Projektna naloga iz statistike

Andraž Čepič

19. 7. 2022

V projektu ves čas uporabljamo Python s paketi Pandas, NumPy, SciPy, Jupyter in Matplotlib. Za izračune gostot, komulativnih funkcij, itd. uporabljamo knjižnico `scipy.stats`. Vsi programi, spisani v namen obdelave podatkov, se nahajajo v mapi `skripte`. Posebej shranjeni rezultati so v mapi `rezultati`, ostali so pa v Jupyter zvezkih, kjer je so se tudi izvedli vsi izračuni.

Naloga 1

V namen obdelave podatkov smo napisali Jupyter zvezek `kibergrad.ipynb`.

(a)

Izberemo enostavni slučanji vzorec velikosti 200 s funkcijo `pandas.DataFrame.sample`. Če so

$$X_1, \ldots, X_{200}$$

števila otrok vsake od vzorčenih družin, je primerna ocena za povprečje enaka

$$\overline{X} = \frac{X_1 + \cdots + X_{200}}{200}.$$

Za naš specifičen vzorec dobimo oceno za povprečno število otrok v mestu Kibergrad:

$$\overline{X} \approx 0{,}755$$

(b)

Ocena za standardno napako je podana s formulo

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

kjer je N velikost populacije in n velikost enostavnega slučanega vzorca. V našem primeru je $N = 43.886$ in $n = 200$. Tako za naš vzorec dobimo

$$\widehat{\mathrm{SE}} \approx 0{,}0753.$$

Za enostavno slučajno vzorčenje so intervali zaupanja ocen povprečja oblike

$$\overline{X} - \widehat{\mathrm{SE}} \cdot F_t^{-1}(1 - \frac{\alpha}{2}) < \mu < \overline{X} + \widehat{\mathrm{SE}} \cdot F_t^{-1}(1 - \frac{\alpha}{2}),$$

kjer je F_t komulativna funkcija Studentove t-porazdelitve z $n-1$ prostostnimi stopnjami in $\alpha = 0.05$ stopnja tveganja. V našem primeru dobimo interval zaupanja

$$0{,}6064 < \mu < 0{,}9036$$

(c)

Pravo populacijsko povprečje se glasi

$$\mu = \frac{x_1 + \ldots + x_N}{N} \approx 0{,}9479.$$

Prava standardna napaka za enostavni slučanji vzorec velikosti $n = 200$ je

$$\mathrm{SE}^2 = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}, \tag{1}$$

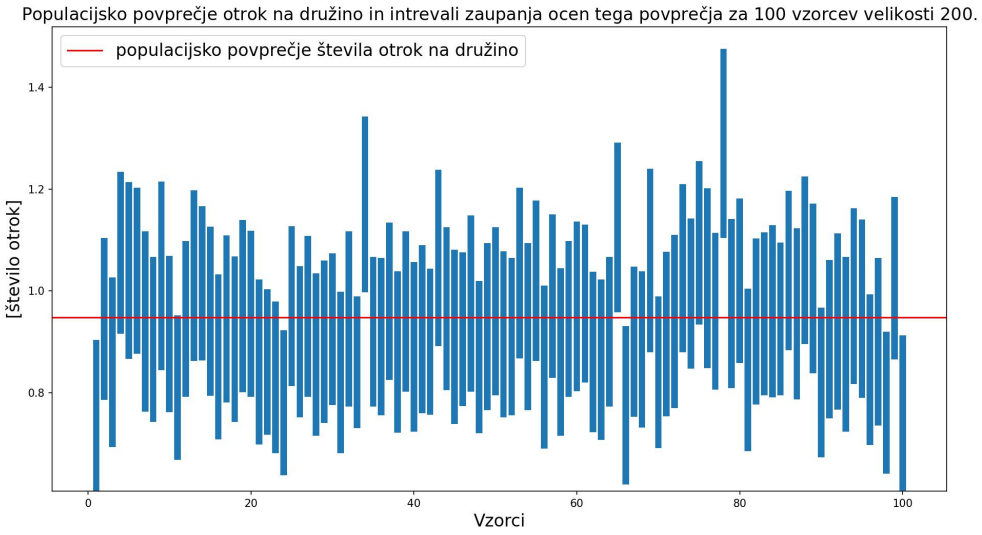
kjer je σ^2 varianca za celo populacijo. Za naše podatke je

$$\mathrm{SE} \approx 0{,}0816.$$

Opazimo, da je ocena za povprečje manjša od pravega povprečja in ocena za standardno napako je prav tako malo manjša, vendar se razlikuje šele v tretji decimalki. Ne, interval zaupanja ne pokrije populacijskega povprečja.

