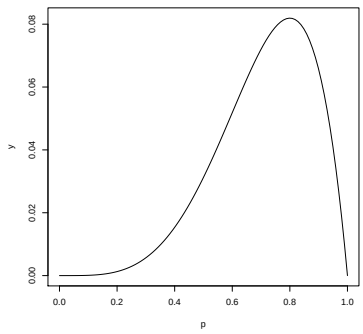
3 Ocenjevanje parametrov - metoda največjega verjetja

3.1 Ocenjevanje deleža

Naj bodo $x_1, \dots x_n$ neodvisne realizacije Bernoullijevo porazdeljene slučajne spremenljivke X. Radi bi ocenili parameter p.

- Recimo, da je $n = 5$ in da smo dobili naslednjih 5 vrednosti: 1,0,1,1,1. Kakšna bi bila verjetnost tega dogodka, če bi bil $p = 0{,}2$? Kaj pa za $p = 0{,}75$? Narišite krivuljo verjetnosti tega dogodka glede na p. Kako bi izračunali njen vrh?

  Verjetnost dogodka izračunamo kot $0{,}2^4 0{,}8^1$, torej $p^k(1-p)^{n-k}$, kjer je k število enk. Označimo z A dogodek $A = \{X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1\}$. Za $p = 0{,}2$ dobimo $P(A) = 0{,}00128$, za $p = 0{,}75$ dobimo $P(A) = 0{,}079$. Narišemo krivuljo za vrednosti p med 0 in 1:



Slika 1: *Verjetnost opaženega dogodka glede na p.*

  Vrh funkcije lahko poiščemo z odvajanjem - odvajamo funkcijo $p^k(1-p)^{n-k}$ po p in izenačimo z 0 (lokalni maksimum). Vrh ni odvisen od vrstnih redov.
  V našem primeru je vrh funkcije dosežen pri $p = 4/5$.

- Podatke, ki jih dobimo na nekem vzorcu, označimo z $x_1, \dots, x_n$ (v zgornjem primeru je bil $n = 5$, $x_1 = 1$ in $x_2 = 0$). Za vsako enoto zapišite

$P(X_i = x_i|p)$, torej verjetnost, da se je zgodil dogodek, ki smo ga videli. Zapišite funkcijo verjetja.

$$P(X_i = x_i|p) = p^{x_i}(1-p)^{1-x_i}$$

Funkcija verjetja je produkt posameznih verjetnosti (predpostavili smo, da so slučajne spremenljivke X_i neodvisne), torej

$$\begin{aligned} L(p,x) = P(X_1 = x_1, \ldots, X_n = x_n|p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

- Poišcite oceno za p po metodi največjega verjetja

  Ker je logaritem monotona funkcija, lahko namesto lokalnega maksimuma te funkcije gledamo raje maksimum logaritma:

$$\begin{aligned} \log L(p,x) &= \sum_{i=1}^{n} x_i \log(p) + (n - \sum_{i=1}^{n} x_i)\log(1-p) \\ \frac{\partial \log L(p,x)}{\partial p} &= \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \\ &= \frac{\sum_{i=1}^n x_i - p\sum_{i=1}^n x_i - p(n - \sum_{i=1}^n x_i)}{p(1-p)} \\ &= \frac{\sum_{i=1}^n x_i - pn}{p(1-p)} \end{aligned}$$

  Odvod logaritma verjetja bo enak 0 pri $\hat{p}n = \sum_{i=1}^n x_i$. Ocena po metodi največjega verjetja je torej $\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i$. Ocena je ravno delež enk v vzorcu.

- Ali je ocena nepristranska?

  Metoda največjega verjetja zagotavlja le doslednost (nepristranost, ko gre $n \to \infty$), v našemo primeru dobimo

$$E(\hat{p}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} p = p$$

V našem primeru je torej ocena nepristranska.

