Projektna naloga iz Statistike

Jan Pantner
Profesor: doc. dr. Martin Raič

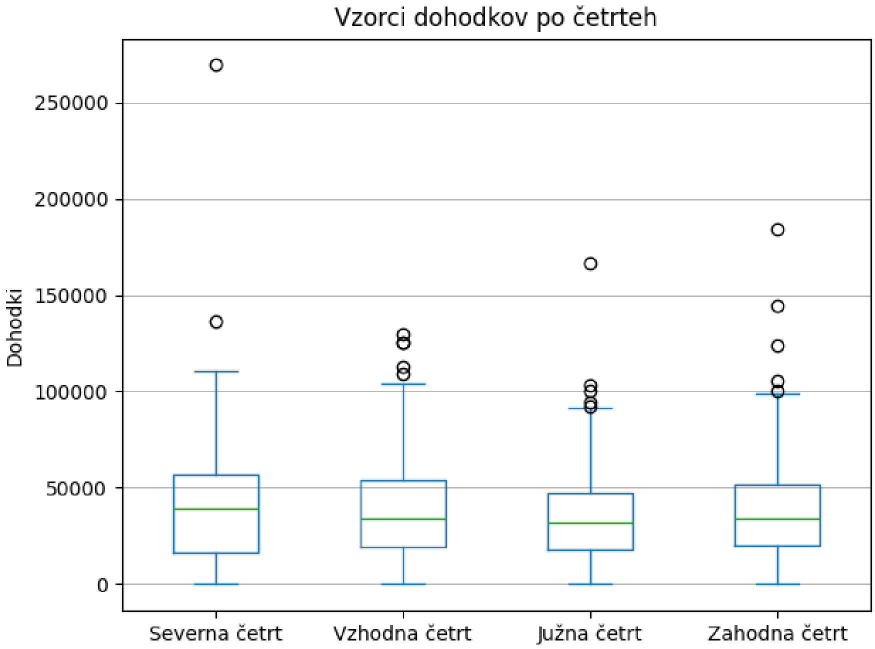
September 2024

Kazalo

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 vzamemo enostavni slučajni vzorec velikosti 100. Dohodke primerjamo s pomočjo škatel z brki, o katerih je več napisano v [3, poglavje 10.6]. Vzporedne škatle z brki so prikazane na sliki 1.

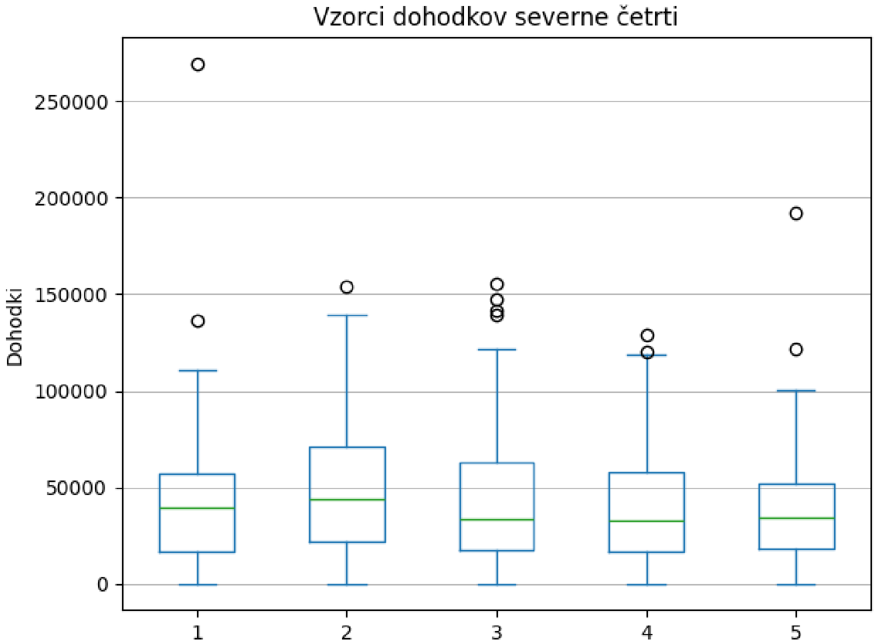


Slika 1: Škatle z brki za dohodke posamezne četrti.