(d)

Intervale zaupanja izračunamo na enak način, kot smo ga za prvi vzorec. Rezultati se nahajajo v mapi `rezultati`, in sicer v `intervali_zaupanja_200.csv`. Naslednja slika prikazuje te intervale zaupanja in populacijsko povprečje:



Izračunamo, da populacijsko povprečje pokrije 92 intervalov zaupanja, oz. delež intervalov, ki pokrijejo populacijsko povprečje, je 0,92.

(e)

Če označimo i-to oceno povprečja iz i-tega vzorca z μ_i in z $\overline{\mu}$ označimo povprečje teh ocen, je potem ocena za varianco teh ocen enaka

$$\widehat{\sigma}^2 = \frac{1}{100-1}\sum_{i=1}^{100}(\mu_i - \overline{\mu})^2.$$

Torej je standardni odklon enak

$$\widehat{\sigma} \approx 0{,}0862.$$

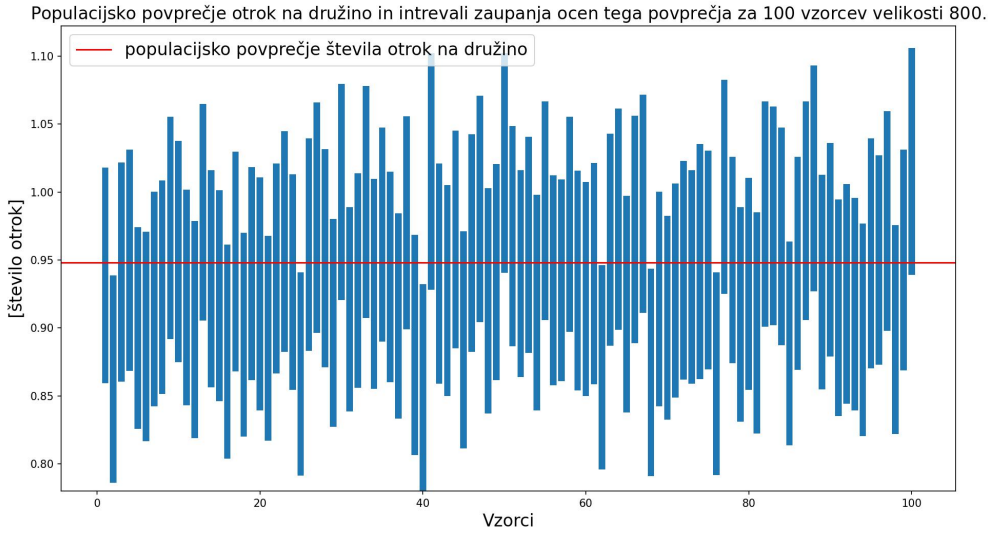
Prava standardna napaka za vzorec velikosti 200 pa je

$$\mathrm{SE} \approx 0{,}0816,$$

kar vemo že od prej. Opazimo, da je standardni odklon ocen povprečja večji od standardne napake.

(f)

Tedaj so rezultati o intervalih zaupanja shranjeni v datoteki `intervali_zaupanja_800.csv` v mapi `rezultati`. Grafično je v tem primeru



Sedaj je delež intervalov zaupanja, ki pokrivajo populacijsko povprečje, enak 0,94. V tem primeru je standardni odklon ocen za povprečje enak

$$\widehat{\sigma}_1 \approx 0{,}037.$$

Prava standardna napaka za vzorec velikosti 800 pa je

$$\mathrm{SE}_1 \approx 0{,}0405.$$

Iz formule (1) je očitno, da je za večje velikosti vzorcev n standardna napaka manjša, zato nas ne preseneča, da je prava stantardna napaka za vzorce velikosti 800 precej manjša od tiste za vzorce velikosti 200. Zanimivo je, da se standardni odklon ocen za povprečje v obeh primerih precej ujema s pravimi napakami, torej je standardni odklon ocen za vzorce velikosti 800 približno pol manjši od tistega, ki pride iz vzorcev velikosti 200. To je smiselno, ker so ocene povprečji pri vzorcih velikosti 800 približno "še enkrat"bližje kot tiste za 200, ker je prava standardna napaka polj manjša. Opazimo tudi, da delež intervalov zaupanja, ki pokrijejo populacijsko povprečje, ni nujno dosti večji, če sploh, od tistega za vzorce velikosti 200. Kljub temu, da so središča intervalov, torej ocene povprečja, bližje populacijskemu povprečju, so širine intervalov manjše. Zato je pridobitev s približanjem središč izgubljena z zmanjšanjem širine intervalov.

