

Errata

Název: Míry diskriminace v kreditním riziku

Autor: Michal Polak

Tento dokument obsahuje opravu k souboru Data.R v příloženém CD a opravu na straně 6 sekce 3.1.

Změna k souboru Data.R v příloženém CD od začátku dokumentu až po řádek

#Vypočet giniho koeficientu:

```
Dataset <- read.table("german.txt",header=FALSE, sep="", na.strings="NA", de
names(Dataset)
attach(Dataset)
#upraveni na 0-1
Dataset$V21 <- abs(Dataset$V21-1)
Dataset$V21 <- abs(Dataset$V21-1)
#prejmenovani
names(Dataset)[c(15)] <- c("Bydleni")
names(Dataset)[c(20)] <- c("Cizinec")
names(Dataset)[c(11)] <- c("DBydleni")
names(Dataset)[c(3)] <- c("Moralka")
names(Dataset)[c(5)] <- c("Objem")
names(Dataset)[c(16)] <- c("PocetUveru")
names(Dataset)[c(8)] <- c("Podil")
names(Dataset)[c(10)] <- c("Ruceni")
names(Dataset)[c(7)] <- c("Souczam")
names(Dataset)[c(21)] <- c("Splatil")
names(Dataset)[c(2)] <- c("Splatnost")
names(Dataset)[c(9)] <- c("Stav")
names(Dataset)[c(12)] <- c("StruktMajetku")
names(Dataset)[c(19)] <- c("Telefon")
names(Dataset)[c(4)] <- c("Ucel")
names(Dataset)[c(1)] <- c("Ucet")
names(Dataset)[c(6)] <- c("Uspory")
names(Dataset)[c(14)] <- c("Uvery")
names(Dataset)[c(13)] <- c("Vek")
names(Dataset)[c(18)] <- c("Vyzivovani")
names(Dataset)[c(17)] <- c("Zamestnani")

#tvorba modelu se vsemi regresory
AllVar <- glm(Splatil ~ Bydleni + Cizinec + DBydleni + Moralka + Objem +
  PocetUveru + Podil + Ruceni + Souczam + Splatnost + Stav + StruktMajetku +
  Telefon + Ucel + Ucet + Uspory + Uvery + Vek + Vyzivovani + Zamestnani,
  family=binomial(logit), data=Dataset)
summary(AllVar)

#vyber regresu pro konecny model
```

```

stepwise(AllVar, direction='backward', criterion='BIC')
stepwise(AllVar, direction='forward', criterion='BIC')
stepwise(AllVar, direction='backward/forward', criterion='BIC')
#vsechny 3 metody vyberou tytez promenne

#tvorba konecneho modelu
MyLogit <- glm(Splatil ~ Moralka +Splatnost +Ucet, family=binomial(logit),
  data=Dataset)
summary(MyLogit)

```

Oprava na straně 6 sekce 3.1, reformulace 2 i 3 odstavce:

Matematicky zápis Označme Y jako závislou binární proměnnou, $\boldsymbol{\beta}$ jako vektor parametrů, které chceme odhadnout a \boldsymbol{x} je vektor regresorů o n složkách. Pak lze model logistické regrese zapsat následujícím způsobem:

$$E(Y | \boldsymbol{x}) = \pi(\boldsymbol{x}), \text{ kde}$$

$$\pi(\boldsymbol{x}) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n\}}$$

znamená pravděpodobnost Y nabytí hodnoty 1 při daných nezávislých proměnných.

Lineární regrese je založená na modelu $y = \boldsymbol{\beta}\boldsymbol{x} + \epsilon$, kde $\boldsymbol{\beta}\boldsymbol{x} = E(Y | \boldsymbol{x})$, y je vysvětlovaná proměnná a ϵ je chyba a vyjadřuje odchylku od podmíněné střední hodnoty. Obvykle se předpokládá, že ϵ má normální rozdělení se střední hodnotou rovnající se nule a nějakým rozptylem, a tedy závislá proměnná má normální rozdělení se střední hodnotou nula a konstantním rozptylem. V modelu logistické regrese, kde odezva podmíněna \boldsymbol{x} se rovná $y = \pi(\boldsymbol{x}) + \epsilon$, nabývá dvě různé hodnoty. Pokud se $y = 1$, potom $\epsilon = 1 - \pi(\boldsymbol{x})$ s pravděpodobností $\pi(\boldsymbol{x})$, pokud $y = 0$, pak $\epsilon = -\pi(\boldsymbol{x})$ s pravděpodobností $1 - \pi(\boldsymbol{x})$. Z výše uvedeného vyplývá, že y má alternativní rozdělení s parametrem $\pi(\boldsymbol{x})$.