

# **Osnova přednášky „Binární logistická regrese“**

- 1. Motivace**
- 2. Odvození modelu**
- 3. Kódování proměnných**
  - 3.1. Příklad dvou kategorií**
  - 3.2. Příklad aspoň tří kategorií**
- 4. Význam parametrů**
- 5. Odhady parametrů**
- 6. Interval 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  $X_1, \dots, X_k$ , 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 bankovníctví, 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:

$$\pi(y) = \begin{cases} \vartheta^y (1 - \vartheta)^{1-y} & \text{pro } y = 0, 1 \\ 0 & \text{jinak} \end{cases} .$$

Jev  $\{Y = 1\}$  často interpretujeme jako úspěch, jev  $\{Y = 0\}$  jako

neúspěch.

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

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

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

Pro predikci hodnot veličiny  $Y$  nelze použít lineární regresní model tvaru

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

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  $\mathbf{x}$ ) 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  $\mathbf{x}$  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  $\mathbf{X} = \mathbf{x}$ , 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  $x_1, \dots, x_k$ .

**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. Příklad 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. Příklad aspoň tří kategorií

Vysvětlující proměnná  $X$  má  $r \geq 3$  kategorií, např.  $X$  udává barvu vlasů a má tři kategorie: černá, blond, rezavá.

**3.2.1. Kódování pomocí referenční kategorie:** zavedeme  $r-1$  nezávislých indikátorových proměnných (indikátorů)  $Z_1, \dots, Z_{r-1}$ . Každý z těchto indikátorů odpovídá jedné kategorii proměnné  $X$ , vynechaná kategorie je referenční, kódujeme ji pomocí 0 a můžeme si ji sami zvolit.

$$Z_1 = \begin{cases} 1 & \text{pro černé vlasy} \\ 0 & \text{jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 & \text{pro blond vlasy} \\ 0 & \text{jinak} \end{cases}. \quad \text{Vyjádřeno tabulkou:}$$

	černé vlasy	blond vlasy	rezavé vlasy
$Z_1$	1	0	0
$Z_2$	0	1	0

V tomto případě roli referenční kategorie hrají rezavé vlasy, jsou kódovány 0.

**3.2.2. Kódování pomocí sigma omezené parametrizace:** na rozdíl od předešlého případu se referenční kategorie kóduje pomocí -1.

$$Z_1 = \begin{cases} 1 \text{ pro černé vlasy} \\ 0 \text{ jinak} \end{cases}, Z_2 = \begin{cases} 1 \text{ pro blond vlasy} \\ 0 \text{ jinak} \end{cases}. \text{ Vyjádřeno tabulkou:}$$

	černé vlasy	blond vlasy	rezavé vlasy
$Z_1$	1	0	-1
$Z_2$	0	1	-1

Součet v každém řádku tabulky je 0.

**3.2.3. Kódování přeparametrizovaného modelu:** v tomto případě máme  $r$  závislých indikátorů  $Z_1, \dots, Z_r$ .

$$Z_1 = \begin{cases} 1 \text{ pro černé vlasy} \\ 0 \text{ jinak} \end{cases}, Z_2 = \begin{cases} 1 \text{ pro blond vlasy} \\ 0 \text{ jinak} \end{cases}, Z_3 = \begin{cases} 1 \text{ pro rezavé vlasy} \\ 0 \text{ jinak} \end{cases}. \text{ Vyjádřeno tabulkou:}$$

	černé vlasy	blond vlasy	rezavé vlasy
$Z_1$	1	0	0
$Z_2$	0	1	0
$Z_3$	0	0	1



## 4. Význam parametrů

Ze vztahu pro logit

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

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**:

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

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  $y_1, \dots, y_n$  závisle proměnné veličiny a příslušných regresorů  $\mathbf{x}_{i1}, \dots, \mathbf{x}_{ik}, i = 1, \dots, n$ . 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.

Logaritmická věrohodnostní funkce pro model binární logistické regrese má tvar:

$$\ell(\boldsymbol{\beta}; \mathbf{x}_i) = \ln \prod_{i=1}^n \frac{(e^{-\mathbf{x}_i^T \boldsymbol{\beta}})^{1-y_i}}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}} = \sum_{i=1}^n \ln \frac{(e^{-\mathbf{x}_i^T \boldsymbol{\beta}})^{1-y_i}}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}} = \sum_{i=1}^n \left[ (y_i - 1) \mathbf{x}_i^T \boldsymbol{\beta} - \ln(1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}) \right].$$
 Řešením systému

věrohodnostních rovnic  $\frac{\partial \ell(\boldsymbol{\beta}; \mathbf{x}_i)}{\partial \beta_j} = 0, j = 0, 1, \dots, k$  získáme odhady  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ . Toto řešení

nelze obecně nalézt v algebraickém tvaru, proto se hledá numericky, např. použitím Newtonovy – Raphsonovy iterační metody.

