Projektna naloga iz Statistike

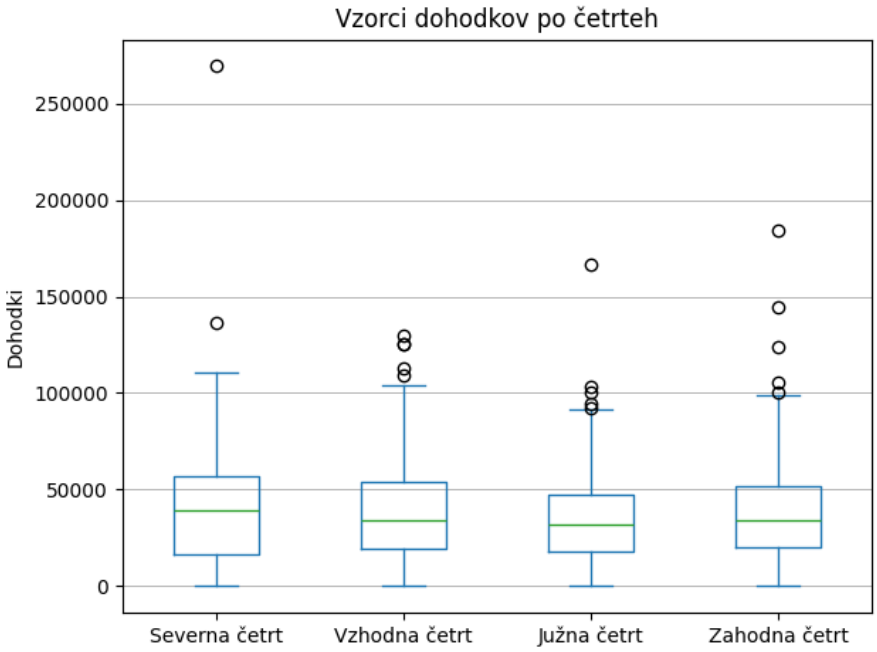
Jan Pantner (jan.pantner@gmail.com)

September 2024

1 Kibergrad

Preučujemo dohodke družin v mestu Kibergrad. Imamo informacije o 43.886 družinah, ki živijo v eni od štirih četrti: v severni četrti stanuje 10.149 družin, v vzhodni 10.390, v južni 13.457 in v zahodni 9.890. Pomagamo si s programom `kibergrad.py`.

Iz vsake četrti vzemimo enostavni slučajni vzorec velikosti 100. Dohodke lahko primerjamo s pomočjo vzporednih škatel z brki, ki so prikazane na sliki 1.

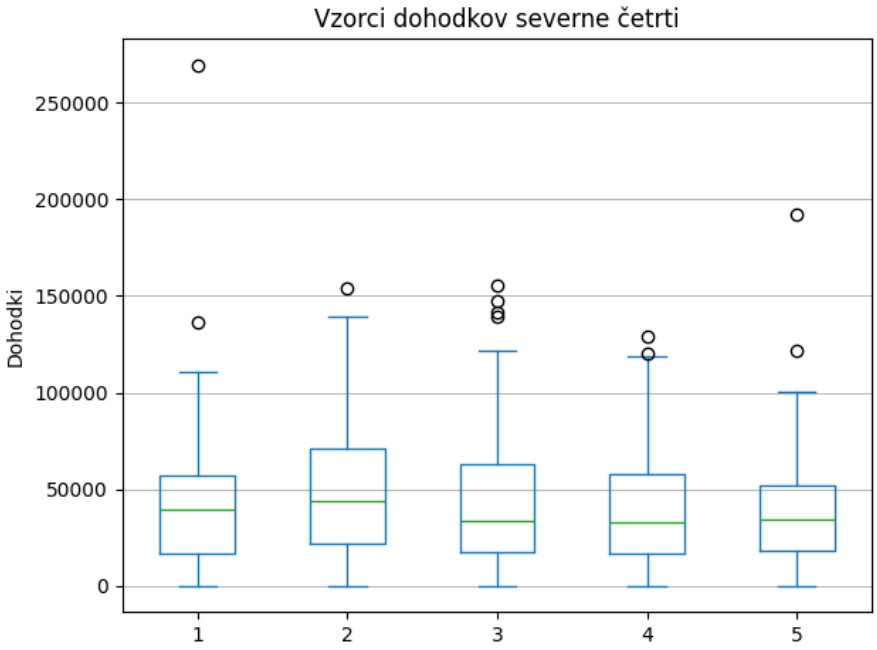


Slika 1: Škatle z brki za dohodke posamezne četrti.

