PROJEKT PRI PREDMETU STATISTIKA

LUKA ANDRENŠEK

1. NALOGA

1.1. **a).** Naj bosta $N = 43.886$ in $n = 400$ velikost populacije in velikost enostavnega slučajnega vzorca. Opazimo torej števila otrok v teh družinah, torej opazimo $X_1, X_2, \ldots, X_n$. Oceno za populacijsko povprečje μ nam da nepristranska cenilka $\bar{X}$. Dobimo:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = 0.93500.$$

Ker imamo enostaven slučajni vzorec tj. vzorec brez ponavljanja, moramo pri oceni SE upoštevati še popravek iz predavanj. Za cenilko za SE^2 uporabimo sledečo nepristransko cenilko s predavanj:

$$\widehat{\mathrm{SE}}_+^2 = \frac{N-n}{N} \cdot \frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{1}$$

Desno stran korenimo in dobimo:

$$\widehat{\mathrm{SE}}_+ = 0.054151.$$

Čeprav X_i niso neodvisne (so pa enako porazdeljene), lahko vseeno uporabimo CLI, saj je n zelo majhen v primerjavi z N. Tako imamo

$$\frac{\overline{X} - \mu}{\sqrt{\mathrm{Var}(\overline{X})}} \overset{(d)}{\approx} \mathrm{N}(0,1),$$

Imenovalec seveda ni nič drugega kot SE, zato ga zamenjamo s $\widehat{\mathrm{SE}}_+$, izračunamo kvatnil normalne porazdelitve $\Phi^{-1}(0.975)$ in dobimo sledeči interval zaupanja pri stopnji tveganja 0.05:

$$(0.82887, 1.04114).$$

1.2. **b).** Naj bo spet N velikost celotne populacije. Tokrat imamo 4 stratume, njihove velikosti označimo z N_1, N_2, N_3 in N_4 in označimo $p_i = N_i/N$. Še vedno imamo vzorec velikosti n, ampak z omejitvijo, da je $n = n_1 + n_2 + n_3 + n_4$, pri čemer je n_i število vzorčenih primerov iz i-tega stratuma. Ker imamo proporcionalno alokacijo, imamo $n_i/n \approx N_i/N$.

Označimo še μ_i populacijsko povprečje i-tega stratuma. Nadalje naj bo $\overline{X}_i$ vzorčno povprečje vsakega stratuma posebej. Nepristranska cenilka za populacijsko povprečje je

$$\hat{\mu} = \sum_{i=1}^{4} p_i \overline{X}_i = 0.99753.$$

Za standarno napako uporabimo formulo

$$\mathrm{SE}^2(\hat{\mu}) = \mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{4} p_i^2 \, \mathrm{SE}^2(\overline{X}_i)$$

Date: June 22, 2025.

Moramo sedaj oceniti $\mathrm{SE}^2(\overline{X}_i)$. Tu spet uporabimo formulo (1), pri čemer za vsak i vzamemo, da je celotna populacija velikosti N_i, velikost vzorca pa je n_i. Dobimo sledeči rezultat:

$$\widehat{\mathrm{SE}}(\hat{\mu}) = 0.060095.$$

Ker imamo sedaj izračunano standardno napako za $\hat{\mu}$, lahko po enakom argumentu kot pri a) izračunamo interval zaupanja s stopnjo tveganja 0.05. Dobimo:

$$(0.87974, 1.11531).$$

Opazimo, da se je standardna napaka povečala pri stratificarnem vzorčenju. Pri stratificiranju tudi dobimo bolj v desno zamaknjen interval zaupanja.

