Linearna regresija

Zadatak A

U članku "Ethylene Synthesis in Lettuce Seeds: Its Physiological Significance" (Plant Physiology, 1972., str. 719-722) se proučava količina etilena (y, u nl/g) koju sadrži sjeme salate kao funkcija vremena izlaganja (x, u min) tvari koja apsorbira etilen. Podaci se nalaze u datoteci `zad51r.dat` (Devore, Jay L., Probability and Statistics for Engineering and the Sciences, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 472).

(a) Prikažite podatke (x, y) u Kartezijevom koordinatnom sustavu.

(b) Provedite prilagodbu kvadratičnog modela $y = \theta_0 + \theta_1 x + \theta_2 x^2$ podacima i dobivenu parabolu prikažite na istom grafu zajedno s empirijskim podacima. Izračunajte statistiku R^2 te testirajte hipotezu $\theta_2 = 0$, naspram dvostrane alternative.

(c) Nacrtajte graf reziduala, graf standardiziranih reziduala (za model iz (b)) te provjerite da li (standardizirani) reziduali dolaze iz jedinične normalne distribucije, i to upotrebom dva kriterija: grafičkog, koji se sastoji od grafa normalnih vjerojatnosti, te Kolmogorov-Smirnovljevog testa.

(d) Transformirajte podatke iz (a) tako da uzmete $y^0 = \ln(y)$. Prikažite točke (x, y^0) u Kartezijevom koordinatnom sustavu. Provedite prilagodbu linearnog modela $y^0 = \theta_0 + \theta_1 x$ transformiranim podacima. Također, provedite analizu reziduala (tj. ponovite (c) dio zadatka) za ovaj model.

(e) Uz pretpostavku da je linearan model dobar za transformirane podatke, napišite kako glasi model za originalne podatke (iz (a)). Nacrtajte pripadnu regresijsku funkciju zajedno sa originalnim podacima. Takoder, prikažite točke $(y, \hat{y})$ zajedno s pravcem $y = x$ u Kartezijevom koordinatnom sustavu ($\hat{y}$ je procjena od y na osnovu modela za originalne podatke).

(f) Nađite gornje i donje krivulje koje definiraju 95% pouzdane intervale, prvo za srednju vrijednost od Y (uz dano x), a zatim i za Y (uz dano x) te ih prikažite zajedno s originalnim podacima (x, y) i regresijskom funkcijom u Kartezijevom koordinatnom sustavu (to napravite za oba modela za originalne podatke - onaj iz (b) i iz (e)). Koji je model bolji za originalne podatke?

Zadatak B

Članak "Determination of Biological Maturity and Effect of Harvesting and Drying Conditions on Milling Quality of Paddy" (J. Agricultural Eng. Research, 1975., str. 353-361) obrađuje podatke o žetvi paddyja, vrste žita u Indiji. Varijabla x predstavlja datum žetve (tj. to je broj dana proteklih od sjetve žita), a y predstavlja urod (u kg/ha). Podaci se nalaze u datoteci `zad55r.dat` (Devore, Jay L., Probability and Statistics for Engineering and the Sciences, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 478).

(a) Prikažite podatke (x, y) u Kartezijevom koordinatnom sustavu.

(b) Prilagodite kvadratični model $y = \theta_0 + \theta_1 x + \theta_2 x^2$ podacima iz (a). Dobivenu krivulju grafički prikažite na grafu iz (a) zajedno s empirijskim podacima. Izračunajte statistiku R^2.

(c) Nacrtajte graf reziduala, te graf standardiziranih reziduala. Provjerite da li (standardizirani) reziduali dolaze iz jedinične normalne distribucije i to upotrebom dva kriterija: grafičkog, koji se sastoji od grafa normalnih vjerojatnosti, te Kolmogorov-Smirnovljevog testa.

(d) Sprovedite test osnovne hipoteze $H_0 : \theta_2 = 0$ u odnosu na alternativu $H_a : \theta_2 \neq 0$. Može li se kvadratični član predloženog modela zanemariti? Ako da, sprovedite prilagodbu linearnog modela podacima i izvršite analizu reziduala kao u (c).

(e) Procijenite 95% pouzdane intervale za parametre prihvaćenog modela. Prikažite (usporedite) ih grafički na istom grafu. Procijenite 95% pouzdano područje za parametre modela. Prikažite grafički dobiveno pouzdano područje.

(f) Odredite gornje i donje krivulje koje definiraju 95% pouzdane intervale, prvo za srednju vrijednost od Y (uz dano x), te zatim i za Y (uz dano x) te ih prikažite zajedno s originalnim podacima (x, y) i regresijskom funkcijom u Kartezijevom koordinatnom sustavu.

Zadatak C

U radu "An Ultracentrifuge Flour Absorption Method" (Cereal Chemistry, 1978., str. 96-101) autori su proučavali odnos između apsorpcije vode pšeničnog brašna i raznih karakteristika tog brašna. Konkretno, promatrali su odnos između apsorpcije z (u %) te proteina brašna x (u %) i gubitka škroba y (u Farrandovim jedinicama). Podaci dobiveni pokusom nalaze se u datoteci `zad57r.dat` (Devore, Jay L., Probability and Statistics for Engineering and the Sciences, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 490).

(a) Prikažite podatke (x, z), (y, z), (x, y) i (x, y, z) u Kartezijevim koordinatnim sustavima.

(b) Izračunajte Pearsonov koeficijent korelacije za podatke (y, z) te provedite pripadni test koreliranosti. Nadalje, izračunajte Spearmanov koeficijent korelacije za podatke (x, y) te također provedite pripadni test koreliranosti.

(c) Sprovedite prilagodbu linearnih modela $z = \alpha_0 + \alpha_1 x$, te $z = \beta_0 + \beta_1 y$ podacima iz (a). Dobivene pravce grafički prikažite na grafovima iz (a) zajedno s empirijskim podacima i izračunajte pripadne R^2 statistike.

(d) Sprovedite prilagodbu linearnog modela $z = \theta_0 + \theta_1 x + \theta_2 y$ podacima iz (a). Sprovedite test o značajnosti ovog modela. Nadalje, usporedite taj prošireni model s reduciranim modelima iz (b) i za svaki od njih sprovedite test osnovne hipoteze H_0 : podaci podržavaju reducirani model uz alternativu H_a : podaci podržavaju prošireni model (testirajte da li su odgovarajući koeficijenti jednaki nula). Izračunajte statistiku R^2.

(e) Nacrtajte graf reziduala te graf standardiziranih reziduala za model iz (d). Provjerite dolaze li (standardizirani) reziduali iz jedinične normalne distribucije i to upotrebom dva kriterija: grafičkog, koji se sastoji od grafa normalnih vjerojatnosti, te Kolmogorov-Smirnovljevog testa.

(f) Odredite gornje i donje plohe koje definiraju 95% pouzdane intervale, prvo za srednju vrijednost od Z (uz dano (x, y)), a zatim i za Z (uz dano (x, y)) te ih prikažite zajedno s originalnim podacima (x, y, z) u trodimenzionalnom Kartezijevom koordinatnom sustavu.

Napomena: Ukoliko imate dodatnih pitanja vezanih uz ovaj konkretan projekt, javite se asistentu Stjepanu Šebeku na stjepan.sebek@fer.hr.