Projektna naloga iz statistike

Lucija Fekonja

August 21, 2023

1. naloga: Kibergrad

Na podlagi enostavnega vzorca 400 enot želimo oceniti povprečen dohodek. Vzorec je enostaven, če iz populacije N enot, v našem primeru je $N = 43886$, izberemo vzorec, ki ga sestavljajo enote $K_1 \dots K_n \in \{1 \dots N\}$, kjer je $n = 400$, pri čemer so vsi možni vzorci enako verjetni. V datoteki `Kibergrad.xlsx` sem enostavni slučani vzorec tvorila tako, da sem vsaki vrstici predpisala naključno število med 1 in 110 $\approx N/n$. Nato sem filtrirala po tem stolpcu in izbrala tisto skupino med 1 in 110, v katero spada natanko 400 vrstic. Če taka skupina ni obstaja, sem postopek ponovila.

Zanimajo nas vrednosti v stolpcu DOHODEK. Na celotni populaciji jih označimo z $x_1 \dots x_N$, vrednosti na vzorcu pa z $X_1 \dots X_n$. Tukaj je $X_i = x_{K_i}$, $i = 1 \dots n$. Ocena za pričakovan dohodek se v teh oznakah poračuna kot:

$$\hat{\mu} = \overline{X} = \frac{X_1 + \dots + X_n}{n} \approx 41047{,}565.$$

Da ocenimo standardno napako naše ocene, moramo oceniti še standardni odklon. Na predavanjih smo izpeljali, da je

$$\hat{\sigma} = \sqrt{\frac{N-1}{N} \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \approx 31604{,}637$$

nepristranska cenilka zanjo kadar iz populacije vzamemo enostavni slučajni vzorec. Nato izračunamo oceno standardne napake

$$\widehat{\mathrm{SE}} = \sqrt{\frac{N-n}{N-1}} \frac{\hat{\sigma}}{\sqrt{n}} = 1573{,}032.$$

Zadnji del (a) naloge nas sprašuje po konstrukciji intervala zaupanja. Ker σ ni znan, ga moramo oceniti, kot smo to storili zgoraj. Za velike n je slučajna spremenljivka $\frac{\overline{X}-\mu}{\hat{\sigma}}\sqrt{n}$ porazdeljena približno standardno normalno. Označimo jo z

$$Z := \frac{\overline{X} - \mu}{\hat{\sigma}} \sqrt{n} \sim \Phi\left(0, 1\right).$$

Sedaj moramo najti tisti δ, za katerega bo veljalo $\overline{X} - \delta \leq \mu \leq \overline{X} + \delta$ v 95% primerov. Pri stopnji tveganja α je to najmanjši δ, za katerega je

$$\begin{aligned} P\left(|\overline{X} - \mu| \geq \delta\right) &\leq \alpha \\ P\left(|Z| \geq \frac{\delta}{\hat{\sigma}}\sqrt{n}\right) &\leq \alpha \\ 2P\left(Z \geq \frac{\delta}{\hat{\sigma}}\sqrt{n}\right) &\leq \alpha \\ 2\left(1 - \Phi\left(\frac{\delta}{\hat{\sigma}}\sqrt{n}\right)\right) &\leq \alpha \\ \frac{\hat{\sigma}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) &\leq \delta. \end{aligned}$$

To pomeni, da je iskani interval zaupanja

$$\overline{X} - \frac{\hat{\sigma}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \mu \leq \overline{X} + \frac{\hat{\sigma}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

oziroma

$$\mu \in [37964{,}423, 44130{,}707].$$

V (b) delu naloge stratificiramo po četrteh. To pomeni, da populacijo razdelimo na stratume, v našem primeru je stratum določena četrt, iz katerih neodvisno izberemo enostavni vzorec. Pri izboru enostavnega vzorca iz vsakega od stratumov sem uporabila isto metodo kot v (a) delu naloge. Naj N_l, $l = 1 \ldots 4$ označuje velikost l-tega stratuma, n_l pa velikost vzorca vzetega iz l-tega stratuma. Velikosti n_l določimo s proporcionalno alokacijo, pri kateri je $n_l = n\frac{N_l}{N} = nW_l$. Tukaj z W_l označujem delež populacije v l-tem stratumu. V našem primeru poračunamo:

$$\begin{aligned} n_1 &= n\frac{10149}{N} \approx 92, \qquad & n_3 &= n\frac{13457}{N} \approx 123, \\ n_2 &= n\frac{10390}{N} \approx 95, \qquad & n_4 &= n\frac{9890}{N} \approx 90. \end{aligned}$$

Pričakovana vrednost dohodka l-tega stratuma je določena z

$$\overline{X}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_{il},$$

kjer X_{il} označuje i-to vrednost v l-tem stratumu. V našem primeru je

$$\begin{aligned} \overline{X}_1 &= 44710{,}065, \qquad & \overline{X}_3 &= 34026{,}423, \\ \overline{X}_2 &= 41644{,}432, \qquad & \overline{X}_4 &= 37798{,}444. \end{aligned}$$

