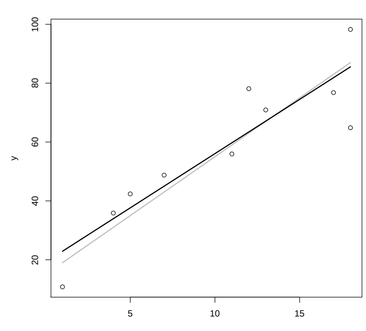
1 Linearna regresija

Zanima nas povezanost števila ur učenja na teden z rezultatom na izpitu iz statistike. Vzemimo, da vemo, da se rezultat na izpitu v populaciji porazdeljuje pogojno normalno: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

- Naj bo X enakomerno porazdeljena spremenljivka (med 0 in 20, zaokrožena navzdol), $\beta_0 = 15$, $\beta_1 = 4$, $\sigma = 10$. Generirajte vzorec velikosti 10, narišite podatke in vrišite populacijsko ter ocenjeno vrednost premice.

```
> set.seed(1)
> n <- 10                                 #velikost vzorca
> beta0 <- 15
> beta1 <- 4
> sigma <- 10
> x <- floor(runif(n)*20)                 #navzdol zaokrozene vrednosti x
> x <- sort(x)                            #uredimo podatke po velikosti x
> y <- rnorm(n,mean=beta0+beta1*x,sd=sigma)  #generiramo iz normalne porazdelitve
> plot(x,y)                               #narisemo tocke
> popul <- beta0 + beta1*x                #populacijska vrednost premice
> lines(x,popul,col="grey",lwd=2)         #dodamo populacijsko vrednost premice v sivi barvi
> fit <- lm(y~x)                          #ocenimo premico na podatkih
> summary(fit)                            #ogledamo si ocene koeficientov
> beta0h <- fit$coef[1]                  #ocenjena beta0
> beta1h <- fit$coef[2]                  #ocenjena beta1
> napoved <- beta0h + beta1h*x
> lines(x,napoved,lwd=2)                       #vrisemo ocenjeno premico na sliko
```



Slika 1: *Točke na vzorcu, ocenjena premica (črna) in populacijska premica (siva).*

- Iz spodnjega izpisa preberite ocene populacijskih parametrov. Interpretirajte rezultate, katere ničelne domneve so testirane in kako?

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-20.683  -4.746   2.844   4.512  14.693

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2049     7.5172   2.555 0.033921 *
x             3.6850     0.6217   5.927 0.000351 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.44 on 8 degrees of freedom
Multiple R-squared: 0.8145,     Adjusted R-squared: 0.7913
F-statistic: 35.13 on 1 and 8 DF,  p-value: 0.0003508
```

Ocene parametrov so $\widehat{\beta}_0 = 19{,}2$, $\widehat{\beta}_1 = 3{,}7$, $\widehat{\sigma} = 11{,}4$. Testirani sta dve ničelni domnevi: $H_{0int} : \beta_0 = 0$ in $H_0 : \beta_1 = 0$. Pri linearni regresiji nas ponavadi zanima le druga - saj ta govori o povezanosti med spremenljivkama v populaciji. Metoda največjega verjetja nam pove, da se ocene parametrov okrog prave vrednosti porazdeljujejo približno normalno (za dovolj velik n). Standardna napaka je ocenjena iz podatkov, uporabimo test t:

$$T = \frac{\widehat{\beta}_1}{\widehat{SE}_{\beta_1}} = \frac{3{,}7}{0{,}6} = 5{,}9$$

Vemo, da je slučajna spremenljivka T porazdeljena približno kot t z 8 stopinjami prostosti (pri ocenjevanju SE porabimo dve stopinji prostosti). Ta test se imenuje Waldov test.

- Kako bi ničelno domnevo $H_0 : \beta_1 = 0$ preverili s testom razmerja verjetij?
  Uporabite rezultat, da ocena $\widehat{\sigma}$ po metodi največjega verjetja ni nepristranska in je enaka

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1)^2}{n}$$

Izračunati moramo vrednost maksimuma funkcije verjetja pod ničelno in alternativno domnevo. Funkcija verjetja je enaka:

$$l(y,x,\beta_0,\beta_1,\sigma) \quad = \quad \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\{-\frac{\sum_{i=1}^n(y_i-\beta_0-\beta_1x_i)^2}{2\sigma^2}\}$$

Maksimum funkcije verjetja je enak

$$\begin{aligned} l(y,x,\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}) &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}{2\widehat{\sigma}^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{n\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}{2\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n}\exp\left\{-\frac{n}{2}\right\} \end{aligned}$$

Logaritem funkcije verjetja v ocenjenih vrednostih je zato enak

$$\begin{aligned} &\log l(y,x,\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}) = \\ &= -\frac{n}{2}\left(\log(2\pi)-\log[\sum_{i=1}^n(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2]+\log n-1\right) \end{aligned}$$

Vrednost maksimuma pod alternativno domnevo izračunamo tako, da vstavimo ocenjene $\widehat{\beta}_0$ in $\widehat{\beta}_1$, za izračun vrednosti pod ničelno domnevo moramo oceniti še β_0 v ničelnem modelu. Dobljeni Wilksov Λ se porazdeljuje kot χ^2_1.

