3u

Milda Pieškutė

2017 m sausis 17 d

## Užduotis 3 (numatyta 30 taškų)

<http://ekonometrija.lt/blog/tiesin%C4%97-regresija>

Faile 'data\_a.csv' yra pateikti duomenys reikalingi tiesinės regresijos tyrimui. Duomenų masyvo aprašymas pateiktas minėto failo pradžioje.

Duomenų apžvalga ir paruošimas. Šioje dalyje jums reikia susipažinti su duomenimis ir paruošti duomenis tolimesnei analizei.

1. Nuskaitykite duomenų masyvą. Jeigu įmanoma, nekoreguokite originalaus failo. Patariu pasiskaityti ?read.table.

datafull<-read.csv("data\_a.csv", stringsAsFactors = FALSE)

1. Paruoškite duomenis naudojimui. Jūs turite užtikrinti, kad:

Neliko įrašų su trūkstamomis reikšmėmis - jeigu šalinote įrašus, tai būtinai raportuokite kiek įrašų pašalinta ir kiek liko.

Duomenų masyve nėra kažkokių klaidų (pvz. negalimos reikšmės, neteisingas duomenų tipas, t.t.). Patartina patikrinti visų duomenų tipus, nes kartais nuskaitymo procedūra suinterpretuoja duoenis ne taip kaip tikimasi.

sapply(df,class)  
#df$sysTime <- as.Date(df$sysTime , "%m/%d/%y %H:$M")  
sapply(df,class)  
ind<-df$user%in%c("1vbutkus","aaa","testas")  
  
sapply(df, anyNA)  
ind<-apply(df,1, anyNA)  
sum(ind)  
df<-df[!ind,]  
dim(df)

#instaliuojamos analizei reikalingos bibliotekos  
#library(car)  
#library(knitr)   
  
#datafull<-read.csv2("data.csv", header = TRUE)   
head(datafull) #atspausdinami duomenys

## X name takeTime sysUser sysName sysTime user  
## 1 1 abs 5 vygantas Linux 2016-08-19 18:56:37 aaa  
## 2 2 abs 5 vygantas Linux 2016-08-19 18:56:37 aaa  
## 3 3 abs 5 vygantas Linux 2016-08-19 18:56:37 aaa  
## 4 4 aaa 5 vygantas Linux 2016-08-19 18:57:02 aaa  
## 5 5 aaa 5 vygantas Linux 2016-08-19 18:57:02 aaa  
## 6 6 aaa 5 vygantas Linux 2016-08-19 18:57:02 aaa  
## course\_name lesson\_name question\_number correct attempt  
## 1 Google Forms Course Lesson 1 2 TRUE 1  
## 2 Google Forms Course Lesson 1 3 TRUE 1  
## 3 Google Forms Course Lesson 1 4 NA 1  
## 4 Google Forms Course Lesson 1 2 TRUE 1  
## 5 Google Forms Course Lesson 1 3 TRUE 1  
## 6 Google Forms Course Lesson 1 4 NA 1  
## datetime skipped  
## 1 2016-08-19 18:56:21 TRUE  
## 2 2016-08-19 18:56:25 FALSE  
## 3 2016-08-19 18:56:27 NA  
## 4 2016-08-19 18:56:52 FALSE  
## 5 2016-08-19 18:56:53 FALSE  
## 6 2016-08-19 18:56:54 FALSE

ind<-apply(datafull, 1, anyNA)  
datafull<-datafull[!ind,]  
#cor(datafull)

1. Pateikite trumpą kintamųjų apžvalgą. Naudokite tik tuos duomenis (ir kintamuosius), kuriuos planuojate taikyti tyrimui, t.y. jei reikėjo ką nors modifikuoti, tai senųjų reikšmių naudoti nebereikia. Šiai apžvalgai užtenka informacijos, kurią pateikia funkcijos plot, summary ir cor.

