Matematika i statistika

Doktorski studij Strojarskog fakulteta

Sveučilište u Slavonskom Brodu

Predavanje 4

Predavač:
izv.prof.dr.sc. Nenad Šuvak, Odjel za matematiku, Sveučilište u Osijeku

VEZA IZMEĐU VARIJABLI

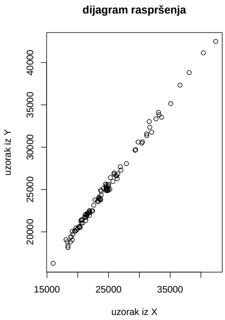
- za parove podataka iz dvije neprekidne varijable želimo zaključivati o postojanju zavisnosti između njih
- kod neprekidnih varijabli zavisnost se može pojaviti na brojne načine - različiti tipovi veza među varijablama

Deterministička veza

- **deterministička veza** između dvije varijable zadana pravilom
$$y = f(x)$$
gdje je y zavisna varijabla, x nezavisna varijabla, a $f : \mathbb{R} \to \mathbb{R}$ zadana funkcija
- primjerice, pravilima $y = x + 54$, $y = x^2 - 14x$ i $y = \sin(3x)$ zadane su determinističke veze među varijablama x i y
- za svaku dopuštenu vrijednost nezavisne varijable x možemo izračunati točnu vrijednost zavisne varijable y

Statistički model s aditivnom greškom

- u praktičnim problemima ne možemo očekivati determinističku vezu
- **dijagram raspršenja** podataka (eng. scatter plot) je prikaz uređenih parova podataka iz dviju varijabli u koordinatnom sustavu



Statistički model s aditivnom greškom

- **regresijska metoda** modeliranja pretpostavlja da možemo uspostaviti funkcijsku vezu ali uz dodanu grešku
- veza između nezavisne varijable x i zavisne slučajne varijable $Y(x)$ će biti oblika

$$Y(x) = f(x) + \varepsilon, \tag{1}$$

  gdje pretpostavljamo da je ε slučajna varijabla koja opisuje grešku u modeliranju
- mnogo nezavisnih slučajnih smetnji u pravilu ima normalnu distribuciju – u primjenama se u klasičnom načinu modeliranja prihvaća da je model adekvatan ako je u njemu postignuta normalna distribuiranost grešaka ε

Statistički model s aditivnom greškom

- sparena mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ dvaju obilježja koja dolaze od slučajnih varijabli $Y_1, \ldots Y_n$ (čije su realizacije realni brojevi $y_1, \ldots y_n$) i nezavisne varijable x (čije su izmjerene vrijednosti $x_1, \ldots x_n$)
- cilj je utvrditi zavisnost između dvije varijable
- **regresijski model** – matematički model oblika

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdje je f realna funkcija jedne realne varijable, a $\varepsilon_1, \ldots, \varepsilon_n$ međusobno nezavisne slučajne varijable takve da je $E\,\varepsilon_i = 0$ i $\mathrm{Var}(\varepsilon_i) = \sigma^2$

Statistički model s aditivnom greškom

- prvi korak u uspostavljanju ovakvih veza među varijablama Y i x prikaz je podataka u dijagramu raspršenja iz kojeg se lako vidi grupiraju li se sparena mjerenja oko pravca (linearna zavisnost) ili neke krivulje (neka druga funkcijska zavisnost - polinomijalna, logaritamska, ...)

Regresijski pravac

- pretpostavimo da je graf funkcije f u modelu pravac
- f možemo prikazati formulom $f(x) = \alpha + \beta x$
- slobodni koeficijent α zove se **odsječak na y-osi**, a koeficijent β uz nezavisnu varijablu x zove se **koeficijent smjera** i važan je iz sljedećeg razloga:

  ako je $\beta < 0$ funkcija $f(x) = \alpha + \beta x$ je padajuća
  ako je $\beta > 0$ funkcija $f(x) = \alpha + \beta x$ je rastuća
- graf funkcije $f(x) = \alpha + \beta x$ nazivamo **regresijskim pravcem**, a koeficijente α i β **regresijskim parametrima**