```
> fit0 <- lm(y~1)                      #ocenimo premico pod nicelno domnevo - le konstanta
> res0 <- y - fit0$coef                #ostanki pod nicelno domnevo
> resA <- y - beta0h - beta1h*x        #ostanki pod alternativno domnevo
> logl0 <- .5*n*(-log(2*pi) - log(sum(res0^2))+log(n) - 1 )   #loglik pod nicelno
> loglA <- .5*n*(-log(2*pi) - log(sum(resA^2))+log(n) - 1 )   #loglik pod alternativno
> Lambda <- 2*(loglA-logl0)            #Wilksov lambda
> 1-pchisq(Lambda,1)                   #likelihood ratio test
[1] 4.048e-05
```

2 Matrično računanje

Vrednosti neodvisnih spremenljivk združimo v matriko X (design matrix), vrednosti odvisne spremenljivke ter koeficientov predstavljajo vektorja Y in β:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix};\ Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix};\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrika X je dimenzije $n \times (p+1)$, kjer je p število spremenljivk. Če naš model ne bi vseboval konstante, bi prvi stolpec X izpustili. V našem primeru imamo le eno neodvisno spremenljivko, matrika X je enaka:

```
> X <- cbind(1,x)                          #zlepimo dva stolpca
> X
          x
 [1,] 1  1
 [2,] 1  3
 [3,] 1  5
 [4,] 1  7
 [5,] 1  7
 [6,] 1  8
 [7,] 1 11
 [8,] 1 16
 [9,] 1 19
[10,] 1 19
> round(y)            #zaokrozimo za vecjo preglednost
 [1] 11 36 42 49 56 78 71 77 65 98
```

- Zapišite vsoto vrednosti $\sum_{i=1}^{n} Y_i^2$ v matrični obliki.

$$Y^T Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = Y_1^2 + Y_2^2 + \ldots + Y_n^2$$

- Kaj dobimo, če matrično pomnožimo $X\beta$?

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = E(Y)$$

- V matrični obliki oceno koeficientov po metodi najmanjših kvadratov (= po metodi največjega verjetja) zapišemo kot $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Pokažite, da za $p = 1$ dobite oceni:

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Izračunajmo najprej $X^T X$:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Inverz te 2×2 matrike je enak:

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

Izračunajmo še $X^T Y$:

$$X^T Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix}$$

Velja torej

$$
\begin{aligned}
(X^TX)^{-1}X^TY &= \frac{1}{n\sum\limits_{i=1}^n x_i^2 - (\sum\limits_{i=1}^n x_i)^2}\begin{bmatrix} \sum\limits_{i=1}^n x_i^2 & -\sum\limits_{i=1}^n x_i \\ -\sum\limits_{i=1}^n x_i & n \end{bmatrix}\begin{bmatrix} \sum\limits_{i=1}^n Y_i \\ \sum\limits_{i=1}^n x_iY_i \end{bmatrix} \\
&= \frac{1}{n\sum\limits_{i=1}^n x_i^2 - (\sum\limits_{i=1}^n x_i)^2}\begin{bmatrix} \sum\limits_{i=1}^n x_i^2 \sum\limits_{i=1}^n Y_i - \sum\limits_{i=1}^n x_i \sum\limits_{i=1}^n x_iY_i \\ -\sum\limits_{i=1}^n x_i \sum\limits_{i=1}^n Y_i + n\sum\limits_{i=1}^n x_iY_i \end{bmatrix}
\end{aligned}
$$

- Izpeljite oceno po metodi najmanjših kvadratov še v matrični obliki. Pri tem boste potrebovali naslednje formule za matrično računanje:

$$(A^T)^T = A;\ (AB)^T = B^TA^T;\ \frac{\partial \beta^T A}{\partial \beta} = A;\ \frac{\partial \beta^T A^T A\beta}{\partial \beta} = 2A^TA\beta$$

Namig: Kaj minimiziramo? Kako zapišemo vsoto kvadriranih ostankov v matrični obliki?

Iščemo minimum funkcije

$$
\begin{aligned}
(Y - X\beta)^T(Y - X\beta) &= (Y^T - \beta^TX^T)(Y - X\beta) \\