- Zapišite oceno standardne napake

Varianca ocene je enaka $\frac{1}{n} I(p)^{-1}$, kjer je

$$I(p) = -E\left[\frac{\partial^2}{\partial p^2}\log(f(X,p))\right] = E\left[\frac{\partial}{\partial p}\log(f(X,p))\right]^2$$

V našem primeru sta izračuna po obeh formulah enako težka, uporabimo prvo formulo:

$$\begin{aligned}
f(X|p) &= p^X(1-p)^{1-X} \\
I(p) &= -E\left[\frac{\partial^2}{\partial p^2}\log(f(X|p))\right] \\
&= -E\left[\frac{\partial^2}{\partial p^2}\left(X\log p + (1-X)\log(1-p)\right)\right] \\
&= -E\left[\frac{\partial}{\partial p}\left(\frac{X}{p} - \frac{1-X}{1-p}\right)\right] \\
&= -E\left[\frac{\partial}{\partial p}\left(\frac{(1-p)X - (1-X)p}{p(1-p)}\right)\right] \\
&= -E\left[\frac{\partial}{\partial p}\left(\frac{X-p}{p(1-p)}\right)\right] \\
&= -E\left[\frac{p(1-p)(-1) - (1-2p)(X-p)}{p^2(1-p)^2}\right] \\
&= -E\left[\frac{-p+p^2-X+2pX+p-2p^2}{p^2(1-p)^2}\right] \\
&= -E\left[\frac{-p^2-X+2pX}{p^2(1-p)^2}\right]
\end{aligned}$$

Pri računanju pričakovane vrednosti upoštevamo, da je $E(X) = p$, ker

je X le v imenovalcu, dobimo

$$\begin{aligned} I(p) &= -E\left[\frac{-p^2 - X + 2pX}{p^2(1-p)^2}\right] \\ &= -\left[\frac{-p + p^2}{p^2(1-p)^2}\right] \\ &= \frac{1}{p(1-p)} \end{aligned}$$

- Oceniti želimo delež volilcev nekega kandidata. Na vzorcu $n = 500$ zanj glasuje 29 % volilcev. Podajte 95 % interval zaupanja za to oceno.

  Vzorčna ocena je $\hat{p} = 0{,}29$. Standardno napako (torej standardni odklon cenilke) na vzorcu ocenimo s pomočjo $\hat{p}$, ocena standardne napake je torej enaka

  $$\widehat{SE} = \sqrt{\frac{1}{nI(\hat{p})}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}02$$

  Teorija nam pove, da je lahko porazdelitev kvocienta $\frac{p-\hat{p}}{\widehat{SE}}$ aproksimiramo z normalno porazdelitvijo, 95% interval zaupanja je enak $[0{,}25, 0{,}33]$.

Predlogi za vaje v R-u:

- Z R-om narišite sliko 1:

```
> p <- seq(0,1,length=100)  #za 100 vrednosi p med 0 in 1
> y <- p^4*(1-p)            #za vsako vrednost izracunam verjetnost
> plot(p,y,type="l")        #narisem in povezem s krivuljo
```

- Generirajte vzorec velikosti 500, v katerem ima vsak posameznik verjetnost 0,3, da glasuje za nekega kandidata. Ocenite verjetnost z deležom na vzorcu. Ponovite poskus 1000x in si oglejte porazdelitev vzorčnih ocen.

- Na vsakem vzorcu ocenjenemu deležu dodajte še 95% interval zaupanja. Kakšen je delež vzorcev, pri katerih interval zaupanja zajema pravo vrednost (0,3)?

3.2 Povezanost dveh spremenljivk

Zanima nas, kako je prihodek podjetja v neki panogi odvisen od števila zaposlenih. Predpostavimo, da je prihodek podjetja normalno porazdeljen s povprečjem $\beta_0 + \beta_1 X$, kjer je X logaritem števila zaposlenih. Denimo, da imamo podatke o številu zaposlenih in prihodku za vzorec podjetij, radi bi ocenili parametra β_0 in β_1.

- Zapišite gostoto porazdelitve prihodka podjetja, če vemo, da je varianca enaka σ^2.

  Predpostavljamo, da je $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$, torej

  $$f(Y, X|\beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y-\beta_0-\beta_1 X)^2}{2\sigma^2}}$$

- Zapišite funkcijo verjetja. Kaj je funkcija, ki jo moramo maksimizirati?