Intuitivno si to razlagamo tako, da večji vzorci bolje reprezentirajo pravo populacijo, torej so napake manjše in ocene povprečij boljše. Vendar ni intuitivno, da bomo z eksperimentom približno tolikokrat zgrešili pravo povprečje, kot pri manjših vzorcih.

Naloga 2

Računalniški izračuni in izrisi grafov so v Jupyter datoteki `naloga_2.ipynb`.

(a)

Naj bo X spremenljivka z dano porazdelitivjo, odvisno od parametra θ. Tedaj je

$$\begin{aligned}\mathrm{E}(X) &= \frac{1}{3}\theta + \frac{4}{3}(1-\theta) + (1-\theta) \\ &= -2\theta + \frac{7}{3}.\end{aligned}$$

Torej je

$$\theta = \frac{7}{6} - \frac{1}{2}\,\mathrm{E}(X),$$

zato je

$$\widehat{\theta} := \frac{7}{6} - \frac{1}{2}\overline{X}$$

cenilka za parameter θ preko ocene prvega momenta $\overline{X}$. Ocena $\overline{X}$ za $\mathrm{E}(X)$ je nepristranska in pričakovana vrednost je linearna, torej je $\widehat{\theta}$ nepristranska cenilka za θ. Po naših opažanjih dobimo oceno

$$\widehat{\theta} \approx 0{,}3667.$$

(b)

Srednja kvadratična napaka te ocene je

$$\mathrm{SE}^2 = \mathrm{E}\left(\left(\widehat{\theta} - \theta\right)^2\right) = \mathrm{var}\left(\widehat{\theta}\right) = \frac{\sigma^2}{10},$$

zato je nepristranska ocena za kvadrat standardne napake pri metodi momentov enaka

$$\widehat{\mathrm{SE}}^2 = \frac{\widehat{\sigma}_+^2}{10} = \frac{1}{10} \cdot \frac{1}{10-1} \sum_{i=1}^{10} (X_i - \overline{X})^2$$

in pri teh konkretnih opažanjih ocena za standardno napako znaša

$$\widehat{\mathrm{SE}} \approx 0{,}3399.$$

(c)

Če so opažene vrednosti označene po vrsti z $x_1, \ldots, x_{10}$ in slučajne spremenljivke teh opažanj z $X_1, \ldots, X_{10}$, je verjetje za naša opažanja zaradi neodvisnosti enako

$$\begin{aligned} \mathrm{L}(\theta; X) &= \mathrm{P}(X_1 = x_1) \cdots \mathrm{P}(X_{10} = x_{10}) \\ &= \frac{2^6}{3^{10}} \cdot \theta^4 (1-\theta)^6. \end{aligned}$$

Iščemo maksimum verjetja. Lažje je opravljati z logaritmom, zato definiramo

$$l(\theta; X) = \log(\mathrm{L}(\theta; X)) = \log\left(\frac{2^6}{3^{10}}\right) + 4\log\theta + 6\log(1-\theta).$$

Logaritem je monotona funkcija, zato bo maksimum $\mathrm{L}(\theta; X)$ dosežen ob istem θ kot za $l(\theta; X)$. Odvod l je potem

$$\frac{\partial l}{\partial \theta} = \frac{4}{\theta} - \frac{6}{1-\theta}.$$

Rešujemo torej enačbo

$$\frac{4}{\theta} - \frac{6}{1-\theta} = 0,$$

zato je

$$\begin{aligned} \frac{4}{\theta} &= \frac{6}{1-\theta} \\ 6\theta &= 4 - 4\theta \\ 5\theta &= 2 \\ \theta &= \frac{2}{5}. \end{aligned}$$

To je edini lokalni ekstrem na notranjosti intervala $[0, 1]$, hkrati pa je L na robovih intervala enak 0 in $L(\frac{2}{5}; X) > 0$, zato je to res globalni maksimum verjetja. Za oceno parametra θ po metodi največjega verjetja torej vzamemo

$$\widehat{\theta} = \frac{2}{5} = 0{,}4.$$

(d)