&= Y^TY - \beta^TX^TY - Y^TX\beta + \beta^TX^TX\beta \\
&= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta,
\end{aligned}
$$

Pri čemer smo v zadnji vrstici uporabili, da je $(\beta^TX^TY)^T = \beta^TX^TY$, saj je matrika dimenzije 1×1. Sedaj odvajamo po β

$$\frac{\partial}{\partial \beta}(Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta) = -2X^TY + 2X^TX\beta$$

in izenačimo z 0:

$$
\begin{aligned}
-2X^TY + 2X^TX\widehat{\beta} &= 0 \\
X^TX\widehat{\beta} &= X^TY \\
\widehat{\beta} &= (X^TX)^{-1}X^TY
\end{aligned}
$$

- Pokažite, da je ocena koeficientov nepristranska (vzemite, da so vrednosti x-ov dane in torej ne slučajne).

$$E(\widehat{\beta}) = (X^TX)^{-1}X^TE(Y) \;=\; (X^TX)^{-1}X^TX\beta = \beta$$

- Izpeljite formulo za standardno napako ocenjenih koeficientov v matrični obliki. Intuitivno razložite od česa je odvisna standardna napaka koeficienta β_1 (za $p = 1$).
  Uporabite, da velja $var(cY) = c\,varY c^T$.

$$\begin{aligned} var(\widehat{\beta}) &= var[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T varY[(X^TX)^{-1}X^T]^T \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \end{aligned}$$

Standardna napaka SE_{β_1} je torej enaka

$$\begin{aligned} SE_{\beta_1} &= \sqrt{\frac{\sigma^2}{(X^TX)^{-1}_{22}}} \\ &= \sqrt{\frac{n\sigma^2}{n\sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2}} \\ &= \sqrt{\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= \sqrt{\frac{\sigma^2}{n\sigma_x^2}} = \frac{\sigma}{\sigma_x\sqrt{n}} \end{aligned}$$

Standardna napaka koeficienta je tako kot vedno odvisna od velikosti vzorca n ter razpršenosti podatkov. Vrednost σ je standardna napaka ostankov okrog premice - večja kot je, bolj se lahko zmotimo pri oceni premice. Vendar tu ni pomembna absolutna velikost variance ostankov, zanima nas variabilnost ostankov glede na variabilnost neodvisne spremenljivke. Če je razpon x-ov majhen, je naša ocena pri isti variabilnosti ostankov manj natančna. Razložimo to na našem primeru - če bi v vzorec zajeli le posameznike, ki so se učili 3-5 ur, bi bila povezanost med spremenljivkama (npr. merjena s korelacijskim koeficientom) pri istih regresijskih dosti manjša in zato možna večja odstopanja pri ocenjevanju.

- Kako bi izračunali interval zaupanja za napovedano premico v našem primeru ($p = 1$)? Dodajte ga na sliko

  Izračunamo standardno napako za vsako točko posebej.

  $$var(\widehat{y_i}) = var(\widehat{\beta}_0 + \widehat{\beta}_0 x_i) = var(\widehat{\beta}_0) + x_i^2 var(\widehat{\beta}_1) + 2x_i cov(\widehat{\beta}_0, \widehat{\beta}_1)$$

  Matrični izračun (za vse točke naenkrat):

  $$var(\widehat{y_i}) = var(X\widehat{\beta}) = Xvar(\widehat{\beta})X^T = \sigma^2 X(X^TX)^{-1}X^T$$

  Izračun v R-u za naš primer:

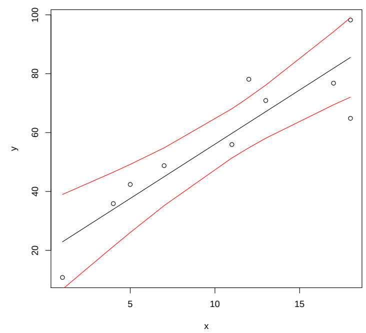
```
> X <- cbind(1,x)
> sigma <- summary(fit)$sigma
> inv <- solve(t(X)%*%X)
> mat <-  X%*%inv%*%t(X)
> se <- sigma*sqrt(diag(mat))
> betah <- c(beta0h,beta1h)
> plot(x,y)
> lines(x,X%*%betah)
> t8 <- qt(.975,8)
> lines(x,X%*%betah - t8*se,col=2)
> lines(x,X%*%betah + t8*se,col=2)
```

- Recimo, da nas zanima, kako sta število ur učenja in spol (0=ženski, 1=moški) povezana z rezultatom na izpitu iz statistike. Predpostavimo model, ki vključuje interakcijo. Kako bi preverili, ali je število ur učenja pri moških povezano z rezultatom na izpitu?