2. NALOGA

2.1. **a).** Q-Q grafikon izgleda takole:

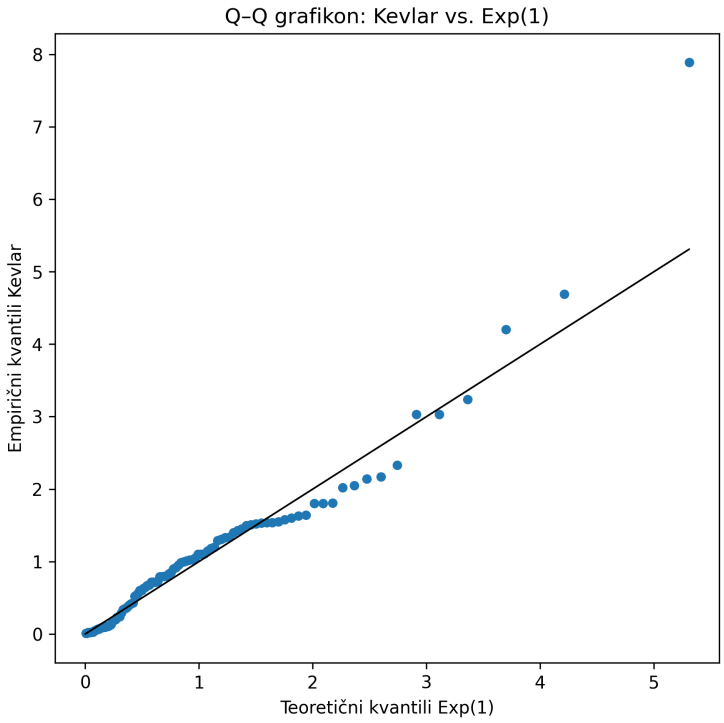

FIGURE 1. Q-Q grafikon empirične porazdelitve Kevlar proti standardni eksponentni

Na začetku izgleda, kot da se lepo ujemajo kvantili, ampak kvantili za verjetnosti, ki so blizu 1, pa se kar razlikujejo.

2.2. **b).** Naj bosta X, Y slučajni spremenljivki. Naj bo X porazdeljena $\mathrm{Exp}(\lambda)$, Y pa $\mathrm{Exp}(\mu)$. Primerjali bomo porazdelitvi X in Y, pri čemer bodo kvantili za X na absicni osi. Dalje je $F_X(x) = 1-e^{-\lambda x}$ in $F_Y(x) = 1-e^{-\mu x}$. Naj bo $\alpha \in (0,1)$. Izračunamo kvatnila $x_\alpha = F_X^{-1}(\alpha)$ in $y_\alpha = F_Y^{-1}(\alpha)$.

Dobimo

$$x_\alpha = -\frac{\ln(1-\alpha)}{\lambda}$$
$$y_\alpha = -\frac{\ln(1-\alpha)}{\mu}.$$

Krivulja je torej sesavljena iz množice točk

$$\{(x_\alpha, y_\alpha) \mid \alpha \in (0,1)\}.$$

Če še zdelimo y_α/x_α, dobimo:

$$y_\alpha = \frac{\lambda}{\mu} x_\alpha.$$

Torej so vsi Q-Q grafikoni premice skozi izhodišče. Narisan je Q-Q grafikon za parametra $\lambda = 3$ in $\mu = 2$.

