Multikolinearita a predikce

3. týden

Multikolinearita

Uvažujte data wages.csv, opět bez odlehlého pozorování.

1. Odhadněte následující modely:

   (A) model pro log mzdu, ve kterém je lineární závislost na vzdělání pro muže a ženy zvlášť a společná závislost na věku,

   (B) model z (A), kde navíc uvažujeme lineárně zkušenost experience.

   Porovnejte koeficienty z obou modelů. Co pozorujeme? Jak byla spočtená proměnná udávající potenciální zkušenost? Jak je možné, že se nám vůbec podařilo odhadnout všechny koeficienty modelu?

2. Spočtěte VIF v modelu (B) pomocí funkce `vif` z knihovny `car`. Následně proveďte totéž pro model (A). Jak vyřešíme problém s multikolinearitou?

Predikce a cross-validace

3. Rozdělte si data na trénovací vzorek o rozsahu $m = 450$ a testovací vzorek o rozsahu $s = n - m$.

```
set.seed(1)
s=sample(1:nrow(data),replace=FALSE)
m=450
train=data[s[1:m],]
test=data[s[-(1:m)],]
```

   Na základě trénovacích dat odhadněte z Vašeho pohledu nejlepší model pro logaritmus mzdy. Spočtěte si vyrovnané hodnoty a predikce pro testovací data.

```
model = lm (log(wage)~.... ,data=train) # ... doplnime nejlepsi model
f=fitted(model)
p=predict(model,newdata=test)
```

   (a) Vykreslete si scatterplot predikovaných a skutečných hodnot logaritmu mzdy. Spočítejte střední kvadratickou chybu

$$MSE = \frac{1}{s}\sum_{i=1}^{s}(\widehat{LW}_i - LW_i)^2$$

   (popřípadě lze uvažovat $RMSE = \sqrt{MSE}$) a střední absolutní chybu

$$MSE = \frac{1}{s}\sum_{i=1}^{s}|\widehat{LW}_i - LW_i|,$$

   kde LW_i je hodnota logaritmu mzdy pro osobu i z testovacího vzorku.

   (b) Porovnejte vypočtené MSE a MAE s MSE a MAE spočtené pro trénovací data.

(c) Uvažujte nějaký horší model, např. model, kde log mzdy závisí pouze na vzdělání (bez rozlišení pohlaví a bez uvažování dalších proměnných). Dostáváme horší predikce?

4. Nyní nás bude zajímat predikce přímo mzdy. Uvažujte následující možnosti:

$$\widehat{W}_i = e^{\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}}, \quad \widetilde{W}_i = e^{\frac{1}{2}\hat{\sigma}^2} e^{\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}}, \quad \breve{W}_i = \hat{\alpha} e^{\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}}$$

kde $\boldsymbol{x}_i$ je vektor regresorů pro i-tou osobu z testovacího vzorku, $\widehat{\boldsymbol{\beta}}$ je odhad parametrů Vašeho nejlepšího modelu a $\hat{\alpha}$ odhadneme z modelu

$$W_j = \alpha e^{\widehat{LW}_j},$$

kde W_j je hodnota mzdy pro osobu j z trénovacího vzorku, $j = 1, \ldots, m$, a $\widehat{LW}_i$ je odpovídající vyrovnaná hodnota logaritmu mzdy.

Porovnejte MAE a MSE pro tyto tři různé predikce. Který přístup vychází jako nejlepší a který jako nejhorší?

5. Co by se dělo, kdybychom volili m jinak? Třeba 530 nebo naopak 30?

6. Leave One Out cross-validace: Model odhadneme z $n-1$ pozorování a na základě něj predikujeme hodnotu pro vynechané pozorování. Provedeme postupně pro všechna pozorování $i = 1, \ldots, n$ a spočteme MSE a MAE.

```
n=nrow(data)
y.hat=numeric(n)
for(i in 1:n){
  m=lm(log(wage)~...,data=data[-i,])   # doplnit model za ....
  y.hat[i]=predict(m,newdata=data[i,])
}
mean(abs(log(data$wage)-y.hat))
mean((log(data$wage)-y.hat)^2)
```

V R existuje knihovna `caret`, která cross-validaci provede za nás.

```
install.packages("caret")
library(caret)

train_control <- trainControl(method = "LOOCV")
CV1 <- train(log(wage) ~..., data = data,  # doplnit model za ....
               method = "lm",
               trControl = train_control)
print(CV1)

CV1$finalModel
CV1$resample
```

7. K-fold cross-validace: Soubor rozdělíme náhodně na K podsouborů podobné velikosti. Postupně vždy jeden podsoubor použijeme jako testovací a ostatní data jako trénovací. MSE a MAE spočítáme jako průměr dílčích MSE a MAE.

```
train_control <- trainControl(method = "cv", number = 5) ## 5-fold CV
CV2 <- train(log(wage) ~..., data = data,   # doplnit model za ....
```

```
                  method = "lm",
                  trControl = train_control)

print(CV2)
```

Pomocí předchozích metod lze porovnávat různé modely a přístupy z hlediska predikce.

Odlehlost v mnohorozměrných datech, metoda hlavních komponent

Uvažujme data `Guns` z knihovny `AER`. Vybereme si pouze rok 1999.

```
data(Guns,package="AER")
data=Guns[Guns$year==1999,]
```

Incidence kriminality (violent, murder, robery) je závisle proměnná, ostatní sloupce tvoří potenciálně vysvětlující proměnné.

8. Zkuste detekovat odlehlá pozorování (z hlediska rozdělení regresorů) pomocí 2D scatter-plotů.

   ```
   reg=data[,c(5:11)]
   rownames(reg)=data$state
   plot(reg)
   ```

   Spočtěte si i výběrové korelační koeficienty. Které proměnné spolu hodně souvisí?

9. Mnohorozměrnou vizualizaci lze provést i pomocí tzv. Chernoffových tváří.

   ```
   library(aplpack)
   faces(reg[,-8],labels=data$state,cex=0.5)
   ```

10. Metoda hlanvích komponent vybírá lineární kombinaci, která vysvětlí nejvíce variability.

    ```
    p.c=prcomp(reg,scale=TRUE)
    summary(p.c)
    # sqrt(eigen(cor(reg))$values)
    plot(p.c)
    p.c$rotation
    biplot(p.c)
    biplot(p.c,choices=c(1,3))
    biplot(p.c,choices=2:3)

    model=lm(data$violent~p.c$x[,1]+p.c$x[,2]+p.c$x[,3])
    summary(model)
    ```