmi nezávisle proměnnými, vypočtěte jeho devianci a významnost poklesu oproti nulovému modelu testujte na hladině významnosti 0,05.

Statistiky kvality modelu:

| | Neklid upraveny - Statistiky kvality modelu (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | |
|---|---|---|---|
| | SV | Stat. | Stat/sv |
| Odchylka | 106 | 118,264744 | 1,115705 |
| Deviance v měřít | 106 | 118,264744 | 1,115705 |
| Pearsonovo Chi2 | 106 | 103,244180 | 0,974002 |
| Scaled P. Chi2 | 106 | 103,244180 | 0,974002 |
| AIC | | 134,264744 | |
| BIC | | 156,154332 | |
| Cox-Snell R2 | | 0,165841 | |
| Nagelkerke R2 | | 0,235436 | |
| Log-věrohodnost | | -59,132372 | |

Deviance modelu se sedmi uvažovanými regresory je 118,2647.

Výsledek testu poměrem věrohodnosti:

| | Testování glonální nulové hypotézy: BETA=0 (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | |
|---|---|---|---|
| | Chí-kvadrát | SV | p |
| Poměr věrohodnos | 20,671725 | 7 | 0,004288 |
| Skóre | 17,575353 | 7 | 0,014040 |
| Wald. | 13,622377 | 7 | 0,058320 |

Vidíme, že došlo k významnému poklesu deviance, protože test poměrem věrohodnosti poskytl p-hodnotu 0,0043.

**Úkol 6.:** 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. Proveďte test poměrem věrohodnosti.

**Návod:** V okně výsledků zvolíme tlačítko Odhady a ve stromové struktuře v levé části Pracovního sešitu vybereme tabulku Neklid upravený – Odhady parametrů.

| Efekt | Neklid upraveny - Odhady parametrů (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, že Neklid upraveny = neklid byl | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Úroveň Efekt | Sloupec | Odhad | Standard chyba | Wald. Stat. | Dolní LS 95,0% | Horní LS 95,0% | p |
| Abs.člen | | 1 | -1,21016 | 1,00336 | 1,454692 | -3,1767 | 0,75639 | 0,227777 |
| Věk | | 2 | -0,60297 | 0,24680 | 5,969086 | -1,0867 | -0,11926 | 0,014559 |
| Hmotnost | | 3 | 0,17389 | 0,06729 | 6,678929 | 0,0420 | 0,30578 | 0,009756 |
| Dávka | | 4 | -0,86261 | 0,55172 | 2,444501 | -1,9440 | 0,21874 | 0,117937 |
| Dávka mg/kg | | 5 | 8,53431 | 15,09212 | 0,319769 | -21,0457 | 38,11432 | 0,571747 |
| Druh léku | Nalbuphin | 6 | -0,53602 | 1,76557 | 0,092169 | -3,9965 | 2,92444 | 0,761437 |
| ASA | I | 7 | 0,24545 | 0,79510 | 0,095297 | -1,3129 | 1,80381 | 0,757548 |
| Novalgin ano/ne | 0 | 8 | 1,36682 | 0,90154 | 2,298540 | -0,4002 | 3,13381 | 0,129496 |
| Měřítko | | | 1,00000 | 0,00000 | | 1,0000 | 1,00000 | |

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.)

V modelu ponecháme proměnné Věk, Hmotnost, Dávka, Novalgin ano/ne.

Dostaneme novou tabulku odhadů parametrů:

| Efekt | Úroveň Efekt | Sloupec | Odhad | Standard chyba | Wald. Stat. | Dolní LS 95,0% | Horní LS 95,0% | p |
|---|---|---|---|---|---|---|---|---|
| Neklid upraveny - Odhady parametrů (Nalbuphin_Rapifen.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, že Neklid upraveny = neklid byl | | | | | | | | |
| Abs.člen | | 1 | -0,790356 | 0,586483 | 1,816080 | -1,93984 | 0,359129 | 0,177781 |
| Věk | | 2 | -0,607933 | 0,243202 | 6,248550 | -1,08460 | -0,131267 | 0,012429 |
| Hmotnost | | 3 | 0,168010 | 0,064711 | 6,740808 | 0,04118 | 0,294842 | 0,009423 |
| Dávka | | 4 | -0,616158 | 0,202831 | 9,228171 | -1,01370 | -0,218616 | 0,002383 |
| Novalgin ano/ne | 0 | 5 | 1,357499 | 0,730567 | 3,452699 | -0,07439 | 2,789385 | 0,063149 |
| Měřítko | | | 1,000000 | 0,000000 | | 1,00000 | 1,000000 | |

