Линеарни статистички модели
припрема 2019.

1. Посматрајмо модел линеарни модел $Y = X\beta+\varepsilon$ са слободном чланом различитим од нуле и дизајн матрицом X $n \times (p+1)$, $\beta = (\beta_0, ..., \beta_p)$ вектор непознатих параметара, и ε има вишедимензиону нормалну $\mathcal{N}_n(0, \sigma^2 I)$ расподелу, при чему је σ^2 такође непознат параметар.

а) Оценити вредност $C \times \beta$, где је $C = (c_0, ..., c_p)$ методом максималне веродостојности. Коју расподелу има та оцена? Одредити 90% интервал поверења за $\beta_0 + \beta_1 + \cdots + \beta_p$.

б) Скуп података `longley` (доступан у R-у) има 7 променљивих: `GNP.deflator`, `GNP`, `Unemployed`, `Armed.Forces`, `Population`, `Year` и `Employed`. Направити линеаран модел који описује зависност `Employed` од осталих променљивих. Одредити 95% интервал поверења за $\beta_{Population} - \beta_{GNP}$ и тестирати хипотезу $H_0 : \beta_{Population} = \beta_{GNP}$.

в) Шта представља мултиколинеарност? Да ли је потребно познавање зависне променљиве за детекцију исте? Одговор образложити. Како се тај проблем превазилази коришћењем назубљене регресије? Од чега зависи оцена непознатих параметара добијена овом методом? Коју расподелу има оцена?

г) Испитати да ли у моделу под б) постоји проблем са мултиколинеарношћу? Ако постоји, решити проблем избацивањем одговарајућих предиктора. Упоредити коефицијенте детерминације почетног и новог модела.

д) Показати да је $\sum_{i=1}^{n} e_i = 0$

ђ) Да ли су резидуали модела независни? А грешке модела ? Одговор образложити.

е) На шта се односи услов хомоскедастичности у линеарном моделу? Како се детектује и како се може решити проблем хетероскедастичности?

ж) Направити линеарни модел који описује зависност променљиве `dist` од променљиве `speed` из скупа података `cars` (уграђен у R). У модел додати и квадратни члан за предиктор `speed`. Испитати графички да ли постоји проблем хетероскедастичности. Уколико постоји проблем, применити Бокс-Кокс трансформацију зависне променљиве. Да ли је проблем ублажен? Који модел (са или без трансформације) боље описује податке?

2. а) Објаснити зашто се регресиона функција зависне променљиве која има Бернулијеву расподелу не може моделирати линеарним моделом. **б)** Навести две линк функције које су погодне за моделирање Бернулијевих зависних променљивих.

б) Написати функцију веродостојности за логистички модел у коме је X дизајн матрица ред $n \times p+1$, а вектор зависних променљивих Y.

в) Која је предност коришћења логит трансформације над онсталим линк функцијама?

г) Скуп података `leukemia` (доступан у пакету `LSMhelp`) садржи једну бинарну променљиву `REMISS` и непрекидне променљиве `CELL`, `SMEAR`, `INFIL`, `LI`, `BLAST` и `TEMP`. Направити модел логистичке регресије којим се описује утицај ових променљивих на променљиву `REMISS`.

д) Написати који сте добили модел уколико се посматрају сми могући предиктори.

ђ) Како изгледа Валдова тест статистика? На основу Валдовог теста, да ли су неки предиктори значајни и који?

е) Тестирати да ли постоји било какав утицај предиктора на зависну променљиву. Одговор образложити.

ж) Да ли постоји значајна разлика између модела који укључује само `LI` као предиктор и полазног модела?

з) Који од ових модела је бољи на основу AIC?