Najprej opazimo, da je v severni četrti osamelec, ki močna izstopa. Prav tako so maksimum, tretji kvartil in mediana v severni četrti največji, vendar razlika ni dovolj velika, da bi lahko pri tako majhnem vzorcu sklepali, da so dohodki severne četrti najvišji. Zdi pa se, da so dohodki severne četrti malenkost višji kot dohodki južne četrti, pri kateri so vse prej omenjene vrednosti najmanjše.

Zdi se, da v splošnem četrt ne vpliva močno na velikost dohodka.

Sedaj vzemimo še štiri enostavne slučajne vzorce velikosti 100 iz severne četrti in poglejmo vzporedne škatle z brki za vzorce iz severne četrti. Prikazane so na sliki 2.



Slika 2: Škatle z brki dohodkov severne četrti.

Vse mediane so večje od 33.000, njihovo povprečje je približno 36.500. Zdi se, da lahko s precejšnjo sigurnostjo sklepamo, da je večina dohodkov višjih od 33.000.

V novih vzorcih ne opazimo tako ekstremnih osamelcev kot pri prvem vzorcu. Večina osamelcev ni zelo oddaljenih od maksimuma.

Za konec si poglejmo še varianco dohodka pojasnjeno s četrtmi in preostalo (rezidualno) varianco. Naj bo N velikost populacije, N_i velikost populacije i-te četrti in w_i velikostni delež i-te četrti. Naj bo μ povprečni dogodek in σ^2 varianca dohodka Kibergrada, μ_i in σ_i^2 pa zaporedoma povprečni dohodek in varianca dohodka i-te četrti. Pojasnjena in nepojasnjena varianca sta zaporedoma določeni s formulama

$$\sigma_B^2 := \sum_{i=1}^{4} w_i(\mu_i - \mu) \quad \text{in} \quad \sigma_W^2 := \sum_{i=1}^{4} w_i\sigma_i^2.$$

Izračunamo, da je

$$\sigma_B^2 = 9.252.923 \quad \text{in} \quad \sigma_W^2 = 1.017.226.451.$$

S četrti pojasnjeni standardni odklon je torej 3.042. Povprečni dohodki četrti, so enaki 45.759, 41.235, 37.473 in 42.158, torej je standardni odklon v primerjavi z njimi majhen. To se ujema z opažanjem, da četrt ne vpliva močno na velikost dohodka.

2 Lomljivost najlonskih palic

Na vzorcu 280 najlonskih palic preizkušamo njihovo lomljivost. Rezultati preizkusa so prikazani v tabeli 1. Pri analizi si pomagamo s programom `najlonske_palice.py`.

št. lomov	0	1	2	3	4	5
št. palic	157	69	35	17	1	1

Tabela 1: Rezultati preizkusa lomljivosti.

Privzemimo, da je število mest, na katerih se je palica zlomila porazdeljeno binomsko $\mathrm{Bin}(5, p)$ za določen neznan p. Privzemimo tudi, da so palice med seboj neodvisne.

Pri teh predpostavkah je logaritem verjetja podan z

$$l(p, x) = \log \prod_{i=1}^{280} \binom{5}{x_i} p^{x_i} (1-p)^{5-x_i}.$$

Izračunamo, da je maksimum dosežen pri približno $\hat{p} = 0{,}14214$. Ta vrednost nam predstavlja oceno p po metodi največjega verjetja.