Ocena pričakovanega dohodka populacije je

$$\overline{X} = \sum_{l=1}^{4} \frac{N_l}{N}\overline{X}_l = \sum_{l=1}^{4} W_l\overline{X}_l = 39150{,}715.$$

Da dobljen rezultat primerjamo z enostavnim slučajnim vzorčenjem, je potrebno pogledati standardno napako ocene. Zanjo pa potrebujemo varianco vsakega od stratumov, ki jo izračunamo na naslednji način:

$$\hat{\sigma}_l^2 = \frac{N_l - 1}{N_l} \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \left(X_{il} - \overline{X}_l \right)^2 .$$

Po danih podatkih je

$$\begin{aligned} \hat{\sigma}_1 &= 32635{,}808, \qquad & \hat{\sigma}_3 &= 26916{,}903, \\ \hat{\sigma}_2 &= 39578{,}556, \qquad & \hat{\sigma}_4 &= 22674{,}367. \end{aligned}$$

Varianca povprečnega dohodka populacije stratificirane s proporcionalno alokacijo je izračunana kot

$$Var\left(\overline{X}\right) = \sum_{l=1}^{4} \frac{W_l^2 \hat{\sigma}_l^2}{n_l}.$$

Potem je standarna napaka ocene povprečja enaka

$$\widehat{SE} = \sqrt{Var\left(\overline{X}\right)} = 1545{,}099.$$

Standardna napaka je pri stratificiranem vzorčenju manjša kot pri enostanvem slučajnem vzorčenju, kar pomeni, da smo v (b) primeru dobili nekoliko boljšo oceno za pričakovan dohodek. Stratificirano vzorčenje da boljšo oceno za pričakovano vrednost, če se vrednosti znotraj posameznih stratumov ne razlikujejo veliko, medtem ko so med različnimi stratumi razlike velike. V dani populaciji Kibergrada se dohodki znotraj stratumov precej razlikujejo, prav tako pa med različnimi stratumi težko ločimo. Zato ocena s stratificiranim vzorčenjem ni veliko boljša od ocene pri enostavnem vzorčenju.

2. naloga: Kevlar

Primerjalni kvantilni (Q–Q) grafikon je v splošnem sestavljen iz točk (x, y), kjer x teče po vseh kvantilih prve, y pa po vseh kvantilih druge porazdelitve za vsako verjetnost $\alpha \in (0, 1)$. Ko primerjamo teoretično in empirično porazdelitev prikažemo le kvantile za verjetnosti $\frac{1}{n+1} \ldots \frac{n}{n+1}$, kjer je n število opaženih podatkov. Opažanja uredimo v ražirno vrsto $X_{\frac{1}{n+1}} \leq \cdots \leq X_{\frac{n}{n+1}}$, kar je v `Kevlar.xlsx` že dano.

Naloga od nas zahteva, da empirično porazdelitev primerjamo z eksponentno. Za prikaz primerjalnega kvantilnega grafikona bomo potrebovali kumulativno funkcijo eksponentne porazdelitve, to je

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Vzemimo $\lambda = 1$. Naj bodo $X_1 \dots X_n, n = 101$, dana opažanja trajanja vlaken pri 90% obremenitvi, preden je prišlo do porušitve. Za vsak $k = 1 \dots n$ narišemo točke

$$\left(\frac{k}{n+1}, F\left(X_k\right) \right).$$

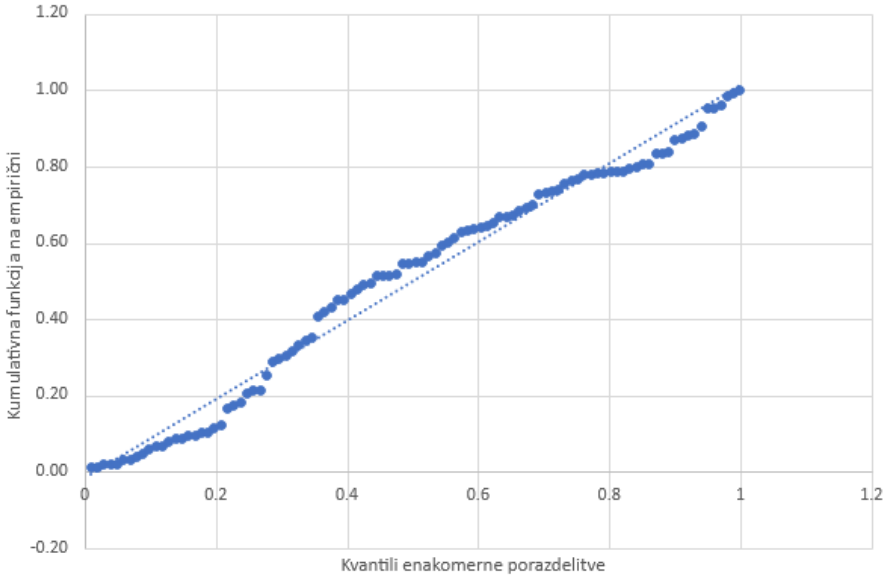

Figure 1: Primerjalni kvantilni (Q–Q) grafikon, ki enakomerno porazdeljenim n točkam na intervalu $(0, 1)$ priredi vrednosti kumulativne funkcije eksponentne porazdelitve.