LINEARNI REGRESIJSKI MODEL

linearni regresijski model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

- $x_1, x_2, \dots, x_n$ – izmjerene vrijednosti varijable x
- $Y_1, Y_2 \dots, Y_n$ – slučajne varijable s izmjerenim vrijednostima $y_1, \dots, y_n$)
- α i β – **nepoznati parametri** linearne veze koje treba **procijeniti** u postupku modeliranja (to zapravo znači da trebamo **procijeniti regresijski pravac** $y = \alpha + \beta x$)

Linearni regresijski model

- $\varepsilon_1, \ldots, \varepsilon_n$ predstavljaju varijable greške koja je dodana na linearnu vezu $(\alpha + \beta x_i)$ – nemjerljive slučajne varijable za koje pretpostavljamo da
  - međusobno su nezavisne
  - sve imaju normalnu distribuciju
  - imaju očekivanje 0
  - sve imaju jednaku varijancu σ^2

METODA NAJMANJIH KVADRATA

- na osnovu podataka želimo procijeniti nepoznate parametre α i β
- tako ćemo dobiti i procjenu nepoznatog regresijskog pravca $y = \alpha + \beta x$
- ako su α i β poznati za svaku izmjerenu vrijednost x_i možemo odrediti broj

$$y_i' = \alpha + \beta x_i$$

- y_i' – **teorijska vrijednost** zavisne varijable u x_i (eng. predicted value)
- y_i – **izmjerena** ili **eksperimentalna** vrijednost zavisne varijable u x_i (eng. observed value)
- y_i se razlikuje od teorijske vrijednosti pa točke (x_i, y_i), $i = 1, \ldots, n$, uglavnom ne leže na pravcu $y = \alpha + \beta x$

METODA NAJMANJIH KVADRATA

- parametre ćemo odrediti tako da razlike između izmjerenih i teorijskih vrijednosti budu što manje
- metoda koju koristimo naziva se **metoda najmanjih kvadrata**
- procjenu treba odrediti tako da funkciju

$$\begin{aligned} D(\alpha, \beta) &= \sum (\text{eksperimentalne v.} \; - \; \text{teorijske v.})^2 \\ &= \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \end{aligned}$$

  učinimo što manjom
- procjene $\widehat{\alpha}$ i $\widehat{\beta}$ regresijskih parametara α i β trebamo odrediti tako da vrijedi:

$$D(\widehat{\alpha}, \widehat{\beta}) = \min_{(\alpha,\beta)\in\mathbb{R}^2} D(\alpha, \beta)$$

METODA NAJMANJIH KVADRATA

- procjene $\widehat{\alpha}$ i $\widehat{\beta}$ nazivamo **procjenama u smislu metode najmanjih kvadrata** (eng. least squares estimates) regresijskih parametara α i β
- procjena nepoznatog regresijskog pravca $y = \alpha + \beta x$ je pravac $y = \widehat{\alpha} + \widehat{\beta} x$

METODA NAJMANJIH KVADRATA

- procjene se mogu eksplicitno izraziti:

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2}, \quad \widehat{\alpha} = \overline{y}_n - \widehat{\beta}\,\overline{x}_n,$$

gdje su

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2, \quad s_y^2 = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2,$$

$$s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)$$

METODA NAJMANJIH KVADRATA

- koristeći formulu procijenjenog regresijskog pravca $y = \widehat{\alpha} + \widehat{\beta}x$, za svaki vrijednost x možemo izračunati pripadnu procjenu teorijske vrijednosti – **predikcija**
- $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ — predikcija zavisne varijable za vrijednost x_i nezavisne varijable
- odstupanje procijenjene vrijednosti $\widehat{y}_i$ od izmjerene vrijednosti y_i zavisne varijable:

$$e_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta}x_i)$$

- $e_1, \ldots, e_n$ zovemo **rezidualima** i možemo ih smatrati procjenama grešaka $\varepsilon_1, \ldots, \varepsilon_n$ iz modela $Y_i = \alpha + \beta x_i + \varepsilon_i$