Sedaj združimo zadnje tri vrednosti in naredimo posplošeni Pearsonov preizkus hi kvadrat. Preizkušamo ničelno domnevo, da je število mest, na katerih se je palica zlomila porazdeljeno binomsko $\mathrm{Bin}(5, p)$ proti alternativni domeni, da ima število lomov katero drugo porazdelitev.

Pearsonova testna statistika je

$$X^2 = \sum_{i=0}^{2} \frac{(x_i - n f_i(\hat{p}))^2}{n f_i(\hat{p})},$$

kjer je $x_0 = 157$, $x_1 = 69$, $x_2 = 35$, $x_3 = 19$ in

$$f_i(\hat{p}) = \begin{cases} \binom{5}{x_i} \hat{p}^{x_j} (1-\hat{p})^{5-x_i}, & i \in \{0, 1, 2\} \\ \sum_{j=3}^{5} \binom{5}{x_j} \hat{p}^{x_j} (1-\hat{p})^{5-x_j}, & i = 3 \end{cases}.$$

Izračunamo, da je $X^2 > 44$. Iz tabela kvantilov porazdelitve hi kvadrat sledi, da ničelno domnevo zavrnemo tako pri stopnji tveganja $\alpha = 0{,}01$ kot tudi pri $\alpha = 0{,}05$.

Recimo sedaj, da imamo za $i = 1, \ldots, 280$ neodvisna opažanja $X_i \sim \mathrm{Bin}(m_i, p_i)$, kjer so parametri m_i znani, p_i pa neznani. S pomočjo razmerja verjetih preizkusimo ničelno domnevo, da so vsi parametri p_i enaki, proti alternativni domnevi, da temu ni tako.

Verjetja sta enaka

$$L_1 = \prod_{i=1}^{280} \binom{m_i}{x_i} p_i^{x_i} (1-p_i)^{m_i - x_i}$$

$$L_0 = \prod_{i=1}^{280} \binom{m_i}{x_i} p_0^{x_i} (1-p_0)^{m_i - x_i}$$

Torej je razmerje verjetij enako

$$\Lambda = \frac{\prod_{i=1}^{280} \hat{p}_i^{x_i}(1-\hat{p}_i)^{m_i-x_i}}{\prod_{i=1}^{280} \hat{p}_0^{x_i}(1-\hat{p}_0)^{m_i-x_i}}$$

Če logaritem verjetja L_1 odvajamo po p_i dobimo

$$\frac{x_i}{p_i} - \frac{x_i - m_i}{1-p_i}.$$

TO DO

3 Spreminjanje temperature v Ljubljani

V tem razdelku preučejemo spreminanje temperature s časom v Ljubljani s pomočjo podatko izmerjenih mesečno med v letih od 1994 do 2023. Pomagamo si s programom `temperature.py`.

Najprej predpostavimo linearni trend in sinusno nihanje s periodo eno leto. To nam da model

$$y_{l,m} = Ax_l + B\sin(x_m\pi/6 + \delta) + C,$$

oziroma

$$y_{l,m} = Ax_l + B\sin(x_m\pi/6) + C\cos(x_m\pi/6) + D,$$

kjer je x_l leto meritve, x_m mesec meritve, $y_{l,m}$ pa izmerjena temperatura v tem mesecu tega leta. Alternativno

$$Y = X_A\beta_A,$$

kjer je $\beta_A^\top = \begin{bmatrix} D & C & B & A \end{bmatrix}$ in

$$X_A = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & \sin(\pi/6) & \cos(\pi/6) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \sin(11\pi/6) & \cos(11\pi/6) \\ 1 & 2 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & \sin(11\pi/6) & \cos(11\pi/6) \\ 1 & 3 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 30 & \sin(11\pi/6) & \cos(11\pi/6) \end{bmatrix}.$$