Vidimo, da se empirični podatki ujemajo z eksponentno porazdelitvijo, saj je dobljen graf skoraj linearen. Črta, ki se podatkom najbolje prilega ima enačbo $y = 1{,}0228x - 0{,}0197$. Oglejmo si še ekvivalentno konstrukcijo primerjalnega kvantilnega grafikona, v kateri za vsak $k = 1 \dots n$ narišemo točke

$$\left(F^{-1}\left(\frac{k}{n+1} \right), X_k \right).$$

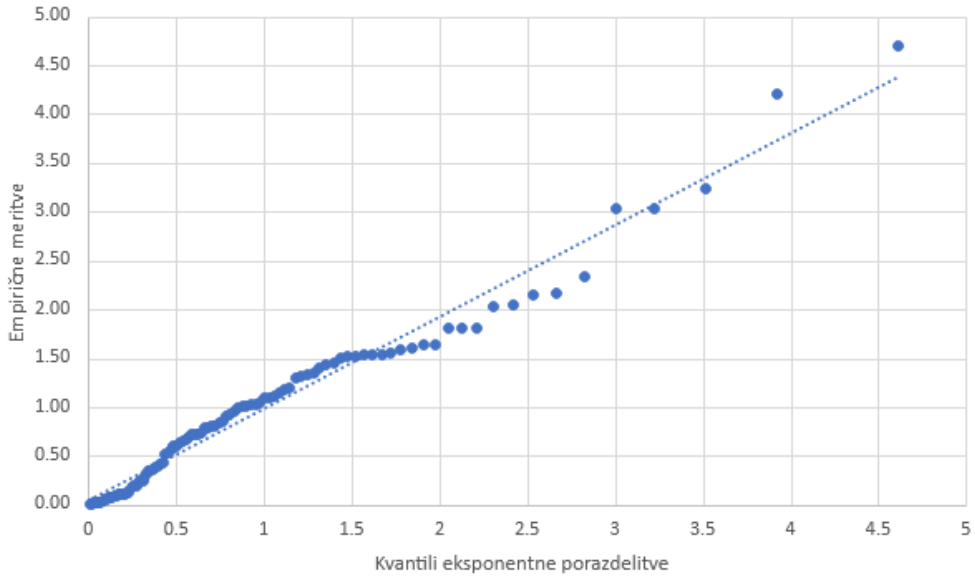

Figure 2: Primerjalni kvantilni (Q–Q) grafikon predstavljen s točkami $\left(F^{-1}\left(\frac{k}{n+1}\right), X_k\right)$.

Tudi v tem primeru se empirična porazdelitev ujema z eksponentno, saj je črta, ki se podatkom najbolje prilega linearna. Njena enačba je $y = 0{,}9401x - 0{,}0361$. Sklepamo, da je trajanje vlaken, preden je prišlo do porušitve, porazdeljeno eksponentno. Namreč ko primerjamo dve eksponentni porazdelitvi, je primerjalni kvantilni grafikon linearen. V bistvu je Q–Q grafikon linearen, če primerjamo dve enaki porazdelitvi.

V nadaljevanju nas naloga sprašuje po eksponentni porazdelitvi, ki se najbolj prilega danim podatkom. Poiskati moramo torej λ v gostoti eksponentne porazdelitve

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

po metodi največjega verjetja. Predpostavimo, da so opažanja $x_1 \dots x_n$ neodvisna. Potem je verjetje definirano kot:

$$L(\lambda|x) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

Če verjetje logaritmiramo, dobimo

$$l(\lambda|x) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i.$$

Iščemo maksimum verjetja L oz. l, zato odvajamo l po λ, kar nam da $\frac{dl}{d\lambda}(\lambda|x) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$, iz česar izrazimo

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i} = 0{,}976695.$$

Porazdelitev, ki se danim podatkom najbolje prilega je torej *Exp(*0,976695*)*.

Da ugotovimo, če se empirična porazdelitev res prilega eksponentni, lahko narišemo tudi viseči histogram iz razlik korenov frekvenc. Pri risanju histograma podatke razdelimo v razrede, katerih širino lahko izračunamo po npr. modificiranem Freedman–Diaconisovem pravilu, pri katerem je

$$\text{širina} = \frac{2.6 \cdot \text{IQR}}{\sqrt[3]{n}}.$$

V našem primeru je interkvartilni razmik enak IQR $= x_{3/4} - x_{1/4} = 1{,}455 - 0{,}235 = 1{,}22$, kjer sta $x_{1/4}$ in $x_{3/4}$ prvi oz. tretji kvartil. Širina intervala po zgoraj navedenem pravilu je torej 0,242789074. Naj bo sedaj x_{j-1} leva meja j-tega intervala, x_j pa desna. Verjetnost, da opažanje pade na j-ti interval je

$$\hat{p_j} = \int_{\frac{x_{j-1}-\hat{\mu}}{\hat{\sigma}}}^{\frac{x_j-\hat{\mu}}{\hat{\sigma}}} \lambda e^{-\lambda x} dx,$$

kjer sta cenilki $\hat{\mu}$ in $\hat{\sigma}$ izračunani z

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{in} \qquad \hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