METODA NAJMANJIH KVADRATA

- suma kvadrata svih reziduala upravo je minimalna postignuta vrijednost za D, tj. $D(\widehat{\alpha}, \widehat{\beta})$, i predstavlja jednu mjeru kvalitete modela koju označavamo SSE (sum of squares of errors):

$$SSE = \sum_{i=1}^{n} e_i^2.$$

- R - primjer 1

Statističko zaključivanje

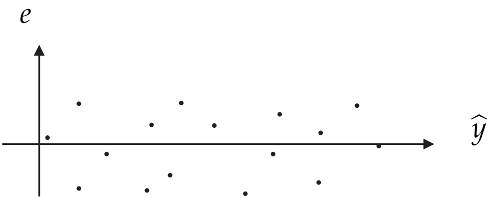
- da bismo mogli koristiti dobiveni model potrebno je napraviti analizu prihvatljivosti modela
- istražujemo jesu li ispunjene osnovne pretpostavke klasičnog regresijskog modela: greške modela trebaju biti međusobno nezavisne i jednako distribuirane slučajne varijable s normalnom distribucijom
- dio analize modela koji se provodi u tu svrhu obično se naziva **analiza reziduala**

Analiza reziduala

Jednakost varijanci grešaka

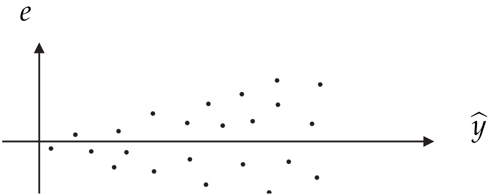
- imaju li $\varepsilon_1, \dots \varepsilon_n$ jednake varijance? - homogenost varijanci grešaka
- zaključujemo na temelju procjena grešaka modela – reziduala $e_1, \dots, e_n$
- grafički prikažemo reziduale u ovisnosti o predikcijama – dijagram raspršenosti za točke $(\widehat{y}_i, e_i)$, $i = 1, \dots, n$
- ako u tom dijagramu uočavamo sustavno povećanje ili smanjenje raspršenosti, to je znak da varijance nisu homogene

Analiza reziduala



parovi $(\widehat{y}_i, e_i)$ koji sugeriraju homogenost varijanci reziduala

ANALIZA REZIDUALA



ovakav raspored parova $(\widehat{y}_i, e_i)$ sugerira stalan rast varijance, dakle varijance nisu homogene

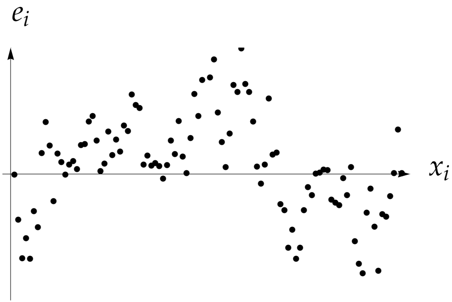
R - primjer 1

Analiza reziduala

Nezavisnost grešaka

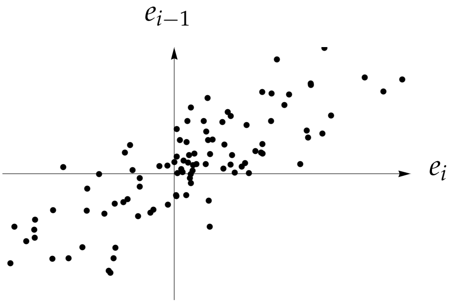
- jesu li $\varepsilon_1, \dots \varepsilon_n$ nezavisne?
- zavisnost grešaka može se manifestirati na razne načine
- koristit ćemo dvije grafičke metode:
  - dijagram raspršenosti reziduala u odnosu na vrijednosti nezavisne varijable
  - dijagram raspršenosti parova susjednih reziduala, tj. parova (e_i, e_{i-1}), $i = 2, \dots, n$
- ako nema nikakve pravilnosti u izgledu dijagrama, nemamo razloga sumnjati u nezavisnost