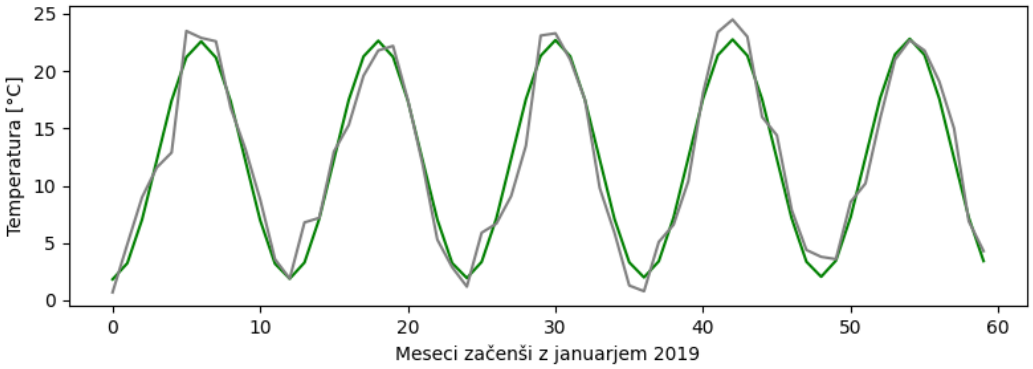
Torej, X_A je matrika velikosti 360×4. Predpostavimo, da je naš model Gaussov. Tedaj je ocena za β_A po metodi največjega verjetja enaka

$$\hat{\beta}_A = (X_A^\top X)^{-1}X^\top Y.$$

Izračunamo

$$\beta_A^\top = \begin{bmatrix} 10.747 & 0.056 & 0.039 & -10.379 \end{bmatrix}.$$

Prva komponenta nam pove, da imamo pozitiven trend, kar se zdi smiselno. Kako dober je model lahko ocenimo s pomočjo slike 3.



Slika 3: Prvi model

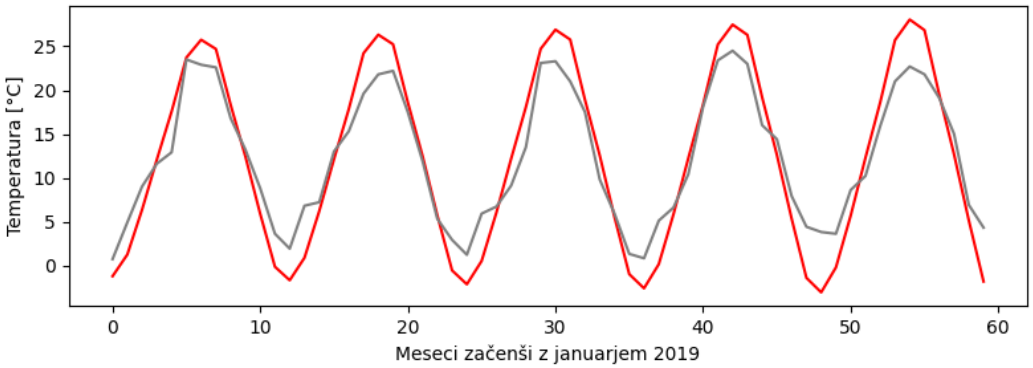
Alternativno predpostavimo linearni trend in spreminanje temperature za vsak mesec v letu posebej. V tem primeru dobimo matriko

$$X_B = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 2 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 1 & 0 & 0 & 0 & \cdots & 0 & 2 \\ 1 & 3 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 1 & 0 & 0 & 0 & \cdots & 0 & 30 \end{bmatrix},$$

ki je velikosti 360×13. Izračunamo

$$\beta_A^\top = \begin{bmatrix} 10.75 & -0.46 & -0.37 & -0.17 & 0.05 & 0.26 & 0.50 & 0.58 & 0.54 & 0.29 & 0.06 & -0.19 & -0.42 \end{bmatrix}.$$

Za model očitno ni preveč dober. To lahko vidimo tudi na sliki 4.



Slika 4: Drugi model

Alternativno si lahko pri izbiri boljšega modela pomagamo z Akaikejevo informacija, ki

je v primeru linearne regresije in Gaussovega modela enaka

$$\mathrm{AIC} := 2p + n \ln \mathrm{RSS},$$

kjer je p število parametrov, n število opažanj in $\mathrm{RSS} = \left|\left|Y - X\hat{\beta}\right|\right|$.