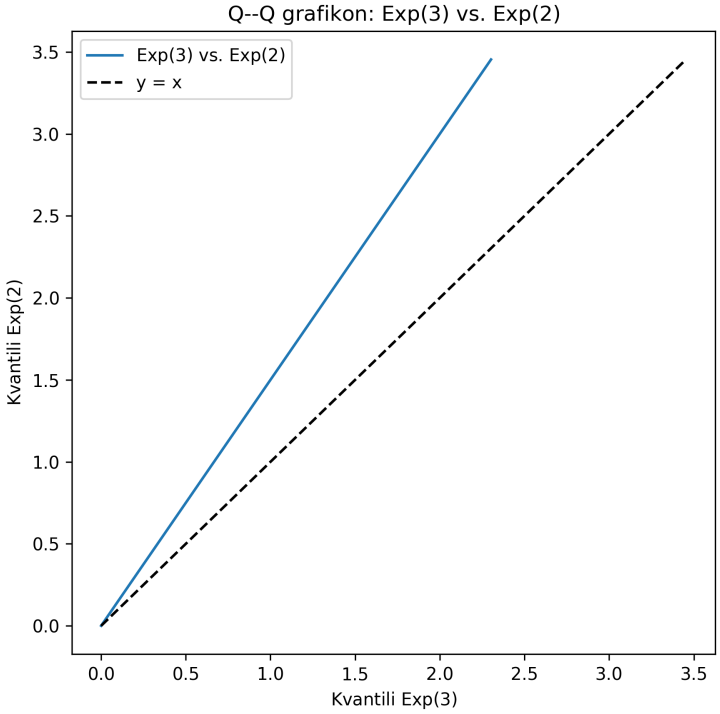

FIGURE 2. Q–Q grafikon: primerjava kvantilov porazdelitev Exp(3) in Exp(2).

2.3. **c).** Opazimo $X_1, X_2, \ldots, X_n$, pri čemer je n število opažanj (vrstic v tabeli). Označimo $f_X(x \mid \lambda) = \lambda e^{-\lambda x}$ za $x \geq 0$. Predpostavimo, da so opažanja $X_1, X_2, \ldots, X_n$ neodvisna in izračunamo verjetje:

$$L(\lambda \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} f_X(X_i \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-n\lambda \overline{X}}.$$

Tu smo upoštevali dejstvo, da so vsa naša opažanja nenegativna. Verjetje logaritmiramo in dobimo

$$l(\lambda \mid X_1, \ldots, X_n) = n \ln \lambda - n\lambda \overline{X}.$$

Ko λ pada proti 0, gre l proti $-\infty$. Ko pa λ narašča preko vseh mej, gre l prav tako proti $-\infty$. Ker je l zvezno odvedljiva, doseže maksimum v stacionarni točki na $(0, \infty)$. Funkcijo odvajamo, odvod enačimo z nič in dobimo cenilko za λ po metodi največjega verjeta:

$$\hat{\lambda} = \frac{1}{\overline{X}}.$$

V našem primeru dobimo:

$$\hat{\lambda} = 0.97669.$$

2.4. **d).** Ker imamo na abscisni osi kvantile porazdelitve Exp(1), in ker sumimo, da je X (opazovana količina v tabeli) porazdeljena Exp($\hat{\lambda}$), kjer je $\hat{\lambda}$ cenilka za λ po metodi največjega verjetja, bomo dodali premico

$$y = \frac{1}{\hat{\lambda}} x.$$

Dobimo

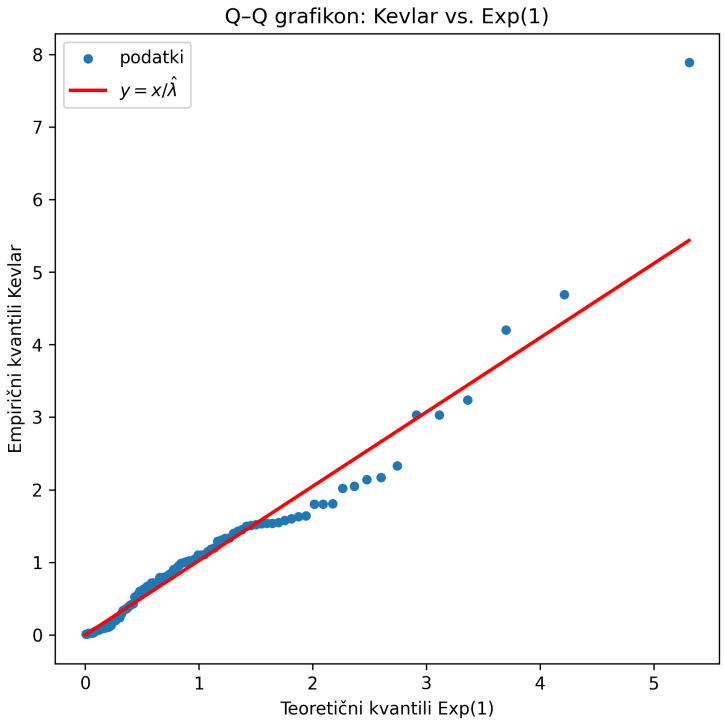

FIGURE 3. Q–Q grafikon: empirični kvantili Kevlar podatkov proti teoretičnim kvantilom standardne eksponentne porazdelitve Exp(1). Dodana je rdeča premica $y = x/\hat{\lambda}$, kjer je $\hat{\lambda} = 1/\bar{x} \approx 0.97669$, ki ponazarja ujemanje s porazdelitvijo Exp($\hat{\lambda}$).