Za oceno standardne napake te ocene uporabimo Fisherjevo informacijo

$$\mathrm{FI}(\theta) = -\,\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}(\theta; X)\right).$$

Ocena za standardno napako se zdaj glasi

$$\widehat{\mathrm{SE}} = \frac{1}{\sqrt{\mathrm{FI}(\theta)}}.$$

V našem primeru je

$$\frac{\partial^2 l}{\partial \theta^2}(\theta; X) = -\frac{4}{\theta^2} - \frac{6}{(1-\theta)^2},$$

zato je

$$\mathrm{FI}(\theta) = \frac{4}{\theta^2} + \frac{6}{(1-\theta)^2}$$

in končno v našem primeru $\theta = \frac{2}{5}$ dobimo oceno za standardno napako:

$$\widehat{\mathrm{SE}} \approx 0{,}1549.$$

(e)

Naj bo f_θ gostota porazdelitve spremenljivke θ. Na intervalu $[0, 1]$ je potem konstantno enaka 1, drugje pa 0. Naj bo H dogodek, da so se zgodila neodvisna opažanja $X_1 = x_1, \ldots, X_{10} = x_{10}$ kot podano. Za pogojno gostoto $f_{\theta|H}$ ob opaženem dogodku H uporabimo Bayesovo formulo za mešane porazdelitve:

$$f_{\theta|H}(t) = \frac{\mathrm{P}(H|\theta = t) \cdot f_\theta(t)}{\mathrm{P}(H)},$$

kjer je verjetnost dogodka H enaka

$$\mathrm{P}(H) = \int_0^1 \mathrm{P}(H|\theta = t) f_\theta(t) dt.$$

Vemo, da je

$$\mathrm{P}(H|\theta = t) = \frac{2^6}{3^{10}} t^4 (1-t)^6,$$

torej je

$$\mathrm{P}(H) = \int_0^1 \frac{2^6}{3^{10}} t^4 (1-t)^6 dt = \frac{2^6}{3^{10}} B(5, 7).$$

Sklepamo, da je

$$f_{\theta|H}(t) = \frac{1}{B(5,7)} t^4 (1-t)^6,$$

kjer je $t \in [0, 1]$, drugje je pa enaka 0. Ugotovili smo, da je aposteriorna porazdelitev $\theta|H$ porazdeljena z beta porazdelitvijo $B(5, 7)$.

Njen modus, tj. točka, v kateri gostota porazdelitve doseže globalni maksimum, dobimo s tem, da poiščemo stacionarne točke. Ker je log strogo naraščajoča funkcija, lahko analiziramo $\log(f_{\theta|H})$. Poleg tega lahko ignoriramo konstanto $\frac{1}{B(5,7)}$. Sedaj je

$$\log(t^4(1-t)^6) = 4 \log t + 6 \log(1-t),$$

torej

$$\log(t^4(1-t)^6)' = \frac{4}{t} - \frac{6}{1-t}.$$

Edina ničla tega izraza je $t = \frac{2}{5}$. Na robu intervala $[0, 1]$ je $f_{\theta|H}$ enaka 0, na notranjosti je pa strogo večja od 0, zato je maksimum dosežen pri $t = \frac{2}{5}$. Modus je potem enak

$$\text{modus} = \frac{2}{5}.$$

Pričakovana vrednost porazdelitve beta $X \sim B(\alpha, \beta)$ se glasi

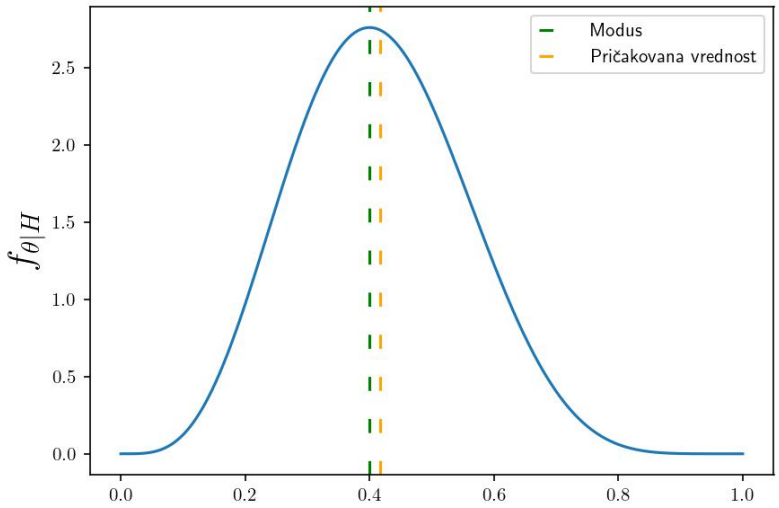
$$\mathrm{E}(X) = \frac{\alpha}{\alpha + \beta}.$$