Najprej opazimo, da je v severni četrti osamelec, ki močno 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 vzamemo še štiri enostavne slučajne vzorce velikosti 100 iz severne četrti in pogledamo 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 pogledamo 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 dohodek in σ^2 varianca dohodka Kibergrada, μ_i in σ_i^2 pa zaporedoma povprečni dohodek in varianca dohodka i-te četrti. Pojma pojasnjene in nepojasnjene variance sta razložena v [1]. Zaporedoma sta določena 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 četrtmi pojasnjen 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.

Privzamemo, da je število mest, na katerih se je palica zlomila, porazdeljeno binomsko $\mathrm{Bin}(5, p)$ za neznan parameter p. Privzamemo 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 $\hat{p} = 0{,}14214$. Ta vrednost nam predstavlja oceno p po metodi največjega verjetja.

Sedaj združimo zadnje tri vrednosti in preizkusimo ničelno domnevo, da je število mest, na katerih se je palica zlomila, porazdeljeno binomsko $\mathrm{Bin}(5, p)$, proti alternativni domneni, da ima število lomov katero drugo porazdelitev. To storimo s Pearsonovim preizkusom hi kvadrat, ki je podrobneje opisan v [3, poglavje 9.5].

Pearsonova testna statistika je

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

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 $\chi^2 > 44$. Iz tabele kvantilov porazdelitve 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. Smiselno predpostavimo, da so vsi parametri m_i vsaj 5. S pomočjo razmerja verjetij preizkusimo ničelno domnevo, da so vsi parametri p_i enaki, proti alternativni domnevi, da temu ni tako.

Verjetji sta enaka

$$L_1 = \prod_{i=1}^{280} \binom{m_i}{x_i} p_i^{x_i} (1-p_i)^{m_i - x_i},$$
$$L_0 = \prod_{i=1}^{280} \binom{m_i}{x_i} p_0^{x_i} (1-p_0)^{m_i - x_i}.$$

Če logaritem verjetja L_1 odvajamo po p_i, dobimo

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

Odtod sledi, da je $\hat{p}_i = \frac{x_i}{m_i}$. Če je $x_i \in \{0, m_i\}$, tedaj vzamemo, da je tisti faktor verjetja enak 1. V primeru L_0 pa lahko numerično izračunamo maksimum verjetja in dobimo $\hat{p}_0$.

Na danih podatkih dobimo

$$\Lambda = \log \frac{\prod_{i=158}^{279} \binom{m_i}{x_i} p_i^{x_i} (1 - p_i)^{m_i - x_i}}{\prod_{i=1}^{280} \binom{m_i}{x_i} p_0^{x_i} (1 - p_0)^{m_i - x_i}} = 444{,}489.$$

Wilksovega izreka tukaj ne moremo uporabiti, saj zahteva, da so ocenjeni parametri v notranjosti parametričnega prostora, mi pa imamo tudi primere, ko je $p_i = 0$ ali pa $p_j = 1$. Namesto tega uporabimo metodo *bootstrap*. Med 10.000 simuliranimi vrednostmi preizkusne statistike pri ničelni domnevi jih 0 preseže vrednost 444,489, torej lahko ničelno domnevo ponovno zavrnemo.

3 Spreminjanje temperature v Ljubljani

Preučujemo spreminjanje temperature v Ljubljani s pomočjo podatkov, izmerjenih mesečno v letih od 1994 do 2023. Pomagamo si s programom `temperature.py`.

Najprej predpostavimo linearni trend in sinusno nihanje s periodo enega 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. Več o sinusnem modelu v [4]. Alternativno je model določen z

$$Y = X_A\beta_A,$$

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

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

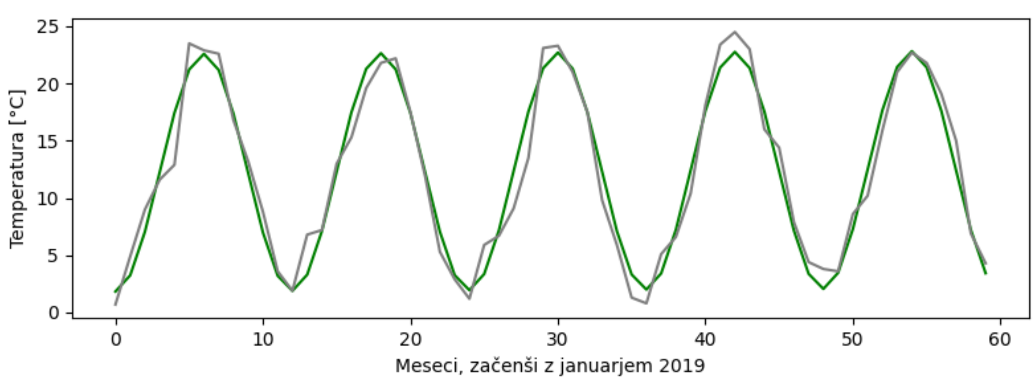
Torej je X_A matrika velikosti 360×4. Predpostavimo, da je naš model Gaussov. Tedaj je ocena za β_A po metodi največjega verjetja enaka