Vtis je malo boljši, ampak še vedno ni Q-Q grafikon zelo prepričljiv.

2.5. **e).** Naša opažanja razdelimo v r razredov, pri čemer je i-ti razred interval $(a_i, b_i]$. Naj bo f_i opažena frekvenca za i-ti razred. Naj bo F porazdelitvena funkcija za porazdelitev, ki naj bi

modelirala opažanja X. Potem je $p_i = F(b_i) - F(a_i)$. Pričakovana frekvenca je tako $\hat{f}_i = p_i n$, pri čemer je n število opažanj. Histogram bo zrisan z vrednostmi

$$d_i = \sqrt{f_i} - \sqrt{\hat{f}_i}.$$

V našem primeru bo porazdelitvena funkcija F porazdelitvena funkcija $\mathrm{Exp}(\hat{\lambda})$ porazdelitve.

Širine razredov določimo v skladu z modificiranim Freedman-Diaconisovim pravilom.

$$\text{širina} \approx \frac{2.6\,\mathrm{IQR}}{n^{1/3}}.$$

Dobimo sledeči viseči histogram iz razlike korenov (rootogram).

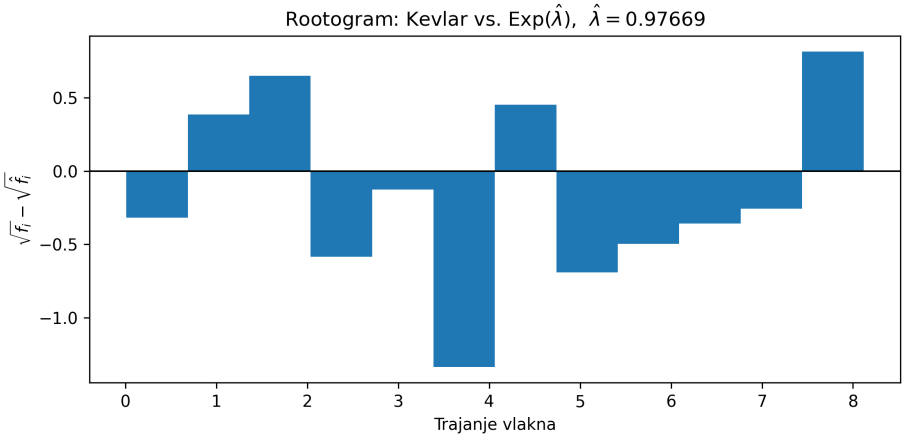

FIGURE 4. Rootogram: primerjava opazovanih in pričakovanih korenov frekvenc za $\mathrm{Exp}(\hat{\lambda})$, kjer $\hat{\lambda} = 1/\bar{x}$.

Prileganje ne zgleda najbolj obetavno, saj imamo en stolpec, ki seže preko 1 po absolutni vrednosti.

2.6. **f).**

2.6.1. *a).* Ponovno zrišemo Q–Q grafikon, tokrat na log-log skali.