Pričakovano število opažanj na j-tem intervalu je $\hat{n}_j = n\hat{p_j}$. Označimo z n_j dejansko število opažanj na j-tem intervalu. Pri visečem histogramu nas zanima razlika $n_j - \hat{n_j}$, pri visečem histogramu iz razlik korenov frekvenc pa $\sqrt{n_j} - \sqrt{\hat{n_j}}$. Narišimo oba viseča histograma in podajmo tabeli razlik frekvenc oziroma razlik njihovih korenov.

Sredina intervala	$n_j - \hat{n_j}$
0.3406	-151.6958539
1.0217	-78.90811836
1.7028	-41.50131883
2.3839	-28.80945688
3.0651	-15.09137051
3.7462	-9.975711295
4.4273	-3.500678381
5.1084	-3.033113146
5.7896	-1.672480053
6.4707	-0.922217289
7.1518	-0.508517086
7.8329	0.719600124
8.514	-0.154614444

Table 1: Tabela vrednosti $n_j - \hat{n_j}$.

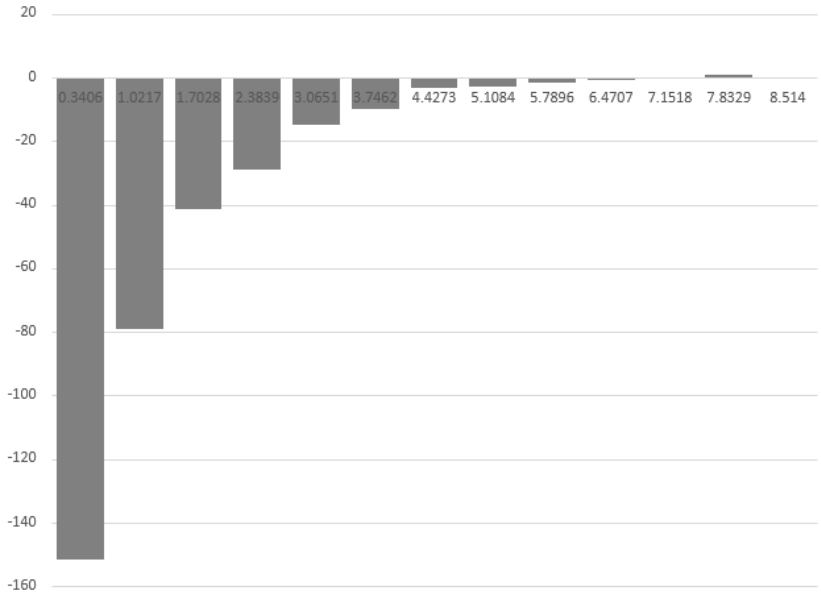

Figure 3: Viseči histogram.

Sredina intervala	$\sqrt{n_j} - \sqrt{\hat{n}_j}$
0.3406	-7.355883841
1.0217	-5.00271844
1.7028	-3.47106911
2.3839	-3.727953988
3.0651	-2.521344366
3.7462	-3.158434944
4.4273	-0.931138944
5.1084	-1.741583517
5.7896	-1.293244004
6.4707	-0.960321451
7.1518	-0.71310384
7.8329	0.470472025
8.514	-0.393210432

Table 2: Tabela vrednosti $\sqrt{n_j} - \sqrt{\hat{n}_j}$.