V našem primeru je pričakovana vrednost enaka

$$\mu = \frac{5}{12}.$$

Opazimo, da je modus natanko enak oceni za θ, ki jo dobimo preko metode največjega verjetja.

Graf gostote $f_{\theta|H}$ je



(f)

Ponovno uporabimo Bayesovo formulo za mešane porazdelitve:

$$f_{\varphi|H}(x) = \frac{\mathrm{P}(H|x) f_{\varphi}(x)}{\mathrm{P}(H)}.$$

V tem primeru je $\theta = \sin^2 \varphi$, zato za parameter x gostote f_{φ} velja

$$\mathrm{P}(H|x) = \frac{2^6}{3^{10}} (\sin^2 x)^4 (\cos^2 x)^6.$$

Spremenljivka φ je porazdeljena enakomerno na intervalu $[0, \frac{\pi}{2}]$, torej velja

$$\begin{aligned} \mathrm{P}(H) &= \int_0^{\frac{\pi}{2}} \frac{2^6}{3^{10}} (\sin^2 x)^4 (\cos^2 x)^6 \frac{2}{\pi} dx, \\ &= \frac{2^6}{3^{10}} \frac{2}{\pi} \int_0^{\frac{\pi}{2}} \sin^8(x) \cos^{12}(x) dx. \end{aligned}$$

Velja

$$\int_0^{\frac{\pi}{2}} \sin^{2p-1}(x)\cos^{2q-1}(x)dx = \frac{1}{2}B(p,q),$$

zato je

$$\mathrm{P}(H) = \frac{2^6}{3^{10}}\frac{1}{\pi}B\left(\frac{9}{2},\frac{13}{2}\right).$$

Dobimo

$$\begin{aligned} f_{\varphi|H}(x) &= \frac{2}{\pi}\frac{\pi}{B\left(\frac{9}{2},\frac{13}{2}\right)}\sin^8(x)\cos^{12}(x), \\ f_{\varphi|H}(x) &= \frac{1}{\frac{1}{2}B\left(\frac{9}{2},\frac{13}{2}\right)}\sin^8(x)\cos^{12}(x) \quad ; x \in \left[0,\frac{\pi}{2}\right]. \end{aligned}$$

Izračunajmo modus. Označimo $C := \frac{1}{\frac{1}{2}B\left(\frac{9}{2},\frac{13}{2}\right)}$ in računamo

$$f'_{\varphi|H} = C\left(8\sin^7(x)\cos^{13}(x) - 12\sin^9(x)\cos^{11}(x)\right) = 0.$$

Sledi

$$\begin{aligned} 8\cos^2 x - 12\sin^2 x &= 0, \\ \tan^2 x &= \frac{8}{12} = \frac{2}{3}, \\ \tan(x) &= \sqrt{\frac{2}{3}}, \end{aligned}$$

saj je $x \in \left[0,\frac{\pi}{2}\right]$. Za ta x je $f_{\varphi|H}(x) > 0$, zato je modus enak $\arctan(\sqrt{\frac{2}{3}})$, oziroma