Akaikejeva modela prvega modela je enaka 1261, drugega pa 1577. Torej ima model A manjšo Akaikejevo informacijo in je primernejši, kar se skleda s prejšnjim opažanjem.

Takšna razlika se verjetno pojavi zaradi avtorjeve napačne interpretacije navodil projektne naloge. Verjetno je mišljeno, da je drugi model določen z matriko

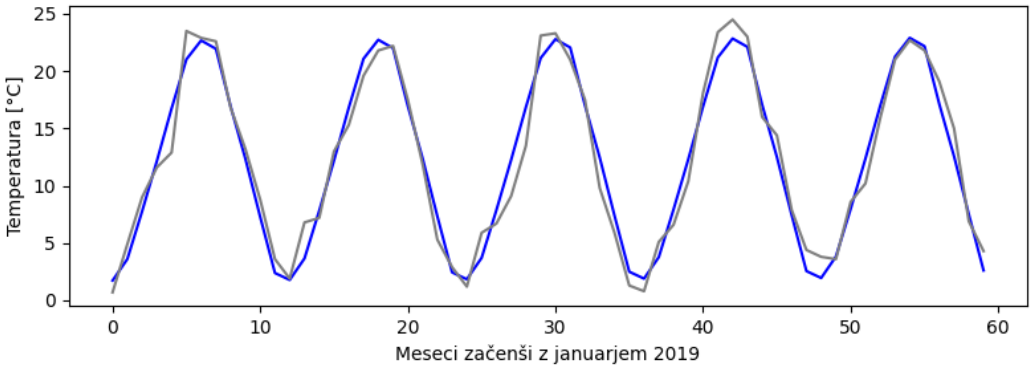
$$X_C = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 2 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 2 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 3 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \\ 30 & 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

V tem primeru je

$$\beta_A^\top = \begin{bmatrix} 0.06 & 0.26 & 2.14 & 6.30 & 10.68 & 15.29 & 19.56 & 21.21 & 20.48 & 15.44 & 10.88 & 5.82 & 0.92 \end{bmatrix}$$

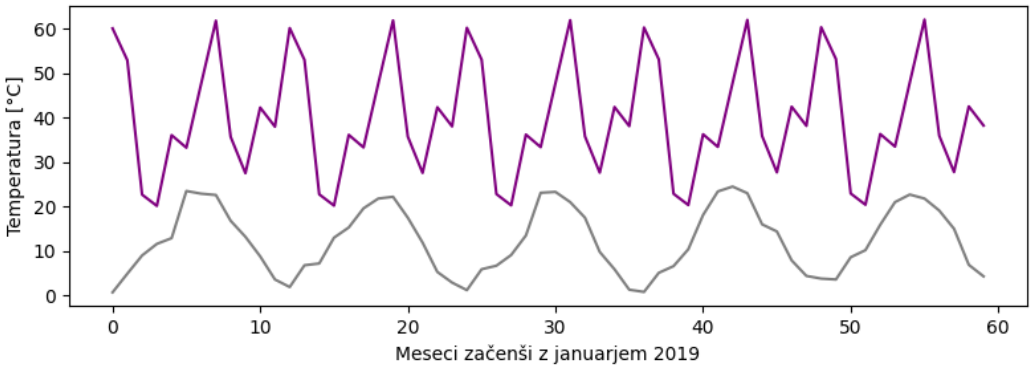
Prva komponenta nam pove, da imamo pozitiven linearen trend, iz ostalih komponent pa je lepo razvidno kako se temperature spreminjajo skozi mesece. Res na primer vidimo, da model predvide najnižje temperature pozimi in najvišje poleti.

Akaikejeva infromacija je enaka 1265. To je malenkost več kot Akaikejeva infromacija prvega modela. Da je model kar dober, vidimo tudi na sliki 5.



Slika 5: Tretji model

Kot zanimivost slika 6 prikazuje kaj se zgodi, če matriki X_C na začetek dodamo še stolpec enic. Nepresenetljivo glede na sliko, je Akaikejeva infromacija tega modela veliko večja od prejšnih, enaka je 2330.



Slika 6: Zelo slab model.