  Dani so podatki (x_i, y_i), $i = 1, \ldots, n$.

  $$\begin{aligned} L(y, x, \beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \end{aligned}$$

  Logaritem te funkcije je

  $$\log L(y, x, \beta_0, \beta_1, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

  Ker nas zanimata le parametra β_0 in β_1, je prvi del funkcije konstanta, maksimizirati je potrebno le izraz

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

- Izračunajte oceni β_0 in β_1 po metodi največjega verjetja

  Najprej za β_0:

$$\begin{aligned}
&\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2\\
&= \quad -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)
\end{aligned}$$

Če zgornji izraz izenačimo z 0, dobimo (izraz je enak nič za posebni vrednosti β_0 in β_1, ki ju označimo s strešico)

$$\begin{aligned}
-2\left(\sum_{i=1}^{n}y_i-n\widehat{\beta}_0-\widehat{\beta}_1\sum_{i=1}^{n}x_i\right) &= 0\\
\widehat{\beta}_0 &= \frac{1}{n}\left(\sum_{i=1}^{n}y_i-\widehat{\beta}_1\sum_{i=1}^{n}x_i\right)
\end{aligned}$$

Sedaj odvajamo še po β_1:

$$\begin{aligned}
&\frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2\\
&= \quad -2\sum_{i=1}^{n}x_i\left(y_i-\beta_0-\beta_1x_i\right)\\
&= \quad -2\left(\sum_{i=1}^{n}x_iy_i-\beta_0\sum_{i=1}^{n}x_i-\beta_1\sum_{i=1}^{n}x_i^2\right)
\end{aligned}$$

Če zgornji izraz izenačimo z 0, dobimo

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}x_iy_i-\widehat{\beta}_0\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}$$

Združimo obe izpeljavi in (po malce premetavanja členov) dobimo

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

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

- Izračunajte standardno napako za obe oceni.

  Za Fisherjevo matriko informacije moramo izračunati druge odvode. Logaritem funkcije verjetja je enak

  $$\log f(Y, X|\beta_0, \beta_1, \sigma) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y - \beta_0 - \beta_1 X)^2}{2\sigma^2}$$

  Prva odvoda sta enaka

  $$\begin{aligned} \frac{\partial}{\partial \beta_0} \log f(Y, X|\beta_0, \beta_1, \sigma) &= \frac{1}{\sigma^2}(Y - \beta_0 - \beta_1 X) \\ \frac{\partial}{\partial \beta_1} \log f(Y, X|\beta_0, \beta_1, \sigma) &= \frac{X}{\sigma^2}(Y - \beta_0 - \beta_1 X) \end{aligned}$$

  Drugi odvodi so potem

  $$\begin{aligned} \frac{\partial^2}{\partial \beta_0^2} \log f(Y, X|\beta_0, \beta_1, \sigma) &= -\frac{1}{\sigma^2} \\ \frac{\partial^2}{\partial \beta_1^2} \log f(Y, X|\beta_0, \beta_1, \sigma) &= -\frac{X^2}{\sigma^2} \\ \frac{\partial^2}{\partial \beta_1 \beta_0} \log f(Y, X|\beta_0, \beta_1, \sigma) &= -\frac{X}{\sigma^2} \end{aligned}$$

  Členi Fisherjeve matrike informacije so negativne pričakovane vrednosti drugih odvodov. Ker pričakovane vrednosti X oziroma X^2 ne poznamo, ju ocenimo iz podatkov:

  $$I(\beta_0, \beta_1) = \frac{1}{\sigma^2} \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Inverz te matrike je potem

$$I^{-1}(\beta_0, \beta_1) = \frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

in zato

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_0) &= \frac{I_{11}^{-1}}{n} = \frac{1}{n} \frac{\sigma^2 \frac{1}{n}\sum_{i=1}^{n} x_i^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{aligned}$$

ter

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_1) &= \frac{I_{22}^{-1}}{n} = \frac{1}{n} \frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \frac{n\sigma^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{aligned}$$

4.5 Posplošeni test razmerja verjetij

Zanima nas ali imajo zares vsi športniki enako variabilnost hemoglobina. Primerjati želimo meritve k športnikov, naj bodo vrednosti i-tega športnika ($i = 1, \ldots, k$) porazdeljene normalno, torej $X_{ij} \sim N(\mu_i, \sigma_i^2)$, kjer $j = 1, \ldots, n_i$ označujejo meritve pri posamezniku. Predpostavimo, da so vse meritve med seboj neodvisne.