Pro každý objekt pak můžeme vypočítat odhad pravděpodobnosti úspěchu či neúspěchu:

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

## 6. Intervaly spolehlivosti

Směrodatná chyba odhadu regresního parametru  $\beta_j$  je  $se(\hat{\beta}_j) = \sqrt{v_{jj}}$ , kde  $v_{jj}$  je j-tý diagonální

prvek matice  $\hat{\mathbf{V}} = \left[ -\frac{\partial^2 \ell(\hat{\boldsymbol{\beta}}; \mathbf{x})}{\partial \beta_j \partial \beta_1} \right]^{-1}$ .

Hodnoty  $se(\hat{\beta}_j)$  větší než 2 indikují numerické problémy, např. multikolinearitu 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- $\alpha$ )% asymptotický interval spolehlivosti pro regresní parametr  $\beta_j$  má meze:

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

### 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- $\alpha$ )% asymptotický interval spolehlivosti pro podíl šancí je

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

## 7. Dílčí testy významnosti regresních parametrů

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

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

### 7.1. Waldův test

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

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

Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$ , 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  $X_1, \dots, X_k$ . 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  $D_M = -2\ell_M$  neboli

$$D_M = \sum_{i=1}^n \left[ (y_i - 1) \mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \ln(1 + e^{-\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}) \right].$$
 Čím je model méně přiléhavý, tím je jeho deviance vyšší.

Model  $M_j$  ... zahrnuje  $k-1$  regresorů  $X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_k$ . 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  $\alpha$ , když  $T_0 \geq \chi^2_{1-\alpha}(1)$ .

**Upozornění:** Test poměrem věrohodnosti lze použít i v případě, kdy porovnáváme model  $M$  a jeho podmodel  $\tilde{M}$ . Předpokládáme, že model  $M$  má  $k_1$  regresorů a model  $\tilde{M}$  má  $k_2 < k_1$  regresorů. Testujeme hypotézu, že dostačující je model  $\tilde{M}$ , tedy  $k_1 - k_2$  regresorů je nulových proti alternativě, že aspoň jeden je nenulový. Testová statistika  $T_0 = D_{\tilde{M}} - D_M$  se za platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(k_1 - k_2)$ . Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $T_0 \geq \chi^2_{1-\alpha}(k_1 - k_2)$ .

### 7.3. Skórový test

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

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

## 8. Test významnosti modelu jako celku

Na hladině významnosti  $\alpha$  testujeme hypotézu  $H_0 : \beta_1 = \dots = \beta_k = 0$  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  $X_1, \dots, X_k$  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  $\hat{v}_i$ ,  $i = 1, 2, \dots, n$ .

#### Pearsonův $\chi^2$ test

Testová statistika má tvar  $\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{v}_i)^2}{\hat{v}_i(1 - \hat{v}_i)}$  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  $\alpha$ , 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 skóreů  $\hat{\vartheta}_i$  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 deviancí  $D_M$ .

**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:  $AIC = -2\ell + 2k$

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

## 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  $\hat{\vartheta}_i$ ,  $i = 1, \dots, n$ . 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:  $TPF = \frac{a}{a + c}$  (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:  $TNF = \frac{d}{b + d}$  (true negative fraction – relativní četnost správně klasifikovaných neúspěšných objektů).



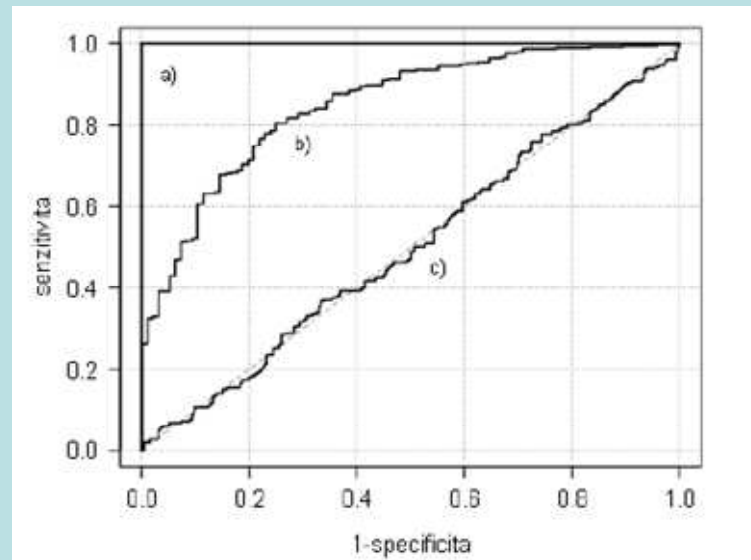
Pro různé hodnoty dělicího bodu C dosáhneme různé hodnoty odhadů senzitivity a specificity. Graficky se jejich vztah reprezentuje právě pomocí ROC křivky. Na vodorovnou osu vynášíme hodnoty  $1 - TNF$  (tj.  $1 -$  odhad specificity) a na svislou osu hodnoty TPF (tj. odhady senzitivity).