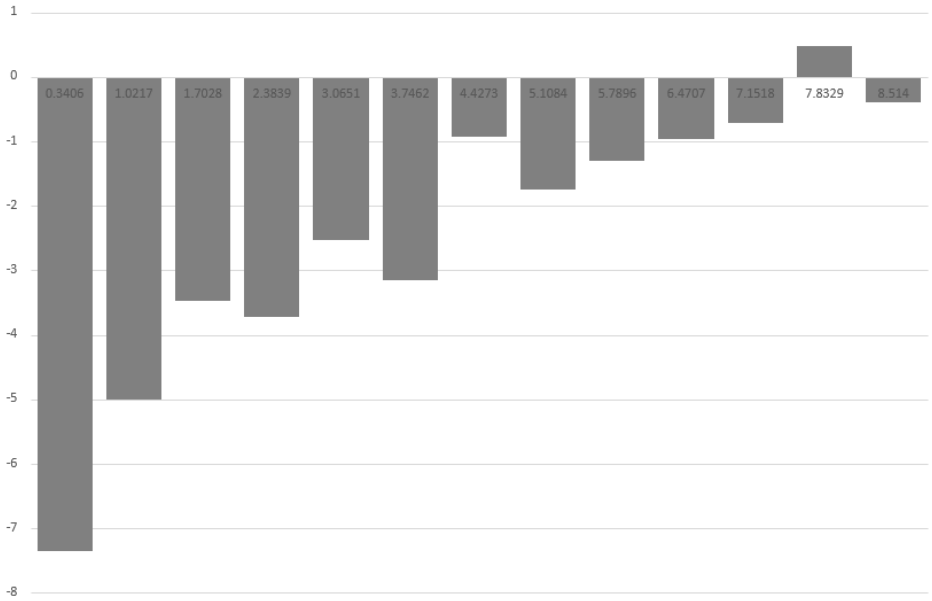

Figure 4: Viseči histogram razlik korenov frekvenc.

Prednost visečega histograma razlik korenov frekvenc je, da se sosednji stolpci smejo v velikosti razlikovati le malo, zato lahko hitro opazimo odstopanja. V našem primeru vidimo velika odstopanja na levi strani histograma. Pove nam, da smo opazili manj vrednosti kot smo jih pričakovali na danem intervalu. Še eno očitno odstopanje nastopi na desni, saj imamo dano eno trajanje 7,89, kar je precej večje od povprečja.

Na koncu predstavimo rezultate še na logaritemski lestvici. Primerjalni kvantilni grafikon, v katerem primerjamo kvantile enakomerne porazdelitve s $F(X_k)$ je na logaritemski lestvici videti tako:

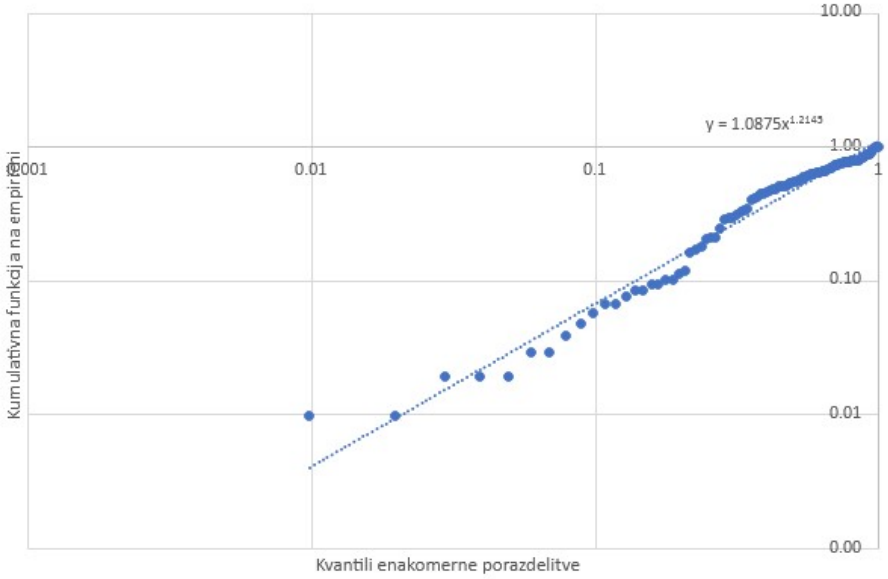

Figure 5: Primerjalni kvantilni (Q–Q) grafikon na logaritemski lestvici.

Črta, ki se podatkom najbolje prilega je $y = 1{,}0875x^{1{,}2145}$. V primeru, ko rišemo točke $\left(F^{-1}\left(\frac{k}{n+1}\right), X_k\right)$, pa je črta, ki se podatkom najbolje prilega na logaritemski lestvici $y = 0{,}9665x^{1{,}1465}$, Q–Q grafikon pa je videti tako:

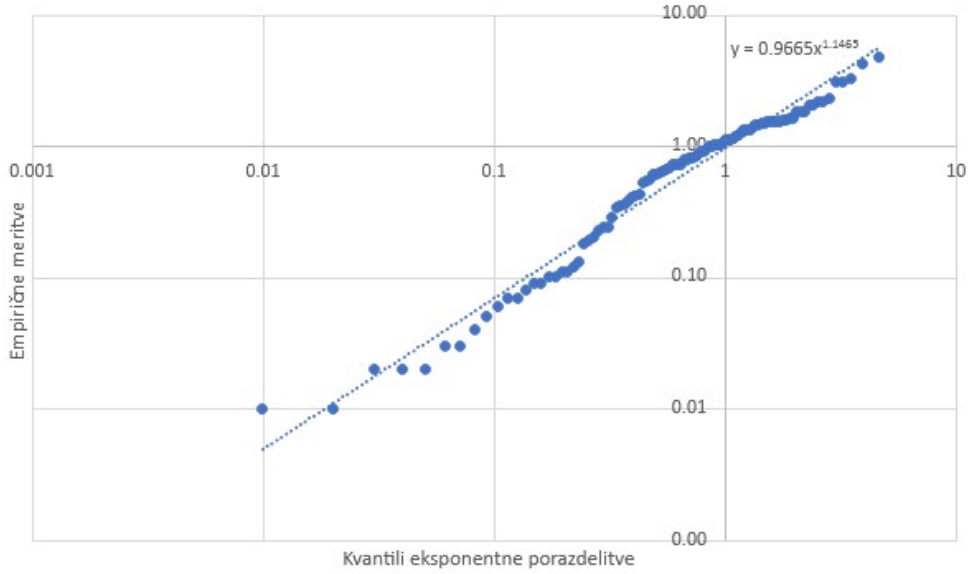

Figure 6: Primerjalni kvantilni (Q–Q) grafikon v drugem primeru na logaritemski lestvici.