  Model, ki ga prilagodimo podatkom, zapišemo kot:

  $$Y = \beta_0 + \beta_1 \text{spol} + \beta_2 \text{ure} + \beta_3(\text{ure} * \text{spol})$$

  Ničelna domneva, ki jo želimo preveriti, je torej $H_0 : \beta_1 + \beta_3 = 0$. Zapišemo v matrični obliki. Naj bo vektor $a^T = [0,0,1,1]$, zanima nas $H_0 : a^T\beta = 0$. Varianca $a^T\beta$ je enaka $var(a^T\beta) = a^T var\beta a$, za preverjanje ničelne domneve uporabimo test t.



Slika 2: *Točke na vzorcu in ocenjena premica z intervalom zaupanja.*

3 Predpostavke linearne regresije

Z osnovnim modelom linearne regresije naredimo štiri predpostavke:

- Ostanki so okrog premice porazdeljeni normalno
- Varianca ostankov ni odvisna od vrednosti neodvisne spremenljivke (homoskedastičnost)
- Ostanki so med seboj neodvisni.
- Povezanost med X in Y je linearna

Kaj se zgodi z ocenami koeficientov, njihovo pričakovano vrednostjo, standardno napako in z intervali zaupanja, če je katera izmed prvih treh predpostavk kršena?

- Kaj se spremeni v izpeljavah, če ostanki okrog premice niso porazdeljeni normalno?

  Najprej vzemimo, da ostanki okrog premice niso porazdeljeni normalno. V tem primeru ocena koeficientov po metodi največjega verjetja ni enaka oceni po metodi najmanjših kvadratov. Ocena po metodi najmanjših kvadratov bo identična kot do sedaj, enaka bo tudi ocena standardne napake. Prav tako bo ocena po metodi najmanjših kvadratov nepristranska ocena populacijskih vrednosti. Ne moremo pa o populacijskih vrednostih sklepati ničesar več, saj ne poznamo porazdelitve ocene okrog prave vrednosti.

- Recimo, da je varianca ostankov odvisna od x.

  Če varianca ostankov ni enaka za vsak x, moramo varianco pisati v matrični obliki, npr.

$$\Sigma = \sigma \begin{bmatrix} w_1 & 0 & \dots & 0 & 0 \\ 0 & w_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & w_n \end{bmatrix}$$

  Varianca sicer ne vpliva na oceno koeficientov po metodi najmanjših kvadratov, vendar pa se spremeni ocena po metodi največjega verjetja, saj moramo maksimizirati funkcijo
  $-2(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)$:

$$\begin{aligned} & (Y - X\beta)^T\Sigma^{-1}(Y - X\beta) = \\ = & \ (Y^T - \beta^T X^T)\Sigma^{-1}(Y - X\beta) \\ = & \ Y^T\Sigma^{-1}Y - \beta^T X^T\Sigma^{-1}Y - Y^T\Sigma^{-1}X\beta + \beta^T X^T\Sigma^{-1}X\beta \\ = & \ Y^T\Sigma^{-1}Y - 2\beta^T X^T\Sigma^{-1}Y + \beta^T X^T\Sigma^{-1}X\beta \end{aligned}$$

  in zato

$$\begin{aligned} -2X^T\Sigma^{-1}Y + 2X^T\Sigma^{-1}X\widehat{\beta} &= 0 \\ X^T\Sigma^{-1}X\widehat{\beta} &= X^T\Sigma^{-1}Y \\ \widehat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \end{aligned}$$

  Ustrezno se spremeni tudi varianca ocene:

$$\begin{aligned} var(\widehat{\beta}) &= var[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y] \\ &= (X^T\Sigma^{-1}X)^{-1}X^T varY[(X^T\Sigma^{-1}X)^{-1}X^T]^T \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1} \end{aligned}$$

Če so vrednosti w_i znane, se v ocenjevanje koeficientov in standardne napake torej le vrine diagonalna matrika. Statistično sklepanje je enako kot do sedaj.

- Kaj pa če ostanki med seboj niso neodvisni?

  Potem variančna matrika Σ ni več diagonalna (je npr. bločno diagonalna). Rezultati bodo podobni tistim v prejšnji točki, bo pa seveda ocenjevanje odvisno od tega, kaj vemo o elementih Σ.