$$\text{modus} \approx 0{,}6847.$$

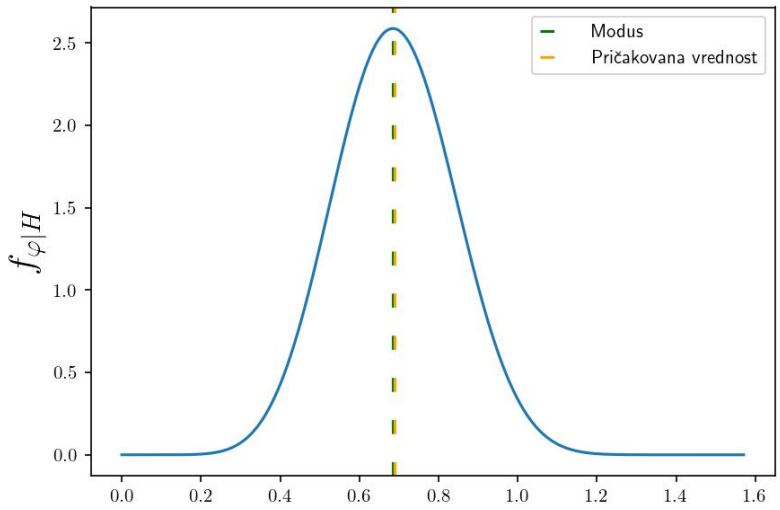
Pričakovana vrednost je enaka

$$\begin{aligned} \mu &= \int_0^{\frac{\pi}{2}} x \cdot f_{\varphi|H}(x)dx \\ &= C\int_0^{\frac{\pi}{2}} x \cdot \sin^8(x)\cos^{12}(x)dx. \end{aligned}$$

Z numerično integracijo, ki smo jo izvedli s funkcijo `scipy.integrate.quad`, smo dobili, da je

$$\mu \approx 0{,}6898.$$

Graf aposteriorne porazdelitve $f_{\varphi|H}$ je



Ko pretvorimo parametra v oceni za θ preko $\theta = \sin^2 \varphi$, dobimo oceni

$$\widehat{\theta}_{\text{modus}} = \theta(\text{modus za } \varphi) = 0{,}4$$

in

$$\widehat{\theta}_{\mu} = \theta(\mu) \approx 0{,}405.$$

Oceni preko modusov se popolnoma ujemata, saj je

$$\sin^2 x = 1 - \cos^2 x = 1 - \frac{1}{1 + \tan^2 x},$$

torej za modus x gostote $f_{\varphi|H}$ velja

$$\sin^2 x = 1 - \frac{1}{1 + \frac{2}{3}} = \frac{2}{5},$$

kar je ravno modus od prej. Pričakovana vrednost od prej je enaka $\frac{5}{12}$, kar je približno enako 0,4167. To je malce večja številka od ocene $\widehat{\theta}_{\mu} = 0{,}405$.

Naloga 3

Izračuni so bili izvedeni v zvezku `temperature.ipynb`. Pri modelu A so pojasnjevalne spremenljivke vseh 420 mesecev in sinus, odvisen od teh mesecev, s periodo eno leto. Oštevilčimo mesece od 1 do 420 in jih označimo z $x_i = i$. Potem se model A glasi

$$Y_i = a + bx_i + c \sin\left(2\pi \frac{x_i - 1}{12}\right) + \varepsilon_i.$$

V matrični obliki je

$$Y = X\beta + \varepsilon,$$

kjer je

$$X := \begin{bmatrix} 1 & x_1 & \sin\left(2\pi\frac{x_1-1}{12}\right) \\ 1 & x_2 & \sin\left(2\pi\frac{x_2-1}{12}\right) \\ \vdots & \vdots & \vdots \\ 1 & x_{420} & \sin\left(2\pi\frac{x_{420}-1}{12}\right) \end{bmatrix},$$

$\beta \in \mathbb{R}^3$ in $\varepsilon \sim N\left(0, \sigma_A^2 I\right)$.