Analiza reziduala



ovakav raspored parova (x_i, e_i) sugerira zavisnost grešaka modela

Analiza reziduala



ovakav raspored parova (e_i, e_{i-1}) sugerira zavisnost grešaka modela

R - primjer 1

Analiza reziduala

Normalna distribuiranost grešaka

- jesu li $\varepsilon_1, \ldots \varepsilon_n$ normalno distribuirane?
- možemo provjeriti KS testom i Shapiro-Wilk testom na rezidualima $e_1, \ldots, e_n$
- nije nužan uvjet, ali ukoliko ne vrijedi treba biti oprezan u statističkom zaključivanju o modelu

ako nemamo razloga sumnjati u ispravnost pretpostavki modela, možemo ga koristiti za zaključivanje o vezi između nezavisne i zavisne varijable

R - primjer 1

Zaključivanje o koeficijentu smjera regresijskog pravca

- je li model $Y_i = \alpha + \beta x_i + \varepsilon_i$ bolji od nul-modela $Y_i = \alpha + \varepsilon_i$ (modela u kojemu je $\beta = 0$)?
- koji od dva modela bolje opisuje promjene u očekivanju slučajnih varijabli Y_i u ovisnosti o vrijednostima x_i?
- ako je $\beta = 0$, takav regresijski pravac bio bi paralelan s x-osi pa promjena vrijednosti nezavisne varijable ne bi rezultirala promjenom očekivanja zavisne varijable
- to možemo utvrditi statističkim testom čije su hipoteze

$$\mathcal{H}_0 : \beta = 0,$$

$$\mathcal{H}_1 : \beta \neq 0$$

Zaključivanje o koeficijentu smjera regresijskog pravca

- test se temelji na test-statistici čiju vrijednost $\hat{t}$ za eksperimentalne vrijednosti x_i i y_i računamo formulom

$$\hat{t} = \frac{s_x \cdot \widehat{\beta}}{s}\sqrt{n-1},$$

gdje je

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}, \quad s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}},$$

a $\widehat{\beta}$ procjena regresijskog koeficijenta β metodom najmanjih kvadrata
- ako je nul-hipoteza istinita, test-statistika ima Studentovu $t(n-2)$ distribuciju

Zaključivanje o koeficijentu smjera regresijskog pravca

- na temelju realizacije $\hat{t}$ test statistike računamo pripadnu p-vrijednost na sljedeći način: $p = P(|T| \geq |\hat{t}|)$ gdje je T slučajna varijabla koja ima Studentovu $t(n-2)$ distribuciju
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku kako slijedi:
  - ako je $p \leq \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. podaci potvrđuju da se promjene u vrijednosti nezavisne varijable odražavaju na promjene zavisne varijable na razini značajnosti α (model ima smisla)
  - ako je $p > \alpha$, nemamo dovoljno argumenata tvrditi da se promjene u vrijednosti nezavisne varijable odražavaju na promjene zavisne varijable na razini značajnosti α
- R - primjer 1

DIO VARIJABILNOSTI OBJAŠNJEN MODELOM

- koliki je dio promjena u eksperimentalnim vrijednostima zavisne varijable objašnjen dobivenim modelom?
- **koeficijent determinacije** – R^2 definiran je izrazom

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \quad R^2 \in [0, 1]$$

- koeficijent determinacije R^2 – u kolikoj mjeri je rasipanje eksperimentalnih vrijednosti zavisne varijable objašnjeno linearnom funkcijom $x \mapsto \alpha + \beta x$, a u kolikoj se mjeri radi o tzv. rezidualnom ili neobjašnjenom rasipanju, a tu informaciju očitavamo iz broja $(1 - R^2)$

Dio varijabilnosti objašnjen modelom

- velika vrijednost koeficijenta determinacije (R^2 blizu 1) ukazuje na to da linearan model objašnjava velik dio raspršenosti u eksperimentalnim vrijednostima zavisne varijable (samo mali dio je ostao neobjašnjen modelom i treba ga pripisati slučajnoj grešci)
- modeli kod kojih je R^2 mali nisu informativni za opis varijable Y korištenjem vrijednosti nezavisne varijable x jer opisuju samo mali dio varijablnosti u podacima iz Y, dok je veliki dio ostao neobjašnjen modelom
- R - primjer 1
- R - primjeri 2 i 3