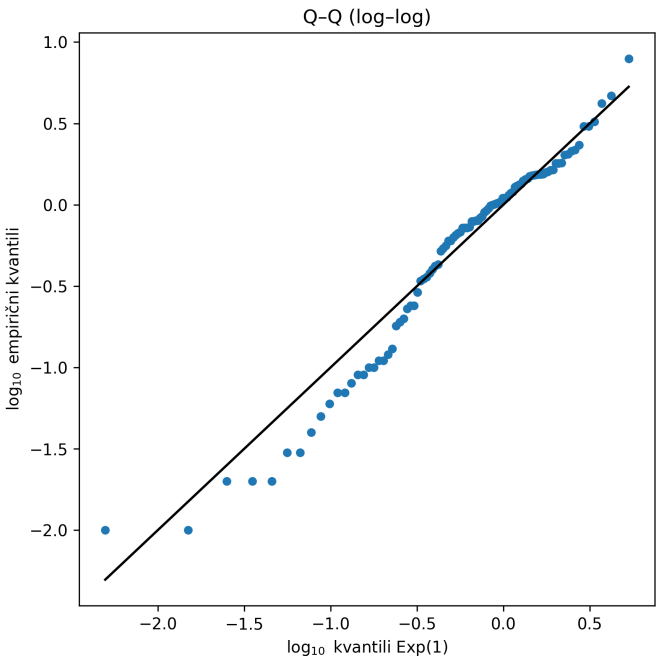

FIGURE 5. Q–Q grafikon na log–log lestvici: Kevlar vs. Exp(1).

Tokrat izgleda Q-Q grafikon bolj obetaven.

2.6.2. *b).* Prej smo imeli enačbo

$$y_\alpha = \frac{\lambda}{\mu} x_\alpha.$$

Uvedemo nove spremenljivke $\widetilde{y}_\alpha = \log y_\alpha$ in $\widetilde{x}_\alpha = \log x_\alpha$ in dobimo novo enačbo

$$\widetilde{y}_\alpha = \log \frac{\lambda}{\mu} + \widetilde{x}_\alpha.$$

Torej je zamaknjena premica s smernim koefecientom 1.

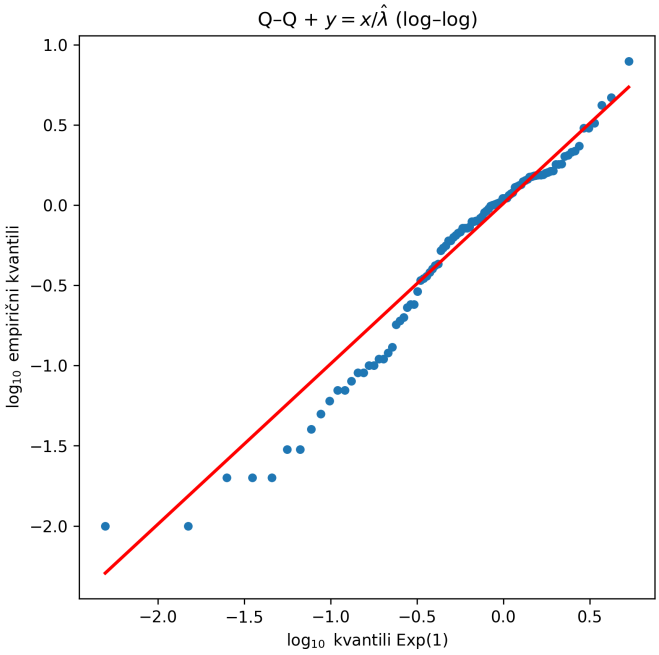

FIGURE 6. Q-Q (log-log) z dodano premico $y = x/\hat{\lambda}$, $\hat{\lambda} = 0{,}97669$.

2.6.3. *d).* Izgleda bolje kot pri nelogaritmirani lestvici.