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

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 sklada s teorijo globalnega segrevanja. Kako dober je model, lahko ocenimo s pomočjo slike 3.



Slika 3: Siva krivulja označuje izmerjene podatke, zelena pa oceno prvega modela.

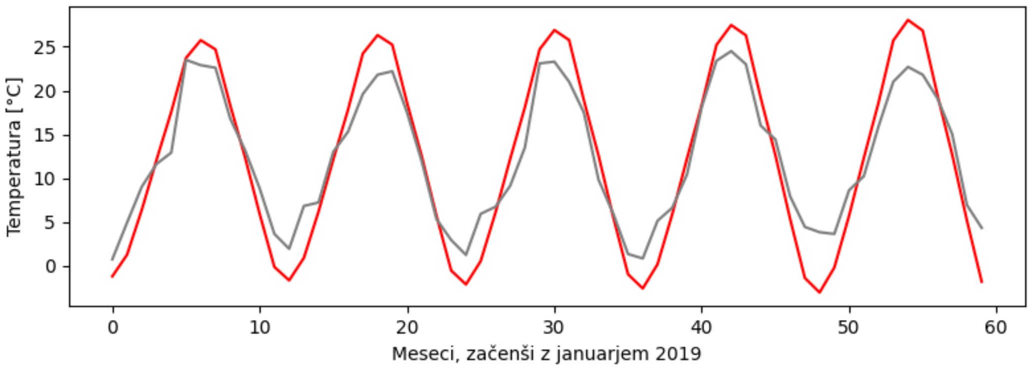
Alternativno predpostavimo linearni trend temperature za vsak mesec v letu posebej. V tem primeru dobimo matriko

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

ki je velikosti 360×13. Izračunamo

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

Na sliki 4 vidimo, da ta model očitno ni preveč dober.



Slika 4: Siva krivulja označuje izmerjene podatke, rdeča pa oceno drugega modela.

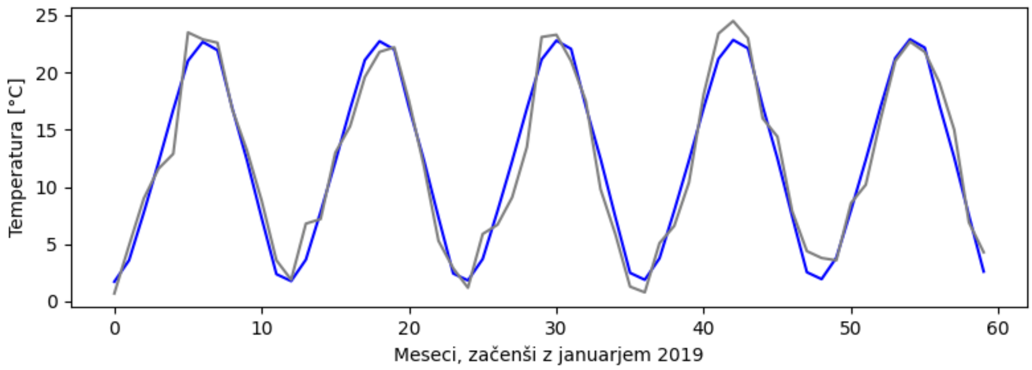
Takšna razlika se verjetno pojavi zaradi avtorjeve napačne interpretacije navodil projektne naloge. Verjetno je mišljeno, da je drugi model določen z matriko

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

To je ponovno matrika velikosti 360×13. 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. Vidimo na primer, da model predvidi najnižje temperature pozimi in najvišje poleti. Model je prikazan na sliki 5.



Slika 5: Siva krivulja označuje izmerjene podatke, modra pa oceno tretjega modela.