е) Шта представља девијација модела? Да ли се у крајњем моделу који сте добили се девијација може искористити за тестирање да ли је модел адекватан?

```r
##### Zadatak 1.
### 1.b)
model <- lm(Employed ~ ., longley)
summary(model)
# Pravimo interval poverenja za b_pop - b_gnp kao na petom casu
l <- c(0, 0, -1, 0, 0, 1, 0)
X <- model.matrix(model)
XtXi <- solve(t(X)%*%X)
n <- nrow(X)
p <- length(model$coefficients) - 1
estim <- t(l) %*% model$coefficients
stdev <- sqrt(sum(model$res^2)/(n-p-1) * t(l) %*% XtXi %*% l)
lwr <- estim - stdev * qt(0.975, n-p-1)
upr <- estim + stdev * qt(0.975, n-p-1)
c(lwr, upr)

# Ovaj interval ukljucuje nulu, sto znaci da ne mozemo da odbacimo hipotezu da
# je b_pop = b_gnp
# Drugi nacin za testiranje H0: b_pop = b_gnp je koriscenjem linear_hypothesis
linear_hypothesis(model, t(c(0, 0, -1, 0, 0, 1, 0)), 0)
# ili anova...
anova(lm(Employed ~ . -GNP - Population + I(GNP + Population), longley), model)

### 1.g) Multikolinearnost
# Mozemo da gledamo vif
library(car)
vif(model)
# Mnoge vrednosti su velike, dakle postoji prisustvo multikolinearnosti

# Promenljive Unemployed i Armed.Forces imaju vidno manji vif od ostalih, tako
# da cemo ih sacuvati.
# Pogledajmo korelacije ostalih promenljivih
cor(longley[, -c(3, 4, 7)])
# Primecujemo izuzetno velike korelacije, mozemo i nacrtati grafike
plot(longley[, -c(3, 4, 7)])
# Velika linearna zavisnost je prisutna, tako da mozemo ih sve zameniti samo
# jednom od promenljivih, npr. Year.
model2 <- lm(Employed ~ Armed.Forces + Unemployed + Year, longley)
summary(model2)
vif(model2)
# U ovom modelu su vif-ovi svi mali (<30) pa nemamo vise problem
# multikolinearnosti, dok je koeficijent determinacije smanjen sa 0.9955 na
# 0.9928, sto je vrlo mala razlika, pa je dobar smanjeni model.

### 1.zh) Heteroskedasticnost
model <- lm(dist ~ speed + I(speed^2), cars)
# Ispitujemo prisustvo heteroskedasticnosti npr. grafikom zavisnosti apsolutnih
# vrednosti standardizovanih reziduala od ocenjenih vrednosti dist.
plot(abs(rstandard(model)) ~ fitted(model))
# Vidi se da disperzija raste sa porastom dist, tako da imamo problem
# heteroskedasticnosti.

# Trazimo odgovarajucu box-cox transformaciju
library(MASS)
bc <- boxcox(model)
lambda <- bc$x[which.max(bc$y)] # = 0.38
# Pravimo funkciju za transformaciju
BC <- function(y) {
  (y^lambda - 1)/lambda
}
# Fitujemo model sa boxcox transformisanom zavisnom promenljivom
model2 <- lm(BC(dist) ~ speed + I(speed^2), cars)
summary(model2)
plot(abs(rstandard(model2)) ~ fitted(model2))
# Sada je heteroskedasticnost vidno ublazena.

##### Zadatak 2.
library(LSMhelp) # instalacija: devtools::install_github("blaza/lsmhelp")
# ili preuzeti  podatke: https://newonlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/leukemia_remission/index.txt
# Pravimo pocetni model sa svim prediktorima
m <- glm(REMISS ~ ., leukemia, family=binomial)
# U summary vidimo vrednosti Valdovog testa (zvezdice)
summary(m)
# Svi koeficijenti su beznacajni

# Testom kolicnika verodostojnosti proveravamo da li postoji uticaj prediktora
# na zavisnu. Poredimo model koji ukljucuje samo slobodan clan i nas model.
anova(glm(REMISS ~ 1, leukemia, family=binomial), m, test = "Chisq")
# p vrednost < 0.05, pa zakljucujemo da postoji znacajan uticaj

# Poredimo model samo sa LI sa nasim modelom
anova(glm(REMISS ~ LI, leukemia, family=binomial), m, test = "Chisq")
# p vrednost je velika, znaci ne postoji znacajna razlika izmedju dva modela
summary(glm(REMISS ~ LI, leukemia, family=binomial))
# AIC za ovaj model je manji nego za prvobitni, dakle bolji je novi model.
```