

Poznámky

- Výpočty môžete v prípade potreby zaokrúhľovať na 3 desatinné miesta.
- Môžete používať výsledky, ktoré boli odvodené na prednáškach alebo cvičeniach.

Úloha 1 (5 bodov). Pre istú premennú y sme vykonali 20 meraní a uvažujeme pre ňu dva modely s rôznymi vysvetľujúcimi premennými:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_4 x_{i,4} + \varepsilon_i, \quad (\text{M1})$$

$$y_i = \beta'_0 + \beta'_1 z_{i,1} + \dots + \beta'_{11} z_{i,11} + \varepsilon'_i. \quad (\text{M2})$$

Tieto modely sme chceli porovnať pomocou Akaikeho informačného kritéria, ale omylom sme ich porovnali pomocou koeficientu determinácie. Vypočítali sme teda $R_1^2 = 0,8$, $R_2^2 = 0,9$ (index 1 zodpovedá (M1), index 2 zodpovedá (M2)), a okrem toho ešte vieme, že $\sum_i y_i^2 = 205$ a $\sum_i y_i = 10$. Vypočítajte AIC_1 a AIC_2 a určte, ktorý model je lepší z pohľadu AIC.

Úloha 2 (8 bodov). Máme model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, kde $\beta = (\beta_0, \beta_1)^T$, $\mathbf{Y} \in \mathbb{R}^n$, $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, a vieme, že

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 9 & 0 \\ 0 & 14 \end{pmatrix}.$$

Chceme testovať hypotézu $H_0 : \beta_0 = \beta_1 + 2$.

- Nájdite \mathbf{R} a r , pomocou ktorých H_0 vieme zapísať ako $H_0 : \mathbf{R}\beta = r$. V ďalších častiach pracujeme s \mathbf{R} a r nájdenými v tejto časti.
- Vypočítajte strednú hodnotu a disperziu odhadu $\mathbf{R}\hat{\beta} - r$.
- Okrem odhadu z časti (b) sme zostrojili konkurenčný odhad hodnoty $\mathbf{R}\beta - r$:

$$V = \mathbf{R} \begin{pmatrix} 1/10 & 0 \\ 0 & 1/15 \end{pmatrix} \mathbf{X}^T \mathbf{Y} - r.$$

Určte, či je V nevychýleným odhadom kvantity $\mathbf{R}\beta - r$.

- Vypočítajte aj disperziu odhadu V a následne určte, ktorý z odhadov z častí (b) a (c) má menšiu disperziu.

Poznámka: Disperzie v častiach (b) a (c) vyjadrite v tvare “číslo (alebo zlomok) krát σ^2 ”, napr. $1543\sigma^2$ a $21,1\sigma^2$.

Bonus (1 bod): Zdôvodnite, prečo/ako sú výsledky tejto úlohy v súlade s teóriou z prednášky, ktorá nám hovorí o “najkvalitnejších” odhadoch v lineárnej regresii.

Úloha 3 (7 bodov). Veľkonočný zajačik pestuje mrkvu a pozoroval, že množstvo vypestovanej mrkvy (*mrkva*) závisí od množstva aplikovaného hnojiva (*hnojivo*) a dodatočného množstva vody zo zavlažovacieho systému (*voda*). Sformuloval model nasledovne:

$$mrkva_i = \beta_0 + \beta_1 \cdot hnojivo_i + \beta_2 \cdot voda_i + \varepsilon_i, \quad (\text{Z1})$$

kde $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, pričom σ^2 je neznáma. Zajačik rozdelil svoju záhradku na štyri časti a množstvá hnojiva, vody a vypestovanej mrkvy na jednotlivých častiach zaznačil do nasledujúcej tabuľky:

mrkva	10	12	9	16
hnojivo	2	7	3	8
voda	3	4	5	6

- Zostrojte maticu plánu \mathbf{X} pre model (Z1).
- Veľkonočný zajačik chcel otestovať hypotézu, že vplyv hnojiva je dvojnásobný voči vplyvu vody. Sformulujte túto hypotézu.
- Sformulujte submodel zodpovedajúci hypotéze z predošlej časti úlohy.
- Zajačik vypočítal $S^2 = 6,25$ v pôvodnom modeli a $RSS_{sub} = 6,266$. Vypočítajte testovú štatistiku, pomocou ktorej vieme otestovať hypotézu z časti (b).
- Zajačik si nepamätal, aké rozdelenie má testová štatistika vypočítaná v predošlej časti za platnosti H_0 , preto z tabuliek vyhľadal tieto hodnoty: $\mathcal{F}_{1,4}(90\%) = 4,54$; $\mathcal{F}_{2,4}(95\%) = 6,94$; $\mathcal{F}_{1,1}(90\%) = 39,86$, $\mathcal{F}_{1,1}(95\%) = 161,44$; $t_1(90\%) = 3,08$; $t_1(95\%) = 6,31$, $t_2(95\%) = 2,92$. $t_4(90\%) = 1,53$. Pomôžte mu, s ktorou hodnotou má porovnať testovú štatistiku, ak chce testovať na hladine významnosti 10% a rozhodnite, či H_0 zamietame alebo nie.