- Zapišite ničelno in alternativno domnevo

  Ničelna domneva:

  $$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

  Alternativna domneva:

  $$H_1 : \sigma_i^2 \text{ niso vse enake}$$

- Najprej vzemimo, da imamo le enega športnika in n njegovih meritev. Kako bi ocenili njegova parametra μ in σ^2 z metodo največjega verjetja? Funkcija verjetja je enaka

$$L(x, \mu, \sigma) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x_j - \mu)^2}{2\sigma^2},$$

del njenega logaritma v katerem nastopata parametra, ki ju želimo oceniti pa je enak

$$\log L(x, \mu, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2.$$

Poiščimo maksimum po μ:

$$\begin{aligned} \frac{\partial \log L(x, \mu, \sigma)}{\partial \mu} &= 0 \\ -\frac{1}{2\widehat{\sigma}^2} \sum_{j=1}^{n} (x_j - \widehat{\mu})(-2) &= 0 \\ \sum_{j=1}^{n} (x_j - \widehat{\mu}) &= 0 \\ \widehat{\mu} &= \frac{1}{n} \sum_{j=1}^{n} x_j \end{aligned}$$

Pa še za varianco:

$$\begin{aligned} \frac{\partial \log L(x, \mu, \sigma)}{\partial \sigma} &= 0 \\ -\frac{n}{\widehat{\sigma}} - \frac{1}{2} \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 \frac{-2}{\widehat{\sigma}^3} &= 0 \\ -\widehat{\sigma}^2 n + \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 &= 0 \\ \widehat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^{n} (x_j - \widehat{\mu})^2 \end{aligned}$$

- Vrnimo se h k športnikom. Utemeljite, da so pod alternativno domnevo ocene parametrov enake

$$\begin{aligned}\hat{\mu}_i &= \frac{1}{n_i}\sum x_{ij} \\ \hat{\sigma}_i^2 &= \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij}-\hat{\mu}_i)^2\end{aligned}$$

  Funkcija verjetja pod alternativno domnevo je enaka

$$L(x,\mu,\sigma) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\sigma_i}\exp -\frac{(x_{ij}-\mu_i)^2}{2\sigma_i^2}$$

  Vsak člen vsote, ki jo dobimo po logaritmiranju gornje funkcije, je sestavljen le iz parametrov enega posameznika, ko odvajamo po tistem parametru torej ostanejo le členi, ki so vezani na tistega posameznika. Za ocenjevanje parametrov za nekega posameznika i torej potrebujemo izključno njegove vrednosti, parametre posameznikov torej ocenimo povsem neodvisno drug od drugega.

- Kakšna je ocena povprečij pod ničelno domnevo?
  Pod ničelno domnevo je σ_i enak za vse i, zato ga v logaritmu funkcije verjetja lahko izpostavimo in ne vpliva na našo oceno posameznih povprečij. Ocena posameznih povprečij je zato enaka kot pod alternativno domnevo.

- Kakšna je ocena variance pod ničelno domnevo?
  Del logaritma funkcije verjetja, ki nas zanima, je enak

$$\log L(x,\mu,\sigma) = -\sum_{i=1}^{k} n_i \log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2.$$

  Odvod po σ izenačimo z 0 in dobimo

$$\hat{\sigma}_0^2 = \frac{1}{\sum_{i=1}^{k} n_i}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\hat{\mu}_i)^2$$

- Kako bi ničelno domnevo preverili s testom razmerja verjetij?

  Zapišemo Wilksov Λ (zgoraj je funkcija verjetja pod alternativno domnevo, spodaj pod ničelno):

$$\begin{aligned}\Lambda &= \frac{\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{\sqrt{2\pi}\widehat{\sigma}_i}\exp\{-\frac{(x_{ij}-\widehat{\mu}_i)^2}{2\widehat{\sigma}_i^2}\}\right)}{\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{\sqrt{2\pi}\widehat{\sigma}_0}\exp\{-\frac{(x_{ij}-\widehat{\mu}_{0i})^2}{2\widehat{\sigma}_0^2}\}\right)} \\ &= \frac{\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\widehat{\sigma}_i}\right)\prod_{i=1}^{k}\exp\{-\frac{\sum_{j=1}^{n_i}(x_{ij}-\widehat{\mu}_i)^2}{2\widehat{\sigma}_i^2}\}}{\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\widehat{\sigma}_0}\right)\exp\{-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(x_{ij}-\widehat{\mu}_{0i})^2}{2\widehat{\sigma}_0^2}\}}\end{aligned}$$