Model B se glasi

$$Y = Z\beta' + \eta,$$

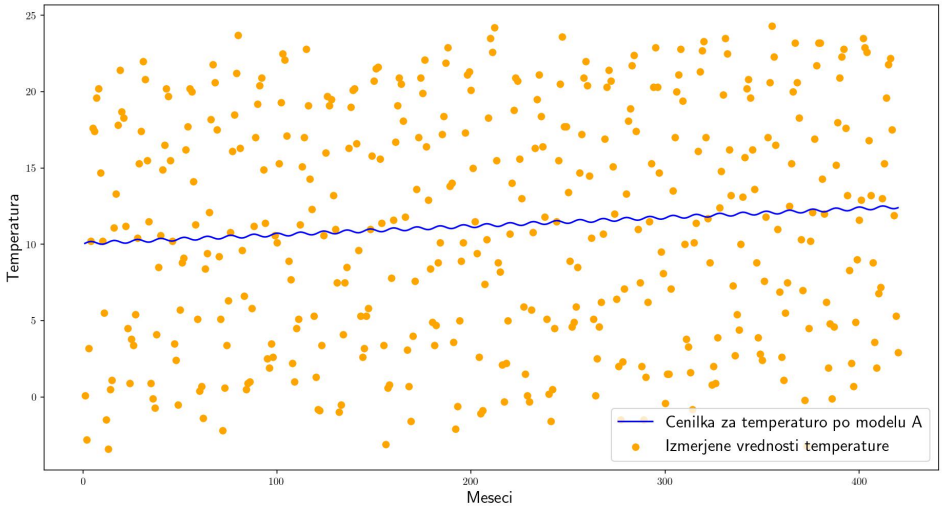
kjer je

$$Z := \begin{bmatrix} x_1 & 1 & 0 & \cdots & 0 & 0 \\ x_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{12} & 0 & 0 & \cdots & 0 & 1 \\ x_{13} & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{420} & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = \begin{bmatrix} x_1 & I_{12} \\ \vdots & I_{12} \\ x_{420} & I_{12} \end{bmatrix},$$

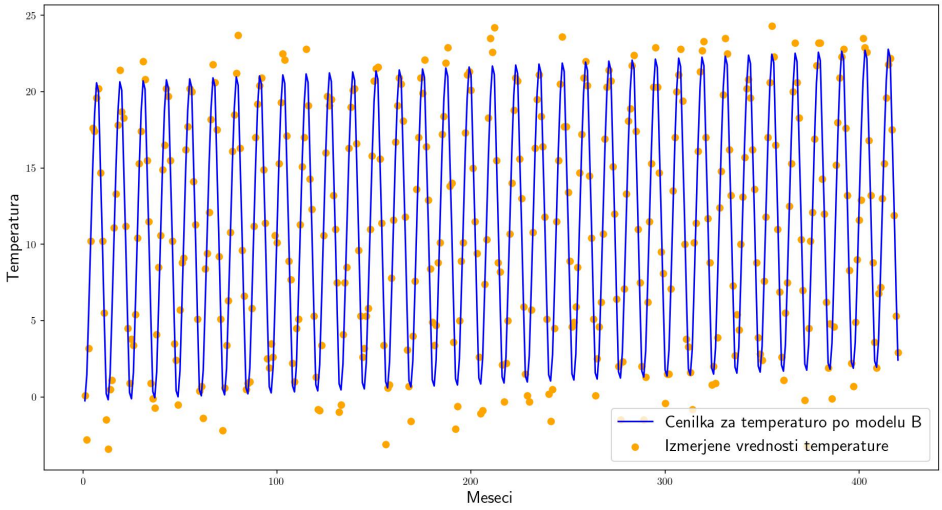
$\beta' \in \mathbb{R}^{13}$ in $\eta \sim N\left(0, \sigma_B^2 I\right)$.

(a)

Cenilke izračunamo po metodi najmanjših kvadratov, za kar smo uporabili `numpy.linalg.lstsq`, ki nam poleg ocene vrne tudi RSS in rang matrik X in Z. Ocenjene temperature po modelu A lahko grafično primerjamo z opaženimi temperaturami:



Podobno storimo za ocenjene temperature po modelu B:



Iz slik se nam zdi, da model A ne bo dobro deloval znotraj modela B. Da preizkusimo model A znotraj modela B, izračunamo Fisherjevo statistiko

$$F = \frac{\frac{\mathrm{RSS}_A - \mathrm{RSS}_B}{\operatorname{rang} Z - \operatorname{rang} X}}{\frac{\mathrm{RSS}_B}{420 - \operatorname{rang} Z}}.$$

Če je $Y = v + \varepsilon$, kjer je $v \in \operatorname{im} Z$, se naša ničelna hipoteza glasi

$$H_0 : v \in \operatorname{im} X.$$

Ovržemo jo, če je

$$F \geq F^{-1}_{\mathrm{Fisher}(p-q,n-q)}(1-\alpha),$$

kjer je $p = \operatorname{rang} Z$, $q = \operatorname{rang} X$, $n = 420$ in α stopnja tveganja. V našem primeru ugotovimo, da ničelno hipotezo ovržemo za obe stopnji tveganja 0,01 in 0,05, torej je model A res preozek.

(b)

Model A ima $\operatorname{rang} X = 3$ parametre in model B ima $\operatorname{rang} Z = 13$ parametrov. Potem je Akaikejeva informacija modela A enaka

$$\mathrm{AIC} = 2 \cdot 3 + 420 \cdot \log(\mathrm{RSS}) \approx 4242,$$

modela B pa

$$\mathrm{AIC} \approx 2988.$$

Model B ima manjšo Akaikejevo informacijo.