Tretji model je širši od prvega. Preizkusimo model A znotraj modela C, torej preizkusimo ničelno domnevo, da velja model A, proti alternativni, da velja model B z odvzetim podmodelom A. To lahko storimo na podlagi statistike

$$F := \frac{(\mathrm{RSS}_A - \mathrm{RSS}_C)(n-p)}{\mathrm{RSS}_C(p-q)},$$

kjer sta p in q prostorski stopnji modelov C in A. Ničelno domnevo zavrnemo, če je

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

kjer je α stopnja tveganja. Izpeljava je na voljo v [2].

V našem primeru je $p = 13$ in $q = 4$. Izračunamo

$$F = 1{,}595, \quad F^{-1}_{\mathrm{fisher}(9,347)}(1-0{,}05) = 1{,}907 \quad \text{in} \quad F^{-1}_{\mathrm{fisher}(9,347)}(1-0{,}01) = 2{,}459.$$

Ničelne domneve torej ne moremo zavrniti.

Določanje optimalnega modela z Akaikejevo informacijo

Pri izbiri boljšega modela si lahko pomagamo z Akaikejevo informacijo, ki je v primeru linearne regresije in Gaussovega modela enaka

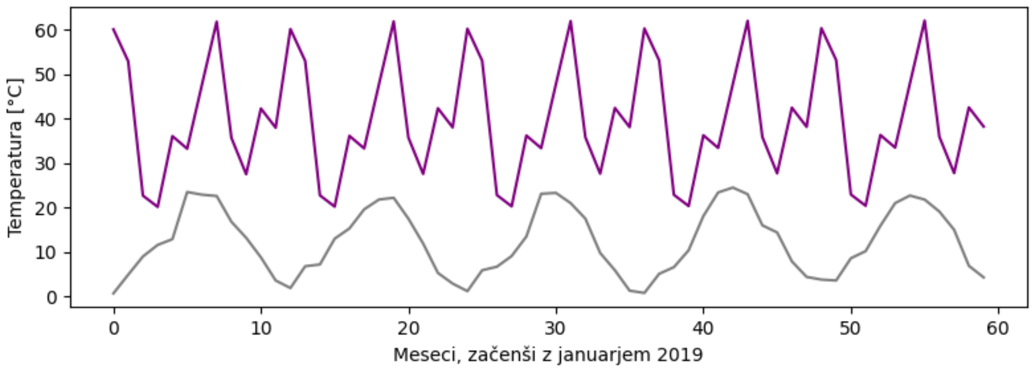
$$\mathrm{AIC} := 2p + n \ln \mathrm{RSS},$$

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

Akaikejeva informacija prvega modela je enaka 1261, tretjega pa 1265. Torej ima model A malenkost manjšo Akaikejevo informacijo in je primernejši.

Akaikejeva infromacija drugega modela je enaka 1577. To se sklada z opažanjem s slike, da je drugi model bistveno slabši od prvega (in tretjega).

Kot zanimivost, slika 6 prikazuje, kaj se zgodi, če matriki X_C na začetek dodamo stolpec enic. Glede na sliko ni presenetljivo, da je Akaikejeva informacija tega modela veliko večja od prejšnjih – enaka je 2330.



Slika 6: Siva krivulja označuje izmerjene podatke, vijolična pa oceno četrtega modela.

Literatura

[1] M. Raič. *Razcep variance*. Dopolnitve predavanj iz statistike v študijskem letu 2023/24. URL: https://ucilnica.fmf.uni-lj.si/mod/resource/view.php?id=63830 (pridobljeno 16. 9. 2024).

[2] M. Raič. *Statistično sklepanje pri linearni regresiji*. Dopolnitve predavanj iz statistike v študijskem letu 2023/24. URL: https://ucilnica.fmf.uni-lj.si/mod/resource/view.php?id=63680 (pridobljeno 19. 9. 2024).

[3] J.A. Rice. *Mathematical Statistics and Data Analysis*. Third Edition. Duxbury, 2007. ISBN: 9780534399429.

[4] Wikipedia. *Sinusoidal model*. URL: http://en.wikipedia.org/w/index.php?title=Sinusoidal%5C%20model&oldid=1176465107 (pridobljeno 17. 9. 2024).

Kvantili Fisherjeve porazdelitve so bili izračunani s kalkulatorjem Stat Trek, dostopnim na https://stattrek.com/online-calculator/f-distribution.