  Vstavimo ocene za variance v eksponent in tako v števcu kot tudi v imenovalcu dobimo $\exp\{-\frac{1}{2}\sum_{i=1}^{k} n_i\}$, ki se zato pokrajša. Logaritem Λ je enak

$$\begin{aligned}\log\Lambda &= -\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log(\widehat{\sigma}_i)\right)+\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\log(\widehat{\sigma}_0)\right) \\ &= \left(\sum_{i=1}^{k} n_i\log(\widehat{\sigma}_0)\right)-\left(\sum_{i=1}^{k} n_i\log(\widehat{\sigma}_i)\right) \\ &= \sum_{i=1}^{k} n_i\left[\log(\widehat{\sigma}_0)-\log(\widehat{\sigma}_i)\right]\end{aligned}$$

  Dvakratna vrednost logaritma verjetij je porazdeljena kot χ^2_{k-1}, saj smo pod alternativno domnevo ocenili $k-1$ parametrov več kot pod ničelno.

5 Linearna regresija

5.1 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 porazde-

ljuje pogojno normalno: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

- 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. Pri iskanju porazdelitve cenilke $\widehat{\beta}_1$ bi se lahko oprli na teorijo metode največjega verjetja, vendar pa v tem primeru aproksimacija ni potrebna. Cenilka je namreč linearna kombinacija vrednosti Y (to smo izpeljali v nalogi 3.2), zato je normalno porazdeljena. Njena varianca (standardna napaka) je ocenjena iz podatkov, zato je standardizirana vrednost cenilke porazdeljena kot t. Testna statistika

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

  je torej porazdeljena 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 posplošenim testom razmerja verjetij?

Namig: Kjer je le mogoče, uporabite rezultate iz prejšnje naloge
Začnimo z ocenami pod ničelno domnevo. Pod ničelno domnevo, je povprečje za vse posameznike enako, neposredno torej lahko uporabimo rezultate iz prejšnje naloge, le da namesto μ pišemo β_0, zato je maksimum funkcije verjetja pod ničelno domnevo enak

$$\begin{aligned} L_0(y, x, \widehat{\beta}_0, \widehat{\sigma}) &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0)^2}{2\widehat{\sigma}^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{n\sum_{i=1}^n (y_i - \widehat{\beta}_0)^2}{2\sum_{i=1}^n (y_i - \widehat{\beta}_0)^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{n}{2} \right\} \end{aligned}$$

Pod alternativno domnevo na enak način uporabimo rezultat, da je ocena $\widehat{\sigma}$ enaka

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

Funkcija verjetja je enaka:

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

In tako je maksimum funkcije verjetja enak

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

Wilksov Λ je enak

$$\begin{aligned}\Lambda &= \frac{L_A}{L_0} = \frac{\frac{1}{(\sqrt{2\pi}\widehat{\sigma}_A)^n}\exp\left\{-\frac{n}{2}\right\}}{\frac{1}{(\sqrt{2\pi}\widehat{\sigma}_0)^n}\exp\left\{-\frac{n}{2}\right\}} \\ &= \frac{\widehat{\sigma}_0^n}{\widehat{\sigma}_A^n} \\ &= \left(\frac{\sum_{i=1}^{n}(y_i-\widehat{\beta}_{00})^2}{\sum_{i=1}^{n}(y_i-\widehat{\beta}_{0A}-x_i\widehat{\beta}_{1A})^2}\right)^n\end{aligned}$$

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.

Predlogi za vaje v R-u:

- 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 y-one
> plot(x,y)                     #narisemo tocke
> popul <- beta0 + beta1*x      #populacijska vrednost premice
> lines(x,popul,col="grey",lwd=2) #dodamo popul. vrednost premice
> 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
```

- Izračunajte posplošeni test razmerja verjetij v R-u

```
> fit0 <- lm(y~1)                  #pod nic. domnevo - le konstanta
> res0 <- y - fit0$coef            #ostanki pod nicelno domnevo
> resA <- y - beta0h - beta1h*x    #ostanki pod alternativno domnevo
#zanima nas razlika log verjetij - konstanto lahko izpustimo:
> logl0 <- -.5*n*log(sum(res0^2))  #loglik pod nicelno
> loglA <- -.5*n*log(sum(resA^2))  #loglik pod alternativno
> Lambda <- 2*(loglA-logl0)        #Wilksov lambda
> 1-pchisq(Lambda,1)               #likelihood ratio test
[1] 4.048e-05
```