Ko s primerjalnim kvantilnim grafikonom primerjamo dve eksponentni porazdelitvi dobimo tudi na logaritemski lestvici prileganje, ki zgleda linearno, v kolikor transformiramo tako abscizno kot tudi ordinatno os.

3. naloga: Temperature v Ljubljani

V tej nalogi bomo konstruirali model linearne regresije in z njegovo pomočjo sklepali o segrevanju podnebja. Enostavna linearna regresija se ravna po predlogi $y = a + bx$, kjer je x pojasnjevalna spremenljivka, y odvisna spremenljivka, a in b pa iskana koeficienta. V našem primeru bomo najprej z X_i označevali leto, z Y_i pa povprečno temperaturo tega leta. Te podatke združimo v tabelo.

Leto	Povprečna temperatura
1986	9.533333333
1987	9.616666667
1988	10.525
1989	10.45
1990	10.675
1991	9.95
1992	11.13333333
1993	10.59166667
1994	11.825
1995	10.75833333
1996	9.775
1997	10.775
1998	11.01666667
1999	10.975
2000	12.175
2001	11.39166667
2002	11.80833333
2003	11.575
2004	10.66666667
2005	10.425
2006	11.4
2007	12.05
2008	11.56666667
2009	11.68333333
2010	10.68333333
2011	11.75
2012	12
2013	11.6
2014	12.64166667
2015	12.15
2016	11.75833333
2017	11.91666667
2018	12.48333333
2019	12.54166667
2020	12.11666667

Table 3: Tabela povprečnih temperatur v letih od 1986 do 2020.

Model enostavne linearne regresije lahko zapišemo tudi z matrično enačbo $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Tukaj je Y slučajni, opažen vektor, X deterministična matrika, β determinističen, a neznan vektor parametrov, ϵ pa slučajni in neznan vektor

rezidualov. Matrična oblika modela enostavne linearna regresije je

$$\begin{bmatrix} Y_1 \\ ... \\ Yn \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ ... & ... \\ 1 & X_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ ... \\ \epsilon_n \end{bmatrix}. \tag{1}$$

Najboljša ocena za iskani β v splošem je

$$\hat{\beta} = \left(\mathbf{X^T X}\right)^{-1} \mathbf{X^T Y},$$

ko upoštevamo 1 pa se $\hat{\beta}$ posploši na

$$\hat{\beta} = \frac{1}{n\sum X_i^2 - \left(\sum X_i\right)^2} \begin{bmatrix} \sum X_i^2 \cdot \sum Y_i - \sum X_i \cdot \sum X_i Y_i \\ -\sum X_i \cdot \sum Y_i + n \sum X_i Y_i \end{bmatrix}, \tag{2}$$

kjer vsote tečejo po $i = 1 \ldots n$. Pri danih podatkih dobimo

$$\hat{\beta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} -118{,}2089496 \\ 0{,}064635854 \end{bmatrix},$$

kjer $\hat{b}$ predstavlja spremembo temperature v enem letu. Iz tega sledi, da bo temperatura narastla za 1 stopinjo v približno 15,5 letih. Ocenimo standardno napako ocene naraščanja temperature $\hat{b}$. Najprej izračunamo

$$\hat{\sigma}^2 = \frac{1}{n-2}||Y - X\hat{b}||^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\epsilon}_i^2,$$

kjer je $\hat{\epsilon}_i = Y_i - \hat{a} - \hat{b}X_i$. Dobimo $\hat{\sigma} = 0{,}52429844$. Če predpostavimo, da je šum porazdeljen večrazsežno normalno, sta cenilki $\hat{\beta}$ in $\hat{\sigma}^2$ neodvisni in velja

$$\frac{\mathbf{c^T}\hat{\beta} - \mathbf{c^T}\beta}{\hat{\sigma}\sqrt{\mathbf{c^T}\left(\mathbf{X^T X}\right)^{-1}\mathbf{c}}} \approx \text{Student}\left(n - m\right),$$

kjer je $\mathbf{c}$ deterministično vektor. Ko ocenjujemo b upoštevamo, da je $\mathbf{c} = (\mathbf{0}, \mathbf{1})$ in dobimo

$$\frac{\hat{b} - b}{\hat{\sigma}}\sqrt{\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n}} \approx Student(n-2).$$