#print(cor(datafull), digits = 3)  
#plot(datafull)  
datafull$garsoIzoliacija<-NULL

1. Duomenų masyvą suskaidykite į du atskirus ir nesikertančius masyvus. Didesnį masyvą turi sudaryti apie 80 proc. turimų stebėjimų. Šį masyvą vadinsime trainSet. Mažesnį masyvą turi sudaryti likę įrašai. Šį masyvą vadinsime testSet. Duomenų suskaidymas turi būti atsitiktinis. Gali būti naudingos šios funkcijos:sample, setdiff, %in%.

n <- length(datafull)  
  
ind <- sample(c(TRUE, FALSE), n, replace=TRUE, prob=c(0.8, 0.2))  
trainSet <- datafull[ind, ]  
testSet <- datafull[!ind, ]

Tiesinio modelio sudarymas. Modelio sudarymui naudokite trainSet masyvą.

1. Įvertinkite tiesinį modelių, kuriame prognozuojamas kintamasis būtų kaina. Ji turi tiesiškai priklausyti nuo kitų kintamųjų.

mod1<-lm(kaina~+plotas+aukstas+silumosLaidumas+atstumas, trainSet)  
kable(summary(mod1)$coef, digits=3)

1. Ar visi įtraukti kintamieji daro reikšmingą įtaką prognozei? Jei ne, koreguokite modelį, kad visi kintamieji darytų reikšmingą įtaką. Šį įvertintą modelį žymėsime fit. Užrašykite šio modelio matematinę lygtį.

fit<-update(mod1, kaina~plotas+aukstas+silumosLaidumas)  
kable(summary(mod2)$coef, digits=3)

kaina=plotas+aukstas+silumosLaidumas*. Atitinkamai įverčiai parodo kaip pakitus vienam vienetui iš šių charakteristikų, pakinta kaina:* **kaina=8035.8 + 600.33 *plotas + 3.18.23* aukstas + 528.82 *silumosLaidumas +*e**\* Interpretacija: 1 kv.m padidėjimas padidina kainą 600,33eurais, kiekvienas papildomas aukštas kainą padidina 318,23 eurais, 1 vnt. šilumos laidumo koeficiento padidėjimas padidina kainą 528,82 eurais. Laisvasis narys - 8035,80 eurai gali būti papildomas išlaidos pvz.:notaro paslaugos, buto lokacija, viešojo transporto prieinamumas ir pan.

1. Patikrinkite ar modelio paklaidas galime laikyti normaliomis.

shapiro.test(fit$res)

Pagal shapiro.test p-value>0.05 priimame H0: liekanų paklaidos yra normalios.

Modelio tyrimas

1. Suskaičiuokite modelio fit1 RMSE (Root Mean Squere Error), kuris apibrėžiamas taip

# Function that returns Root Mean Squared Error  
  
rmse <- function(error)  
{  
 sqrt(mean(error^2))  
}  
  
# Calculate error  
#error <- actual - predicted  
pre<-predict(fit,trainSet)  
error1<-trainSet$kaina-pre  
rmse(error1)  
  
  
pre2<-predict(fit,testSet)  
error2<-testSet$kaina-pre2  
rmse(error2)

RMSE suskaičiuokite abiem duomenų masyvams trainSet ir testSet. Palyginkite rezultatus ir padarykite išvadas.

1. Išbrėžkite sklaidos diagramą, kurioje ašyje būtų fit1 modelio testSet prognozė, o ašyje faktinės testSet kainų reikšmės. Ant šios sklaidos diagramos išbrėžkite dvi linijas, kurios žymėtų prognozės pasikliautinus intervalus (pasikliovimo lygmuo 90 proc.).

plot(pre2, testSet$kaina, xlab = "prognoze", ylab = "faktines")  
testSet$kaina  
  
  
#quantile(testSet$kaina, probs = c(0.05, 0.95))  
#a <- predict(fit, interval="confidence")  
  
newx <- seq(min(dt$x), max(dt$x), length.out=100)  
preds <- predict(quadratic.model, newdata = data.frame(x=newx), interval = 'confidence', level = 90)  
  
  
lines(newx, preds[ ,3], lty = 'dashed', col = 'red')  
lines(newx, preds[ ,2], lty = 'dashed', col = 'red')

1. Kiek procentų prognozuotų reikšmių nepateko į pasikliautinus intervalus?