Prosta linearna regresija

Blagoje Ivanović

October 8, 2018

Prosta regresija

- U slučaju proste regresije, tražimo neku funkcionalnu vezu izmedju obeležja Y, koje nazivamo zavisnom promenljivom i obeležja X koje nazivamo nezavisnom promenljivom ili prediktorom.
- Ova veza je oblika
$$Y = f(X) + \varepsilon,$$
gde je ε slučajna greška modela, koja je nezavisna od X.
- Prava funkcija $f(x)$ nam je nepoznata i treba da je ocenimo na osnovu uzorka $(x_1, y_1), \ldots, (x_n, y_n)$.
- Pretpostavljamo da za uzorak važi
$$y_i = f(x_i) + \varepsilon_i,$$
i potrebna nam je ocena $\hat{f}(x)$.

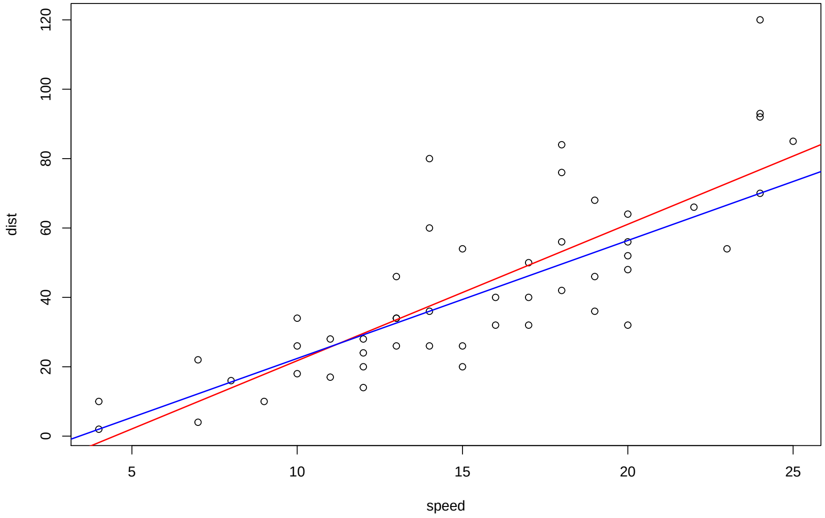
Linearna regresija

- Prostor svih mogućih funkcija $f(x)$ je prevelik i nije moguće istražiti ga u potpunosti. Zato uvek pretpostavljamo odredjen (obično parametarski) oblik funkcije $f(x)$, čime smanjujemo prostor funkcija nad kojim tražimo $f(x)$.
- U linearnoj regresiji, pretpostavljamo da je $f(x) = \beta_0 + \beta_1 x$, odnosno imamo model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- Ostaje zadatak odrediti parametre modela β_0 i β_1 koji najbolje opisuju uzorak.

Koja je prava bolja?



Ocena metodom najmanjih kvadrata

- Da bismo dobili što bolji model, cilj je odabrati parametre tako da je neka vrsta greške modela najmanja. Mi ćemo se truditi da minimizujemo srednjekvadratnu grešku.
- Kako je model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, želimo da minimizujemo

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

- Minimizacijom tog izraza dobijamo ocene metodom najmanjih kvadrata $\hat{\beta}_0$ i $\hat{\beta}_1$.
- Kod proste linearne regresije one su jednake

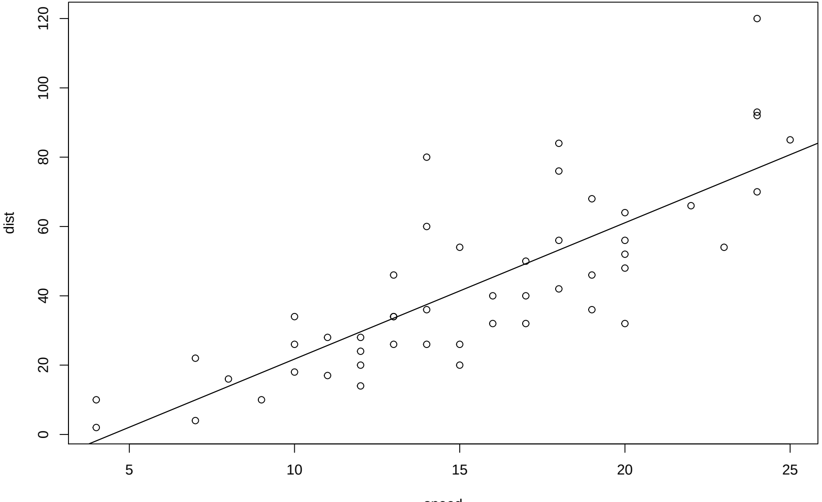
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

Primer ocene MNK

```r
plot(cars) # nacrtamo skup podataka
model <- lm(dist ~ speed, cars) # odredimo linearni model, tj. koeficijente MNK
model$coef
```

```
## (Intercept)       speed
##  -17.579095    3.932409
```

```r
abline(-17.579, 3.9324) # nacrtamo pravu odredjenu modelom
abline(model) # jednostavnije crtanje prave
```