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 nejbližší 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 0 pro odpověď „ne“, hodnoty 1 pro odpověď „ano“)
- 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  $X_5$ ).

a) Proveďte jednorozměrnou analýzu dat.

b) Sestavte model binární logistické regrese, který umožní odhadnout pravděpodobnost návštěvy dané rekreační oblasti v závislosti na vhodně vybraných veličinách  $X_1, \dots, X_5$ .

Odhadněte regresní parametry a vypočtete odhady podílů šancí

Sestrojte 95% intervaly spolehlivosti pro regresní parametry a pro podíly šancí.

Na hladině významnosti 0,05 proveďte dílčí testy významnosti.

c) Na hladině významnosti 0,05 proveďte celkový test významnosti.

d) Proveďte hodnocení kvality modelu

## Řešení:

Ad a)

Testy normality proměnných  $X_1, \dots, X_5$  v daných dvou skupinách:

Pro skupinu rodin, které danou rekreační oblast nenavštěvují: Statistika – Základní statistiky/tabulky – Select cases – ID=0 – OK – Tabulky četností – Proměnné X1 až X5 – OK – Normalita – zaškrtneme S-W test – Testy normality

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

Pro skupinu rodin, které danou rekreační oblast navštěvují: Statistika – Základní statistiky/tabulky – Select cases – ID=1 – OK – Tabulky četností – Proměnné X1 až X5 – OK – Normalita – zaškrtneme S-W test – Testy normality

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.

Dvouvýběrové t-testy:

Statistiky – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK, Proměnné – Závislé proměnné X1 až X5, Grupovací proměnná ID – OK.

Po kliknutí na tlačítko Výpočet dostaneme tabulku

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<sub>1</sub>, X<sub>3</sub> a X<sub>5</sub>. Do modelu logistické regrese však zahrneme všechny proměnné, protože i u proměnných X<sub>2</sub> a X<sub>4</sub> jsou odpovídající p-hodnoty menší než 0,25.

Ad b)

Provedení logistické regrese

Statistiky – Pokročilé lineární/nelineární modely – Zobecněné lineární/nelineární modely – Logitový model – OK – Proměnné – Závislé ID, Spojité nezáv. prom.  $X_1, \dots, X_5$  – OK – na záložce Detaily zaškrtneme Parametrizace Ref. – OK.

V okně výsledků zvolíme tlačítko Odhady a ve stromové struktuře v levé části Pracovního sešitu vybereme tabulku ID – Odhady parametrů.

ID - Odhady parametrů (dovolena.sta)								
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT								
Modelovaná pravděpodobnost, že ID = návštěva ano								
Efekt	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.

## Nová tabulka:

ID - Odhady parametrů (dovolená.sta)								
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT								
Modelovaná pravděpodobnost, že ID = návštěva ano								
Efekt	Úroveň Efekt	Sloupec	Odhad	Standard chyba	Wald. Stat.	Dolní LS 95,0%	Horní LS 95,0%	p
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

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

Např. :

Parametr 0,3185 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 na  $e^{0,3185} = 1,375$  krát.

Model zařadí rodinu do skupiny s  $\text{ID} = 1$ , pokud odhadnutá pravděpodobnost je aspoň 0,5.

Dále se podíváme na tabulku ID – Poměry šancí.

ID - Poměry šancí (dovolena.sta)						
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT						
Modelovaná pravděpodobnost, že ID = návštěva ano						
Efekt	Úroveň Efekt	Sloupec	Šance Poměr	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1				
X1		2	1,375118	1,155175	1,595062	0,004532
X2		3	2,268945	1,403815	3,134075	0,063429
X3		4	2,056078	1,299518	2,812638	0,061856
X5		5	1,271861	1,073123	1,470600	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 je 1, tj. pro test hypotézy, že jednotlivé regresory neovlivňuje návštěvu oblasti. Povšimněte si, že p-hodnoty pro testy těchto hypotéz jsou totožné s p-hodnotami pro testy nevýznamnosti regresních parametrů - viz tabulka ID – Odhady parametrů.



Ad c)

Celkový test významnosti provedeme takto: Návrat do GLZ-Výsledky – Kval. proložení – vybereme tabulku Testování globální nulové hypotézy.