Tabulka podílů šancí:

| Efekt | Úroveň Efekt | Sloupec | Šance Poměr | Dolní LS 95,0% | Horní LS 95,0% | p |
|---|---|---|---|---|---|---|
| Neklid upraveny - Poměry šancí (Nalbuphin_Rapifen.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, že Neklid upraveny = neklid byl | | | | | | |
| Abs.člen | | 1 | | | | |
| Věk | | 2 | 0,544475 | 0,338037 | 0,87698 | 0,012429 |
| Hmotnost | | 3 | 1,182949 | 1,042038 | 1,34291 | 0,009423 |
| Dávka | | 4 | 0,540015 | 0,362874 | 0,80363 | 0,002383 |
| Novalgin ano/ne | 0 | 5 | 3,886462 | 0,928313 | 16,27101 | 0,063149 |
| Měřítko | | | 1,000000 | | | |

Pokud se věk pacienta zvýší o rok, poklesne šance na výskyt neklidu 0,54x. Podobně pro ostatní parametry.

Test poměrem věrohodnosti:

| | Chí-kvadrát | SV | p |
|---|---|---|---|
| Testování glonální nulové hypotézy: BETA=0 (Nalbuphin_Rapifen.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | | |
| Poměr věrohodnos | 20,203311 | 4 | 0,000455 |
| Skóre | 17,223083 | 4 | 0,001749 |
| Wald. | 13,674513 | 4 | 0,008410 |

**Úkol 7.:** Proveďte H-S test a Pearsonův test dobré shody.

Výsledky Hosmerova-Lemeshowova testu jsou uvedeny v tabulce Neklid upravený – Kvalita proložení.

| | Neklid upraveny - Kvalita proložení: Hosmer-Lemeshow Test (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Hosmer Lemeshow = 6,3157, p hodn. = 0,611911 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Odezva | Skupi1a | Skupi2a | Skupi3a | Skupi4a | Skupi5a | Skupi6a | Skupi7a | Skupi8a | Skupi9a | Skupi10 | Row Tot. |
| 0: Pozorov. | 11,0 | 10,0 | 9,0 | 8,0 | 10,0 | 8,0 | 6,0 | 8,0 | 3,0 | 7,0 | 80 |
| Očekáv. | 10,6 | 10,2 | 9,5 | 8,6 | 9,5 | 7,6 | 7,0 | 6,5 | 5,6 | 4,8 | |
| 1: Pozorov. | 0,0 | 1,0 | 2,0 | 3,0 | 3,0 | 3,0 | 5,0 | 3,0 | 8,0 | 6,0 | 34 |
| Očekáv. | 0,4 | 0,8 | 1,5 | 2,4 | 3,5 | 3,4 | 4,0 | 4,5 | 5,4 | 8,2 | |
| Vš. skup. | 11,0 | 11,0 | 11,0 | 11,0 | 13,0 | 11,0 | 11,0 | 11,0 | 11,0 | 13,0 | 114 |

H-S test poskytl p-hodnotu 0,6119, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že model souhlasí s daty.

Výsledky Pearsonova testu jsou uvedeny v tabulce Neklid upravený – Statistiky kvality modelu.

| | Neklid upraveny - Statistiky kvality modelu (Nalbuphin_Rapifen.sta)<br>Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT<br>Modelovaná pravděpodobnost, žeNeklid upraveny = neklid byl (Vzorek pro analýzu) | | |
|---|---|---|---|
| | SV | Stat. | Stat/sv |
| Odchylka | 109 | 118,733158 | 1,089295 |
| Deviance v měřít | 109 | 118,733158 | 1,089295 |
| Pearsonovo Chi2 | 109 | 104,620692 | 0,959823 |
| Scaled P. Chi2 | 109 | 104,620692 | 0,959823 |
| AIC | | 128,733158 | |
| BIC | | 142,414151 | |
| Cox-Snell R2 | | 0,162406 | |
| Nagelkerke R2 | | 0,230560 | |
| Log-věrohodnost | | -59,366579 | |

Testová statistika Pearsonova testu nabyla hodnoty 104,6207.

Kritický obor $W = \langle \chi^2_{0,95}(109), \infty) = \langle 134{,}3688, \infty)$, tedy nulovou hypotézu nezamítáme na hladině významnosti 0,05.

**Úkol 8.:** Vypočtěte Nagelkerkův koeficient.
Nagelkerkův koeficient je uveden v tabulce Neklid upravený – Statistiky kvality modelu. Nabývá hodnoty 0,2306, což svědčí o tom, že náš model není příliš vzdálen od nulového modelu.

