

Datový soubor cigarety.sta

Datový soubor obsahuje údaje o 25 značkách cigaret. Jsou sledovány tyto proměnné:

X_1 ... obsah dehtu (v mg)

X_2 ... obsah nikotinu (v mg)

X_3 ... hmotnost cigarety (v g)

Y ... obsah oxidu uhelnatého v uvolněném kouři (v mg)

Úkoly

1. Popište datový soubor pomocí číselných charakteristik a vhodných grafů.
2. Normalitu proměnných Y , X_1 , X_2 , X_3 posuďte pomocí N-P plotu a S-W testu s hladinou významnosti 0,05.
3. Závislost mezi dvojicemi proměnných (Y, X_1) , (Y, X_2) , (Y, X_3) znázorněte dvourozměrnými tečkovými diagramy.
4. Vypočtete výběrovou korelační matici všech čtyř proměnných a pro $\alpha = 0,05$ otestujte významnost jednotlivých korelačních koeficientů.
5. Vypočtete výběrové parciální korelační koeficienty $r_{Y, X_1.(X_2, X_3)}$, $r_{Y, X_2.(X_1, X_3)}$, $r_{Y, X_3.(X_1, X_2)}$ a porovnejte je s výběrovými párovými korelačními koeficienty r_{YX_1} , r_{YX_2} , r_{YX_3} . Na hladině významnosti $\alpha = 0,05$ testujte hypotézy o nevýznamnosti parciálních korelačních koeficientů $\rho_{Y, X_1.(X_2, X_3)}$, $\rho_{Y, X_2.(X_1, X_3)}$, $\rho_{Y, X_3.(X_1, X_2)}$.
6. V první fázi zpracování předpokládejte, že je vhodný regresní model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.
Vypočtete index determinace a interpretujte ho. Proved'te celkový F-test. Odhadněte parametry regresního modelu. Proved'te dílčí t-testy pro regresní koeficienty. Zjistěte odhad rozptylu. (Hladinu významnosti volte $\alpha = 0,05$.)
Posuďte pomocí beta koeficientů vliv jednotlivých nezávisle proměnných veličin na regresní model.
Posuďte pomocí VIF koeficientů, zda proměnné X_1 , X_2 , X_3 mohou způsobit multikolinearitu v uvedeném modelu.
7. Z regresního modelu odstraňte ty proměnné, jejichž regresní koeficienty se neprokázaly významné pro $\alpha = 0,05$. Sestavte nový regresní model a proved'te v něm tytéž úkoly jako v bodě 5.
Normalitu reziduí v tomto novém regresním modelu posuďte S-W testem na hladině významnosti $\alpha = 0,05$.
V novém regresním modelu najděte 95% interval spolehlivosti pro teoretickou regresní funkci a 95% predikční interval.