	Testování glonální nulové hypotézy: BETA=0 (dovolena.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, žeID = 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

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

Ad d)

Hodnocení kvality modelu provedeme několika způsoby.

Ve výstupní tabulce GLZ – Výsledky vybereme Kvalita proložení a zvolíme tabulku ID – Statistiky kvality modelu.

	ID - Statistiky kvality modelu (dovolena.sta)		
	Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT		
	Modelovaná pravděpodobnost, žeID = 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	

Pearsonův  $\chi^2$  test dobré shody (nulová hypotéza tvrdí, že naměřené a predikované hodnoty se neliší): na řádce Pearsonovo Chi2 je uvedena hodnota testové statistiky, v našem případě 19,6.

Kritický obor  $W = \langle \chi^2_{0,95}(45), \infty \rangle = \langle 61,6562; \infty \rangle$ , tedy nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Nagelkerkův koeficient nabývá hodnoty 0,8208, což je hodnota dosti blízká 1 a svědčí o dobré kvalitě našeho modelu.

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

ID - 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

Hosmerův \_ Lemeshowův test poskytl p-hodnotu 0,8709, tedy na hladině významnosti 0,05 nelze zamítnout hypotézu, že pozorované a modelované hodnoty se neliší.

Klasifikační tabulku najdeme v systému STATISTICA takto: Ve výstupní tabulce GLZ – Výsledky vybereme záložku Rezid. 1 – Klasif & odds ratio.

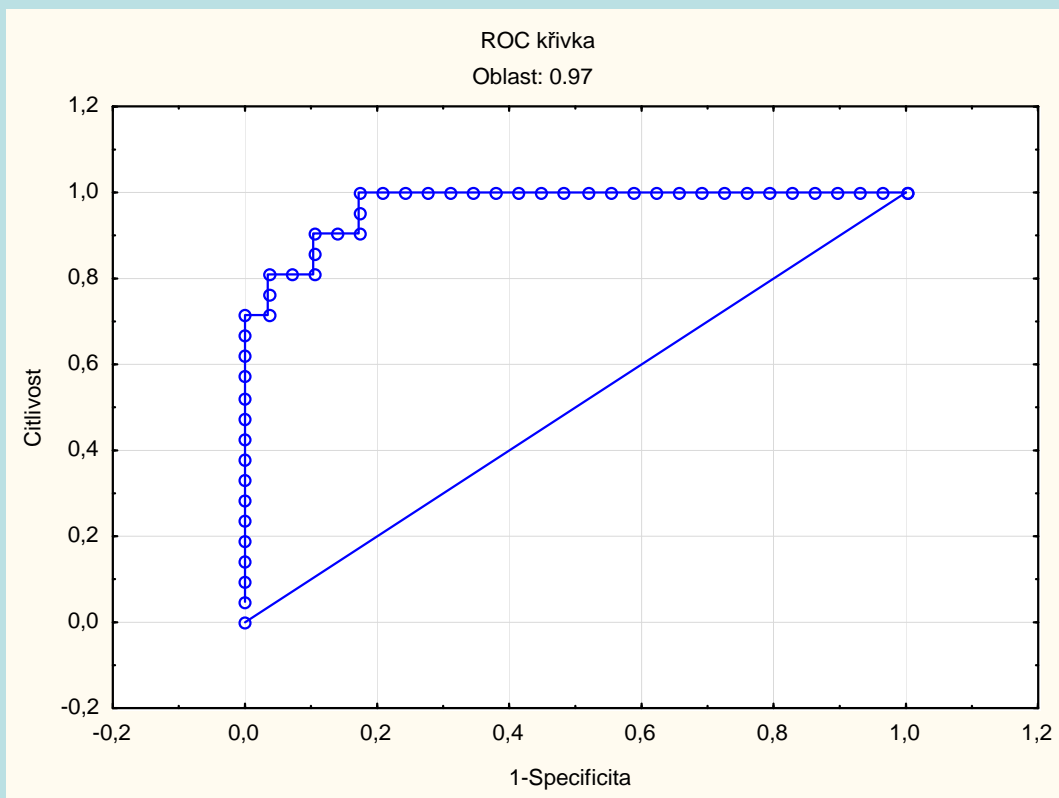
	Klasifikace případů (dovolena.sta)		
	Odds ratio: 52,000000		
	Log odds ratio: 3,951244		
	Předpovězená: návštěva ano	Předpovězená: návštěva ne	Procento správných
Pozorované: návštěva ano	18	3	85,7142857
Pozorované: návštěva ne	3	26	89,6551724

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  $\frac{18 + 26}{50} = 88\%$  .

ROC křivku získáme takto: Ve výstupní tabulce GLZ – Výsledky vybereme záložku Rezid. 1 – ROC křivka



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).