Tako dobimo oceno standardne napake naraščanja temperature:

$$\Delta_b = \text{Student}_{1-\frac{\alpha}{2}}\left(n-2\right)\frac{\hat{\sigma}}{\sqrt{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2}} \approx 0{,}018126. \tag{3}$$

Ocena za strmino naraščanja temperature $\hat{b}$ je pozitivna, njena standardna napaka pa majhna, zato ni dvoma, da je linearni trend segrevanja statistično značilen. V splošnem se standardna napaka izračuna kot

$$\text{SE}\left(\mathbf{c^T}\hat{\beta}\right) = \text{Student}_{1+\frac{\alpha}{2}}\left(n-m\right)\hat{\sigma}\sqrt{\mathbf{c^T}\left(\mathbf{X^T X}\right)^{-1}\mathbf{c}}.$$

Za standardno napako ocene koeficienta a vstavimo $\mathbf{c} = (\mathbf{1}, \mathbf{0})$ in jo izračunamo kot

$$\Delta_a = \text{Student}_{1-\frac{\alpha}{2}}\,(n-2)\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(\frac{1}{n}\sum X_i\right)^2}{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2}} \approx 36{,}30595 \qquad (4)$$

Napovejmo še povprečno temperaturo leta 2040 in zanjo izračunajmo napovedni interval pri stopnji tveganja $\alpha = 0{,}05$. Točkasta napoved za Y je

$$\hat{Y} = \hat{a} + \hat{b}X = 13,6482, \qquad (5)$$

kjer je $X = 2040$. Napovedni interval je potem enak $\hat{Y} - \Delta \leq Y \leq \hat{Y} + \Delta$, kjer je

$$\Delta = \text{Student}_{1-\frac{\alpha}{2}}\,(n-2)\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(X - \frac{1}{n}\sum X_i\right)^2}{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2}}. \qquad (6)$$

V našem primeru izračunamo $\Delta = 1{,}2869$, torej je interval zaupanja enak $[12{,}3613, 14{,}9351]$.

V drugem delu naloge bomo konstruirali model linearne regresija, pri čemer bomo vključili tudi letno nihanje temperature. V bistvu bomo konstruirali dvanajst modelov linearne regresije – za vsak mesec enega – podobno kot smo to storili zgoraj. Naslednja tabela nam podaja povprečne mesečne temperature v letih od 1986 do 2020.

Leto	Jan	Feb	Mar	Apr	Maj	Jun	Jul	Avg	Sep	Okt	Nov	Dec
1986	0.1	-2.8	3.2	10.2	17.6	17.4	19.6	20.2	14.7	10.2	5.5	-1.5
1987	-3.4	0.5	1.1	11.1	13.3	17.8	21.4	18.7	18.3	11.2	4.5	0.9
1988	3.8	3.4	5.4	10.4	15.3	17.4	22	20.8	15.5	11.5	0.9	-0.1
1989	-0.7	4.1	8.5	10.6	14.9	16.5	20.2	19.7	15.5	10.2	3.5	2.4
1990	-0.5	5.7	8.8	9.1	16.2	17.7	20.2	20	14.1	11.3	5.1	0.4
1991	0.7	-1.4	8.4	9.4	12.1	18.2	21.8	20.6	17.5	9.2	5.1	-2.2
1992	0.6	3.4	6.3	10.8	16.1	18.5	21.2	23.7	16.3	9.6	6.6	0.5
1993	0.9	1	5.8	11.2	17	19.2	20.4	20.9	14.9	11.4	2.5	1.9
1994	3.5	2.6	10.6	10.1	15.3	19.3	22.5	22.1	17.1	8.9	7.7	2.2
1995	1	4.5	5.1	11.3	15.1	17	22.8	19.1	14.3	12.3	5.3	1.3
1996	-0.8	-0.9	3.4	10.6	16	19.7	19.1	19.5	13.2	11	7.5	-1
1997	-0.5	4.1	7.5	8.5	16.3	19	20.1	20.2	16.6	9.6	5.3	2.6
1998	3.2	5.3	5.8	11	15.8	20.7	21.5	21.6	15.6	11.4	3.4	-3.1
1999	0.6	0.8	7.8	11.6	16.7	19.1	20.9	20.5	18.1	11.8	3.1	0.7
2000	-1.6	4	7.6	13.6	17	20.9	19.9	22.1	16.4	12.9	8.4	4.9
2001	3.4	4.7	8.8	10.1	17.2	18.4	21.9	22.9	13.8	14	3.6	-2.1
2002	-0.6	5	8.9	10.1	17.3	21.1	21.3	20.1	15	11.5	9.4	2.6
2003	-1.1	-0.9	7.4	10.3	18.3	23.5	22.6	24.2	15.5	8.8	8.2	2.1
2004	-0.3	2.2	5	10.7	14	18.8	20.9	20.7	15.6	13	5.9	1.5
2005	0.1	-0.3	5.7	10.8	16.3	19.5	21.1	18.4	16.4	11.8	5.1	0.2
2006	-1.6	0.5	4.5	11.5	15.5	20.5	23.6	17.7	17.7	13.4	8.9	4.6
2007	4.9	5.9	8.5	14.7	17.2	20.9	22	20.4	14.5	10.4	5.1	0.1
2008	2.5	4.6	6.2	10.7	16.9	20.3	21.4	20.7	15.1	12	6.4	2
2009	-1.5	2.3	7.1	13.3	18.1	18.9	21.7	22.4	17.4	11	7.5	2
2010	-1.5	1.3	6.2	11.5	15.3	20.3	22.9	20.3	14.7	9.5	8.1	-0.4
2011	1.5	1.5	7.1	13.5	17	20	21.1	22.8	19.4	10	3.8	3.3
2012	1.6	-0.8	10.1	11.4	16.1	21.3	22.7	23.3	17	11.7	8.8	0.8
2013	2	0.9	3.9	12.4	14.8	19.8	23.5	22.5	16.2	13.2	7.3	2.7
2014	5.4	4.4	10	13.1	15.7	20.2	20.8	19.6	16.2	13.6	8.8	3.9
2015	2.8	2.4	7.6	11.8	17	20.6	24.3	22.3	16.5	11	6.9	2.6
2016	1.1	5.5	7.5	12.5	15.3	20	23.2	20.6	18.3	10.3	7	-0.2
2017	-3.2	4.5	10.2	12.1	16.9	21.7	23.2	23.2	14.3	12	6.2	1.9
2018	4.8	-0.1	4.6	15.2	18	20.9	22.3	22.8	17.6	13.2	8.3	2.2
2019	0.7	4.9	9	11.6	12.9	23.5	22.9	22.6	16.8	13.2	8.8	3.6
2020	1.9	6.8	7.2	13	15.3	19.6	21.8	22.2	17.5	11.9	5.3	2.9