Reziduali

- Kada imamo ocene $\hat{\beta}_0$ i $\hat{\beta}_1$, ocenjene vrednosti modela, $\hat{y}_i$, za prediktore x_i su
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$
- Odstupanja ocenjene vrednosti od stvarnih vrednosti y_i nazivamo rezidualima:
$$e_i = y_i - \hat{y}_i.$$
- Korisna mera odstupanja ocenjenog modela od stvarnog je suma kvadrada reziduala - SSE:
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Koeficijent determinacije - R^2

- Jedna mera koja govori koliko linearni model dobro predvidja je koeficijent determinacije

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

- R^2 nam govori koliko je dobijeni linearni model bolji u predvidjanju y_i u odnosu na najjednostavniji model koji prosto y_i ocenjuje sa prosekom $\bar{y}$.
- Oznaka R^2 potiče od toga što je on jednak kvadratu korelacije izmedju y i $\hat{y}$.
- Kod proste regresije je $R^2 = \rho(x, y)$.

Primer R^2

```r
#plot(Petal.Length ~ Petal.Width, data = iris)
model <- lm(Petal.Length ~ Petal.Width, data = iris)
#abline(model)
summary(model)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
## Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

```r
cor(iris$Petal.Length, model$fitted.values)^2 # == R^2
```

```
## [1] 0.9271098
```

```r
# cor(iris$Petal.Length, predict(model, iris))^2 # isto

cor(iris$Petal.Length, iris$Petal.Width)^2 # == R^2
```

```
## [1] 0.9271098
```

Uslovi Gaus-Markova

- Da bi ocene koeficijenata modela metodom najmanjih kvadrata bile dobre, od grešaka ε_i modela zahtevamo da ispunjavaju uslove:
  - $E(\varepsilon_i) = 0, \forall i$,
  - $D(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2, \forall i$ - homoskedastičnost
  - $E(\varepsilon_i \varepsilon_j) = 0, \forall i \neq j$ - nekoreliranost
- Pod ovim uslovima, nepristrasna ocena za σ^2 je

$$\hat{\sigma}^2 = \frac{SSE}{n-2}.$$

Primer za $\hat{\sigma}^2$

```r
#plot(mpg ~ wt, data = mtcars)
model <- lm(mpg ~ wt, data = mtcars)
#abline(model)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```r
n <- length(mtcars$wt)
sqrt(sum(model$residuals^2) / (n - 2)) # == Residual standard error
```

```
## [1] 3.045882
```

Svojstva ocena pri G-M uslovima

- Očekivanje i disperzija ocena $\hat{\beta}_0$ i $\hat{\beta}_1$, pod uslovima Gaus-Markova su

$$
\begin{aligned}
E(\hat{\beta}_0) &= \beta_0 \\
D(\hat{\beta}_0) &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \\
E(\hat{\beta}_1) &= \beta_1 \\
D(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

- Ocene disperzija dobijamo uzimanjem umesto σ^2 ocenu $\hat{\sigma}^2$.

Primer za standardne greške koeficijenata

```
#plot(mpg ~ wt, data = mtcars)
model <- lm(mpg ~ wt, data = mtcars)
#abline(model)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
n <- length(mtcars$wt)

s2 <- sum(model$residuals^2) / (n - 2) # ocena sigma^2

sqrt(s2*(1/n + (mean(mtcars$wt)^2)/(var(mtcars$wt)*(n-1)))) #ocena std. greske beta_0
```

```
## [1] 1.877627
```

```
sqrt(s2/(var(mtcars$wt)*(n-1))) # ocena std. greske beta_1
```

```
## [1] 0.559101
```

Normalno raspodeljene greške

- Pretpostavimo, pored G-M uslova, da su greške i normalno raspodeljene, tj. da važi

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- Tada je $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Pokazuje se da je

$$(\hat{\beta}_j - \beta_j)/s.e.(\hat{\beta}_j) \sim t_{n-2}, \ j = 0, 1,$$

gde je $s.e.(\hat{\beta}_j)$ standardna greška ocene parametra $\hat{\beta}_j$.

Testiranje statističke značajnosti koeficijenta

- Zanima nas da li je neki od koeficijenata β_j statistički neznačajan, tj. da li može da se izbaci iz modela, a da se ne izgubi na kvalitetu ocena.
- To proveravamo testiranjem hipoteze
$$H_0 : \beta_j = 0.$$
- Na osnovu prethodno rečenog, pod H_0 važi
$$t(\hat{\beta}_j) = \hat{\beta}_j / s.e.(\hat{\beta}_j) \sim t_{n-2}$$
- Zaključak o statističkoj značajnosti donosimo proveravanjem p-vrednosti testa
$$P(|t| > |t(\hat{\beta}_j)|).$$
Ako je ona izrazito mala, odbacujemo H_0 i zaključujemo da je koeficijent značajan.

Primer za testiranje hipoteza

```
#plot(mpg ~ wt, data = mtcars)
model <- lm(mpg ~ wt, data = mtcars)
#abline(model)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
n <- length(mtcars$wt)
s2 <- sum(model$residuals^2) / (n - 2) # ocena sigma^2
se1 <- sqrt(s2/(var(mtcars$wt)*(n-1))) # ocena std. greske beta_1
t_stat <- unname(model$coef[2]) / se1
p_val <- 2*(1-pt(abs(t_stat), df=n-2))
t_stat
```

```
## [1] -9.559044
```

```
p_val
```

```
## [1] 1.293958e-10
```