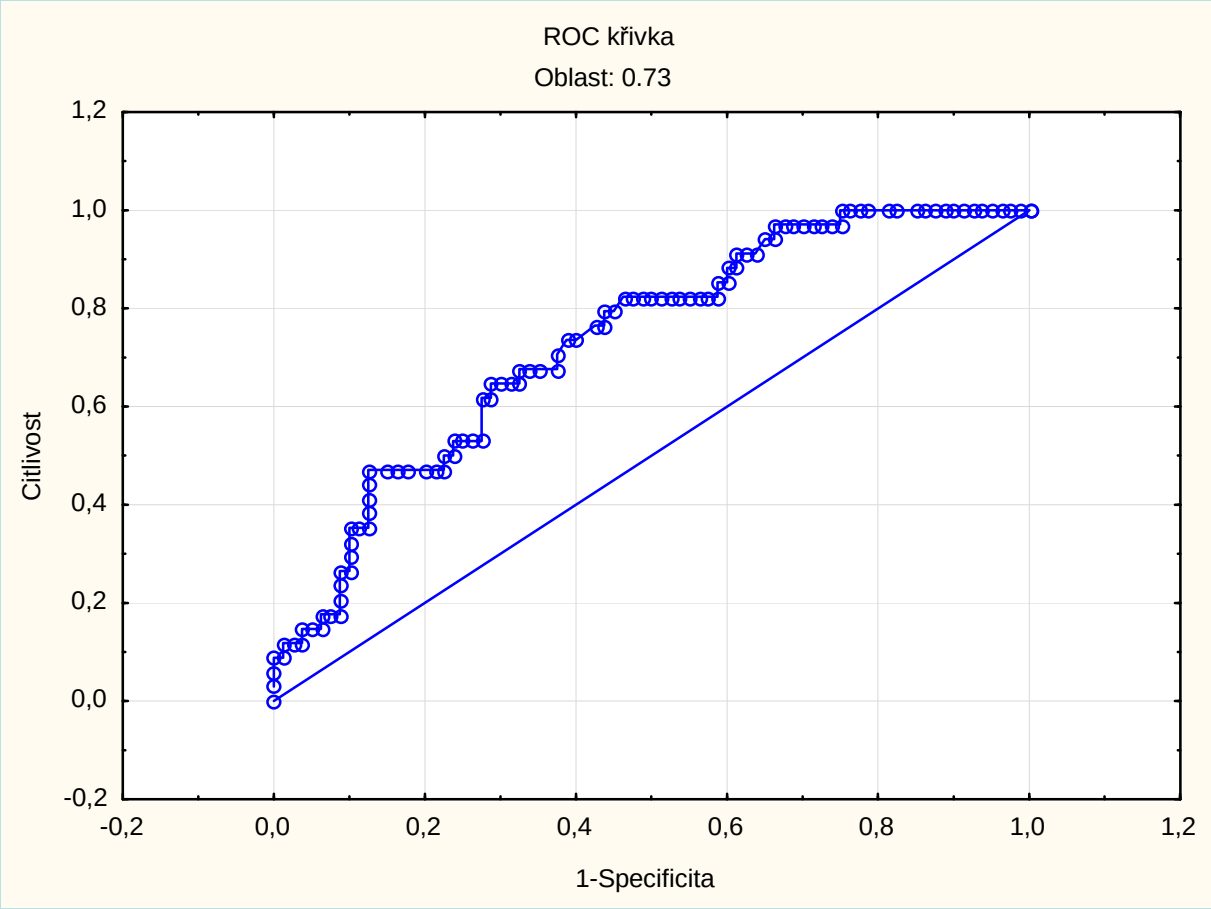
**Úkol 9.:** Sestavte klasifikační tabulku.
Ve výstupní tabulce GLZ – Výsledky vybereme záložku Rezid. 1 – Klasif & odds ratio.

Klasifikační tabulka:

| | Klasifikace případů (Nalbuphin_Rapifen.sta)<br>Odds ratio: 4,304348<br>Log odds ratio: 1,459626 | | |
|---|---|---|---|
| | Předpovězená: neklid byl | Předpovězená: neklid nebyl | Procento správných |
| Pozorované: neklid byl | 11 | 23 | 32,3529412 |
| Pozorované: neklid nebyl | 8 | 72 | 90 |

Z 34 pacientů, u nichž se vyskytl neklid po CA, model správně zařadil 11 pacientů, tj. odhad senzitivity = 32,4 %.
Z 80 pacientů, u nichž se neklid po CA nevyskytl, model správně zařadil 72,
tj. odhad specificity = 90 %.
Celkové procento úspěšné klasifikace je tedy 83/114 = 72,8 %.

**Úkol 10.:** Sestrojte ROC křivku.
Ve výstupní tabulce GLZ – Výsledky vybereme záložku Rezid. 1 – ROC křivka




Vidíme, že ROC křivka je poněkud vzdálena ideálnímu tvaru a plocha pod ní je AUC = 0,73, tedy predikční schopnost modelu je pouze dobrá.

**Shrnutí:**

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

Nagelkerkův koeficient determinace nabývá hodnoty 0,23.

Úspěšnost správné klasifikace je 72,8 %.

Plocha AUC pod ROC křivkou je 0,73.

Lze soudit, že pravděpodobnost výskytu neklidu po celkové anestézii lze uspokojivě vysvětlit působením čtyř sledovaných proměnných (Věk, Hmotnost, Dávka, Novalgin ano/ne).