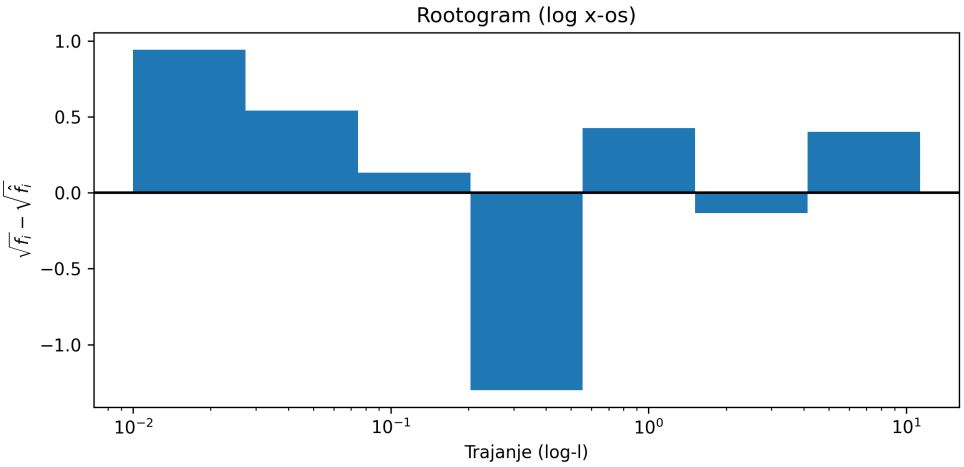

FIGURE 7. Rootogram z logaritemsko absciso za Exp($\hat{\lambda}$).

2.6.4. *e).* Še vedno ne izgleda najbolj obetavno.

2.7. **Razmisleki.**

2.7.1. *1. vprašanje.* Če je fiksno število razredov, recimo r, opažanja $X_1, \ldots, X_n$ so neodvisna in X_i so enako porazdeljene slučajne spremenljivke, pričakovane frekvence so dovolj velike in razredi dovolj ozki, je število opažanj v nekem razredu binomsko. Bolj natančno, naj bodo $F_1, F_2, \ldots, F_r$ števila opaženih frekvenc v razredih. Potem je

$$(F_1, \ldots, F_r) \overset{(d)}{\approx} \text{Multinom}(n, (p_1, \ldots, p_r)),$$

pri čemer je $p_i = F(b_i) - F(a_i)$, pri čemer je F porazdelitvena funkcija za opažanja in a_i, b_i meji i-tega razreda.

2.7.2. *2. vprašanje.* Zanima nas porazdelitev $\sqrt{F_i}$. Ker je $F_i \approx \text{Bin}(n, p_i)$, bo po CLI

$$F_i \overset{(d)}{\approx} \text{N}(np_i, np_i(1-p_i)).$$

Dodatno še uporabimo aproksimacijo $x(1-x) \approx x$ za majhne x in dobimo:

$$F_i \overset{(d)}{\approx} \text{N}(np_i, np_i).$$

Approksimiramo kvadratni koren okoli točke np_i do linearnega reda in dobimo:

$$\sqrt{F_i} \approx \sqrt{np_i} + \frac{1}{2\sqrt{np_i}} \cdot (F_i - np_i).$$

Porazdelitev slučajne spremenljivke na desni je (približno) znana. Dobimo

$$\sqrt{F_i} \overset{(d)}{\approx} \text{N}(\sqrt{np_i}, 1/4).$$

Pričakovana kvadratična napaka, ki je prišla iz Taylorjeve aproksimacije za kvadratni koren, je

$$\begin{aligned} R_2 &= |-\frac{1}{8}(np_i)^{-3/2} \cdot \mathbb{E}[(F_i - np_i)^2]| \approx \frac{1}{8}(np_i)^{-3/2} \cdot \mathbb{E}[(\text{N}(0, np_i))^2] \\ &= \frac{1}{8}(np_i)^{-3/2} \cdot np_i \cdot \mathbb{E}[(\text{N}(0,1))^2] = \frac{1}{8\sqrt{np_i}}, \end{aligned}$$

pri čemer smo upoštevali, da je drugi moment standardne normalne porazdelitve enak 1. Ker je standardni odklon $\sqrt{F_i}$ približno 0.5, vidimo, da je R_2 res majhna v primerjavi s standardnim odklonom $\sqrt{F_i}$.

2.7.3. *3. vprašanje.* V visečem rootogramu so višine (ali globine?) stolpcev enake

$$d_i = \sqrt{F_i} - \sqrt{np_i}.$$

Torej je $d_i \overset{(d)}{\approx} \text{N}(0, 1/4)$. Torej, če modeliramo opažanja s pravilno teoretično porazdelitvijo, bi se morale višine stolpcev gibati okoli 0, s standardno deviacijo 1/2 po normalni porazdelitvi.