Table 4: Tabela povprečnih mesečnih temperatur v letih od 1986 do 2020.

Tako kot prej bomo s formulo 2 ocenili koeficiente za vsakega od modelov. Označimo z β_j koeficiente modela j-tega meseca. Če sedaj X_{ij} označuje j-ti mesec i-tega leta, Y_{ij} pa pripadajočo temperaturo, oceno za koeficiente izračunamo z

$$\hat{\beta_j} = \frac{1}{n\sum X_{ij}^2 - (\sum X_{ij})^2} \begin{bmatrix} \sum X_{ij}^2 \cdot \sum Y_{ij} - \sum X_{ij} \cdot \sum X_{ij}Y_{ij} \\ -\sum X_iij \cdot \sum Y_{ij} + n\sum X_{ij}Y_{ij} \end{bmatrix},$$

kjer vse vsote spet tečejo po letih $i = 1 \dots n$, $n = 35$. V naslednji tabeli so zbrane ocene koeficientov $\hat{\beta}_i = \begin{bmatrix} \hat{a_i} \\ \hat{b_i} \end{bmatrix}$, ter standardne napake teh ocen po formulah 3 in 4.

	Jan	**Feb**	**Mar**	**Apr**
$\hat{a_i}$	-94.19288515	-100.2831092	-114.0293838	-173.6722969
$\hat{b_i}$	0.04745098	0.051344538	0.060364146	0.092408964
$SE(\hat{a_i})$	855.482328	963.9493216	839.0534516	461.6688106
$SE(\hat{b_i})$	0.427095084	0.481246664	0.418893053	0.230485742
	Maj	**Jun**	**Jul**	**Avg**
$\hat{a_i}$	-23.22411765	-197.3568627	-113.4804202	-95.12683473
$\hat{b_i}$	0.019579832	0.108347339	0.067478992	0.058039216
$SE(\hat{a_i})$	563.4320175	468.8558869	406.9318144	590.9713987
$SE(\hat{b_i})$	0.281290492	0.234073853	0.203158583	0.295039384
	Sep	**Okt**	**Nov**	**Dec**
$\hat{a_i}$	-62.89501401	-91.2472549	-202.607395	-150.3918207
$\hat{b_i}$	0.039439776	0.051232493	0.104201681	0.075742297
$SE(\hat{a_i})$	563.7060539	521.614813	727.0058193	690.5996253
$SE(\hat{b_i})$	0.281427303	0.260413471	0.362953858	0.344778256

Table 5: Ocene koeficientov in njihove standardne napake

Hitro vidimo, da so napake koeficientov v tem primeru precej večje kot v prejšnjem modelu. To nam pove, da temperature v, na primer, januarju skozi leta nihajo in je zanje težko predpisati linearno črto, h kateri strmijo, kot smo to naredili za povprečne letne temperature. Kljub temu dajmo napovedati pričakovano januarsko temperaturo za leto 2040, kot to od nas zahteva naloga. Z uporabo formul 5 in 6 izračunamo, da leta 2040 v januarju pričakujemo 2,6071 stopinj, z verjetnostjo 0,95 pa bo povprečna temperatura padla med 2,0856 in 3,1287 stopinj.