Poznámky

- Výpočty môžete v prípade potreby zaokrúhľovať na 3 desatinné miesta.
- Môžete používať výsledky, ktoré boli odvodené na prednáškach alebo cvičeniach.

Úloha 1 (7 bodov). Veľkonočný zajačik pestuje mrkvu a pozoroval, že množstvo vypestovanej mrkvy (*mrkva*) závisí od množstva aplikovaného hnojiva (*hnojivo*) a dodatočného množstva vody zo zavlažovacieho systému (*voda*). Sformuloval model nasledovne:

$$mrkva_i = \beta_0 + \beta_1 \cdot hnojivo_i + \beta_2 \cdot voda_i + \varepsilon_i, \quad (Z1)$$

kde $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, pričom σ^2 je neznáma. Zajačik rozdelil svoju záhradku na štyri časti a množstvá hnojiva, vody a vypestovanej mrkvy na jednotlivých častiach zaznačil do nasledujúcej tabuľky:

mrkva	10	15	13	18
hnojivo	4	6	5	9
voda	5	6	7	8

- Zostrojte maticu plánu \mathbf{X} pre model (Z1).
- Veľkonočný zajačik chcel otestovať hypotézu, že vplyv hnojiva je trojnásobný voči vplyvu vody. Sformulujte túto hypotézu.
- Sformulujte submodel zodpovedajúci hypotéze z predošlej časti úlohy.
- Zajačik vypočítal $S^2 = 2,333$ v pôvodnom modeli a $RSS_{sub} = 2,347$. Vypočítajte testovú štatistiku, pomocou ktorej vieme otestovať hypotézu z časti (b).
- Zajačik si nepamätal, aké rozdelenie má testová štatistika vypočítaná v predošlej časti za platnosti H_0 , preto z tabuliek vyhľadal tieto hodnoty: $\mathcal{F}_{1,4}(95\%) = 7,71$; $\mathcal{F}_{2,4}(97,5\%) = 10,65$; $\mathcal{F}_{1,1}(95\%) = 161,45$, $\mathcal{F}_{1,1}(97,5\%) = 647,79$; $t_1(95\%) = 6,31$; $t_1(97,5\%) = 12,71$, $t_2(95\%) = 2,92$. $t_4(95\%) = 2,13$. Pomôžte mu, s ktorou hodnotou má porovnať testovú štatistiku, ak chce testovať na hladine významnosti 5% a rozhodnite, či H_0 zamietame alebo nie.

Úloha 2 (5 bodov). Pre istú premennú y sme vykonali 10 meraní a uvažujeme pre ňu dva modely s rôznymi vysvetľujúcimi premennými:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_7 x_{i,7} + \varepsilon_i, \quad (M1)$$

$$y_i = \beta'_0 + \beta'_1 z_{i,1} + \dots + \beta'_5 z_{i,5} + \varepsilon'_i. \quad (M2)$$

Tieto modely sme chceli porovnať pomocou Akaikeho informačného kritéria, ale omylom sme ich porovnali pomocou koeficientu determinácie. Vypočítali sme teda $R_1^2 = 0,7$, $R_2^2 = 0,6$ (index 1 zodpovedá (M1), index 2 zodpovedá (M2)), a okrem toho ešte vieme, že $\sum_i y_i^2 = 210$ a $\sum_i y_i = 30$. Vypočítajte AIC_1 a AIC_2 a určte, ktorý model je lepší z pohľadu AIC.

Úloha 3 (8 bodov). Máme model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, kde $\beta = (\beta_0, \beta_1)^T$, $\mathbf{Y} \in \mathbb{R}^n$, $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, a vieme, že

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 8 & 0 \\ 0 & 9 \end{pmatrix}.$$

Chceme testovať hypotézu $H_0 : \beta_1 = \beta_0 + 3$.

- a) Nájdite \mathbf{R} a r , pomocou ktorých H_0 vieme zapísať ako $H_0 : \mathbf{R}\beta = r$. V ďalších častiach pracujeme s \mathbf{R} a r nájdenými v tejto časti.
- b) Vypočítajte strednú hodnotu a disperziu odhadu $\mathbf{R}\hat{\beta} - r$.
- c) Okrem odhadu z časti (b) sme zostrojili konkurenčný odhad hodnoty $\mathbf{R}\beta - r$:

$$V = \mathbf{R} \begin{pmatrix} 1/9 & 0 \\ 0 & 1/10 \end{pmatrix} \mathbf{X}^T \mathbf{Y} - r.$$

Určte, či je V nevychýleným odhadom kvantity $\mathbf{R}\beta - r$.

- d) Vypočítajte aj disperziu odhadu V a následne určte, ktorý z odhadov z častí (b) a (c) má menšiu disperziu.

Poznámka: Disperzie v častiach (b) a (c) vyjadrite v tvare “číslo (alebo zlomok) krát σ^2 ”, napr. $1543\sigma^2$ a $21,1\sigma^2$.

Bonus (1 bod): Zdôvodnite, prečo/ako sú výsledky tejto úlohy v súlade s teóriou z prednášky, ktorá nám hovorí o “najkvalitnejších” odhadoch v lineárnej regresii.