MULTIVARIJATNI LINEARNI REGRESIJSKI MODEL

multivarijatni linearni regresijski model

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(m)} + \varepsilon_i, \quad i = 1, \ldots, n$$

- $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}$ - vrijednosti nezavisnih (prediktorskih) varijable $x^{(k)}$, $k = 1, \ldots, m$
- $Y_1, Y_2 \ldots, Y_n$ slučajne varijable (njihove izmjerene vrijednosti su $y_1, \ldots, y_n$)
- $\alpha, \beta_1, \ldots \beta_m$ - regresijski koeficijenti koje procjenjujemo metodom najmanjih kvadrata
- $e_i = y_i - \hat{\alpha} - \widehat{\beta}_1 x_i^{(1)} + \ldots + \widehat{\beta}_k x_i^{(m)}$ - reziduali (procjene grešaka), gdje su $\hat{\alpha}, \hat{\beta}_1, \ldots \hat{\beta}_m$ procjene regresijskih koeficijenata

Statističko zaključivanje - više varijabli

- dolazi li do promjene obilježja koje proučavamo zbog provođenja neke aktivnosti, u nekom drugom trenutku ili općenito u nekim drugim uvjetima?
- za dvije varijable (uzorka) možemo koristiti ranije obrađene metode
- za više od dvije varijable postoje analogne metode

Usporedba očekivanja

- promatrat ćemo jednu neprekidnu varijablu Y (zavisna varijabla)
- druga varijabla X (**faktor**) bit će diskretna s k različitih mogućih realizacija, $k \in \mathbb{N}$
- uzorak je veličine $n = n_1 + \ldots + n_k$

$$\begin{gathered} y_{11}, \ldots, y_{1n_1} \\ y_{21}, \ldots, y_{2n_2} \\ \vdots \\ y_{k1}, \ldots, y_{kn_k} \end{gathered}$$

gdje je n_i veličina uzorka karakterizirana kategorijom faktora i, $i \in \{1, \ldots, k\}$

Usporedba očekivanja

- pretpostavka: uzorci su nezavisni i iz normalnih distribucija s jednakim varijancama, odnosno

$$\begin{array}{ll} y_{11}, \ldots, y_{1n_1} & \text{je uzorak iz } \mathcal{N}(\mu_1, \sigma^2) \\ y_{21}, \ldots, y_{2n_2} & \text{je uzorak iz } \mathcal{N}(\mu_2, \sigma^2) \\ \vdots & \vdots \\ y_{k1}, \ldots, y_{kn_k} & \text{je uzorak iz } \mathcal{N}(\mu_k, \sigma^2), \end{array}$$

- želimo analizirati efekte faktora X na varijablu Y (slično kao kod jednostavne linearne regresije uz bitnu razliku što varijabla X predstavlja kategorije)

Usporedba očekivanja

- nul-hipoteza je

$$\mathcal{H}_0 : \ \mu_1 = \mu_2 = \ldots = \mu_k$$

(očekivanje obilježja Y se ne mijenja u ovisnosti o kategorijama obilježja X)
- alternativnu hipotezu formuliramo kao dvostranu hipotezu

$$\mathcal{H}_1 : \ \text{postoje} \ \ i, j \in \{1, \ldots, k\} \ \ \text{takvi da je} \ \ \mu_i \neq \mu_j$$

- statistički test za ovu svrhu temelji se na analizi varijance cijelog uzorka i uzoraka po kategorijama faktora, te na F distribuciji (F-test)
- poznat je kao (**jednofaktorska**) **analiza varijance** (*one-way ANOVA*)
- provoditi usporedbu očekivanja višestrukom primjenom t-testa po parovima varijabli nije dobro jer rezultira većom pogreškom prvog tipa

Usporedba očekivanja

- test je relativno robustan na odstupanja od normalnosti pogotovo ako se radi o velikim uzorcima
- u slučaju da je homogenost varijanci narušena može se koristiti Welchova varijanta F-testa
- R - primjeri 4, 5 i 6

LITERATURA

- Benšić, M. i Šuvak, N., *Primijenjena statistika*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2013.
- Benšić, M. i Šuvak, N., *Uvod u vjerojatnost i statistiku*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2014.