Dobro prileganje bi torej pomenilo, da so višine stolpcev v pasu $(-0.98, 0.98)$, torej odstopanje 1.96 standardnih odklonov, če želimo stopnjo zaupanja 95%. Če so višine stolpcev izven tega pasu, je to znak slabega prileganja.

3. Naloga

3.1. **a).** Naj bo T_i izmerjena temperatura, L_i leto meritve in M_i mesec meritve. Opazimo podatke $(T_i, (L_i, M_i))$. Če modeliramo z enostavno linearno regresijo, predpostavljamo, da podatki sledijo predlogi

$$T_i = \beta_0 + \beta_1 L_i + \beta_2 M_i + \epsilon_i,$$

pri čemer $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ in so neodvisni med seboj. Naj bo še X matrika, ki ima v 1. stolpcu same enke, v 2. stolpcu ima leta meritev in v 3. ima mesce meritev. Cenilko za $\beta = [\beta_0, \beta_1, \beta_2]^T$ dobimo po metodi najmanjših kvadratov. Dobimo

$$\hat{\beta} = \begin{bmatrix} -104.32912 \\ 0.056446 \\ 0.39679 \end{bmatrix}.$$

Letni trend oz. podnebno spreminjanje nam pove $\hat{\beta}_1$, torej 2. element v zgornjem vektorju, torej naj bi se pobnebje vsako leto segrelo za 0.056446 stopinje celzija.

Sedaj moramo preizkusiti domnevo $H_0; \beta_1 = 0$ porit alternativni domnevi $H_1; \beta_1 \neq 0$. Poslužimo se sledeče statistike:

$$t_1 = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}_{\hat{\beta}_1}}.$$

Pri veljavnosti H_0 ima statistika t_1 Studentovo porazdelitev z $n-3$ prostostnimi stopnjami.

Moramo še izračunati $\widehat{\mathrm{SE}}_{\hat{\beta}_1}$. Vemo, da je $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2 a_{2,2})$, kjer je $(X^T X)^{-1} = (a_{ij})$. Od tod dobimo

$$\mathrm{SE}^2_{\hat{\beta}_1} = \sigma^2 a_{2,2}.$$

Ker pa σ^2 ne poznamo, jo bomo morali oceniti. To storimo s cenilko:

$$\hat{\sigma}^2_+ = \frac{\mathrm{RSS}}{n-3}. \tag{2}$$

Količino RSS pa izračunamo po formuli:

$$\mathrm{RSS} = \|T - X\hat{\beta}\|^2.$$

Od tod dobimo cenilko $\widehat{\mathrm{SE}}_{\hat{\beta}_1}$ za $\mathrm{SE}_{\hat{\beta}_1}$:

$$\widehat{\mathrm{SE}}_{\hat{\beta}_1} = \frac{\|T - X\hat{\beta}\|}{\sqrt{n-3}} \sqrt{a_{2,2}}.$$

Za t_1 statistiko dobimo vrednost 1.24556, pripadajoča p vrednost pa je 0.21374, zato ne zavrnemo ničelne domneve. Zato sklep o segrevanju podnebja ni statistično značilen s stopnjo tveganja 0.05.

3.2. **b).** V modelu bomo odstranili spremenljivko M_i in dodali spremenljivki

$$S_i = \sin\frac{2\pi M_i}{12} \qquad \text{in} \qquad C_i = \cos\frac{2\pi M_i}{12}.$$

Model bo potem sledil predlogi

$$T_i = \beta_0 + \beta_1 L_i + \beta_2 S_i + \beta_3 C_i + \epsilon_i,$$

pri čemer so $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. Od tu dalje postopamo enako kot v primeru a). Morda je vredno omembne, da bo število prostostnih stopenj enako $n-4$ in pa matrika X se bo seveda spremenila.

Vektor $\hat{\beta}$ se glasi:

$$\hat{\beta} = \begin{bmatrix} -101.74995 \\ 0.056446 \\ -5.15602 \\ -9.00812 \end{bmatrix}.$$

Tokrat naj bi se ozračje (spet) letno segrelo za 0.056446. Tokratna vrednost statistike t_1, če testiramo ničelno domnevo $H_0; \beta_1 = 0$ proti alternativni domnevi $H_1; \beta_1 \neq 0$, je 5.37818, medtem ko pa je p vrednost tokrat enaka $1.36552 \cdot 10^{-7}$. Opazimo, da se je p vrednost bistveno zmanjšala

na pram p vrednosti iz primera a). V tem primeru zavrnemo ničelno domnevo s stopnjo tveganja 0.05.

3.3. **c).** Za napoved januarske temperature bomo v model iz b) vstavili vrednosti $L = 2044$, $S = \sin(2\pi/12)$ in $C = \cos(2\pi/12)$. Dobimo napoved

$$\widehat{T}_1^* = 3.24651.$$

Za napoved povprečja $\widehat{\overline{T}}^*$ bomo izračunali $\widehat{T}_i^*$ za $1 \leq i \leq 12$, pri čemer je $\widehat{T}_i^*$ napoved povprečne temperature v i-tem mescu v letu 2044. Dobimo

$$\widehat{\overline{T}}^* = \frac{\widehat{T}_1^* + \cdots + \widehat{T}_{12}^*}{12} = 13.62578.$$

Sedaj moramo še skonstruirati intervale zaupanja. Torej imamo model

$$T = X\beta + \epsilon.$$

Za napoved $T^* = X^*\beta + \epsilon^*$ pri X^*, uporabimo

$$\widehat{T}^* = X^*\hat{\beta}.$$

Ob predpostavki, da so šumi ϵ in ϵ^* neodvisni in normalno porazdeljeni, še velja

$$\frac{T^* - X^*\hat{\beta}}{\hat{\sigma}_+\sqrt{1 + X^*(X^TX)^{-1}(X^*)^T}} \sim \text{Student}(n-p), \tag{3}$$

pri čemer je p število parametrov modela ozirom število stolpcev matrike X, medtem pa je $\hat{\sigma}_+$ dobljena na analogen način kot v (2) (imenovalec se seveda spremeni v $n-p$). Omenimo, da je X^* pravzaprav kovektor oz. vrstica.

Na podlagi zgornjega konstruiramo interval zaupanja za $\widehat{T}_1^*$ (vstavimo $X^* = [1, 2044, \sin(2\pi/12), \cos(2\pi/12)]$ in $p = 4$). Dobimo interval zaupanja s stopnjo tveganja 0.05:

$$(-0.23529, 6.72831).$$

Za interval zaupanja za povprečje uporabimo podobno metodo. Definirajmo $X_i^* = [1, 2044, \sin(2\pi i/12), \cos(2\pi i/$ Potem je

$$\begin{aligned}\widehat{\overline{T}}^* &= \frac{\widehat{T}_1^* + \cdots + \widehat{T}_{12}^*}{12} \\ &= \frac{1}{12}\sum_{i=1}^{12} X_i^*\hat{\beta} \\ &= \overline{X}^*\hat{\beta},\end{aligned}$$

pri čemer je $\overline{X}^*$ povprečni kovektor kovektorjev X_i^*. Sedaj bomo uporabili formulo (3), a moramo upoštevati, da pri napovedi povprečja seštejemo 12 neodvisnih kopij spremenljivk ϵ_i^*, zato se formula spremeni v

$$\frac{\overline{T}^* - \overline{X}^*\hat{\beta}}{\hat{\sigma}_+\sqrt{\frac{1}{12} + \overline{X}^*(X^TX)^{-1}(\overline{X}^*)^T}} \sim \text{Student}(n-p). \tag{4}$$

Sedaj dobimo sledeči interval zaupanja s stopnjo tveganja 0.05:

$$(12.39032, 14.86124).$$