Opakovanie

1. Zopakujte si otázky a príklady z predchádzajúcej série príkladov na precvičenie a príklady z prvej písomky. ;-)

2. Uvažujme dva regresné modely pre tie isté dáta $y_1, \ldots, y_n$. Vieme, že (pri splnení potrebných predpokladov) odhady variancie náhodných chýb ε_{1i} (v prvom modeli) a ε_{2i} (v druhom modeli) sú
$$\hat{\sigma}_1^2 = \frac{RSS_1}{n-k_1}, \ \ \hat{\sigma}_2^2 = \frac{RSS_2}{n-k_2}.$$
Ukážte, že ak $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$, tak prvý model má väčší upravený koeficient determinácie.

3. Dokážte, že po vynechaní premennej z modelu sa upravený koeficient determinácie zväčší práve vtedy, keď je absolútna hodnota t-štatistiky z testovania signifikancie príslušného parametra menšia ako 1.

4. Uvažujme model, v ktorom sa snažíme vysvetliť detskú úmrtnosť v štátoch sveta y pomocou premenných $x_1, \ldots, x_k$. Okrem toho do modelu zaradíme premenné, ktoré popisujú polohu štátu: Premenná d_1 sa rovná 1, ak je daný štát v Afrike, inak sa rovná 0. Premenná d_2 sa rovná 1, ak daný štát nie v Afrike, inak sa rovná 0. Prečo sa model
$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} d_1 + \beta_{k+2} d_2 + \varepsilon$$
nedá odhadnúť vzorcom $(X^T X)^{-1} X^T Y$?

5. Uvažujme dva lineárne regresné modely pre dáta Y:
   - V prvom z nich je matica vysvetľujúcich premenných X.
   - V druhom sú premenné transformované, matica vysvetľujúcich premenných je XA, kde A je regulárna matica.

   Označme $\hat{\beta}_X$ odhad parametrov prvého modelu a $\hat{\beta}_{XA}$ odhad parametrov druhého modelu. Dokážte, že

   (a) $\hat{\beta}_{XA} = A^{-1}\hat{\beta}_X$
   (b) Rezíduá z obidvoch modelov sú rovnaké.
   (c) $\widehat{Var}[\hat{\beta}_{XA}] = A^{-1}\widehat{Var}[\hat{\beta}_X](A^T)^{-1}$

Multikolinearita

1. Uvažujme model, v ktorom sa snažíme vysvetliť spotrebu elektrickej energie domácnosti pomocou počtu členov domácnosti, ich celkového príjmu a rozlohy bytu. Prečo sa dá v tomto modeli čakať problém s multikolinearitou?

2. Aké dôsledky má prítomnosť mulikolinearity v modeli? Aké kritériá sa dajú použiť na jej zistenie?

3. V prípade multikolinearity v modeli sa niekedy odporúča použiť tvz. hrebeňovú regresiu (ridge regression): vektor parametrov β odhadneme ako $\hat{\beta} = (X^T X + kI)^{-1} X^T Y$, kde $k > 0$ je parameter, ktorý treba zvoliť.[1] Takýto odhad je vychýlený, ale pre malé hodnoty k má menšiu varianciu ako MNŠ odhad. Prečo nie je toto tvrdenie o variancii v spore s Gauss-Markovovou vetou?

[1] V numerike je tento postup známy ako Tichonovova regularizácia.

Heteroskedasticita a zovšeobecnená metóda najmenších štvorcov

1. Vysvetlite Whitov test heteroskedasticity.
   (a) Na čo sa používa tento test? Akú hypotézu ním testujeme?
   (b) Akú pomocnú regresiu spravíme? Ako z jej výsledkov vypočítame hodnotu testovacej štatistiky? (obidve verzie testu - "cross terms" aj "no cross terms")
   (c) Aké je asymptotické pravdepodobnostné rozdelenie štatistiky, ak platí nulová hypotéza?
   (d) Aká je kritická oblasť testu, t.j. kedy nulovú hypotézu zamietame?
   (e) Kedy je v modeli heteroskedasticita - ak nulovú hypotézu zamietame alebo ak ju nezamietame?
2. Goldfeld-Quandtov test
   (a) Odhadujeme model $Y = X\beta + \varepsilon$, pričom ε_i majú normálne rozdelenie a nulovú strednú hodnotu. Máme však dve skupiny dát, pričom v prvej sa variancia ε_i rovná σ_1^2 a v druhej σ_2^2. Ako otestujeme Goldfeld-Quandtovym testom hypotézu $\sigma_1^2 = \sigma_2^2$? Napíšte, ako vypočítame testovaciu štatistiku, aké má rozdelenie za platnosti nulovej hypotézy a kedy nulovú hypotézu zamietame.
   (b) Odhadujeme regresný model, pričom máme podozrenie, že jedna z premenných spôsobuje heteroskedasticitu. Domnievame sa, že so zväčšujúcimi sa hodnotami tejto premennej sa zväčšuje variancia náhodnej zložky ε. Ako sa o tom môžeme presvedčiť pomocou Goldfeld-Quandtovho testu? Napíšte postup a povedzte, aký výsledok testu potvrdí naše podozrenie o uvedenom tvare heteroskedasticity?
3. Breusch-Paganov test
   (a) Kedy použijeme na testovanie heteroskedasticity Breusch-Paganov test?
   (b) Akú pomocnú regresiu spravíme? Ako z jej výsledkov vypočítame hodnotu testovacej štatistiky?
   (c) Aké je asymptotické pravdepodobnostné rozdelenie štatistiky, ak platí nulová hypotéza?
   (d) Aká je kritická oblasť testu, t.j. kedy nulovú hypotézu zamietame?
   (e) Kedy je v modeli heteroskedasticita - ak nulovú hypotézu zamietame alebo ak ju nezamietame?
4. Čo je Whitov odhad kovariančnej matice $\hat{\beta}$ - ako sa vypočíta a na čo ho potrebujeme?
5. Waldov test: Uvažujme model $Y = X\beta + \varepsilon$. Testujeme hypotézu $R\beta = r$.
   (a) Kedy môžeme použiť na testovanie tejto hypotézy F-štatistiku?
   (b) Napíšte Waldovu štatistiku na testovanie tejto hypotézy. Aké má asymptotické rozdelenie a kedy hypotézu zamietame?
   (c) Odvoďte Waldov test na testovanie signifikancie parametra. Kedy použijeme tento test namiesto t-štatistiky?
6. Predpokladajme, že v modeli $Y = X\beta + \varepsilon$ platí $E[\varepsilon] = 0$, $Var[\varepsilon] = \sigma^2\Omega$, kde Ω je známa symetrická kladne definitná matica. Ktoré z nasledujúcich vlastností MNŠ odhadu $\hat{\beta} = (X^T X)^{-1} X^T Y$ zostávajú v platosti?
   (a) $\hat{\beta}$ je nevychýlnený odhad parametra β
   (b) Kovariančná matica odhadu je $\sigma^2(X^T X)^{-1}$
   (c) $\hat{\beta}$ má spomedzi lineárnych nevychýlených odhadov parametra β najmenšiu varianciu.

7. Uvažujme model $Y = X\beta + \varepsilon$ platí $E[\varepsilon] = 0$, $Var[\varepsilon] = \sigma^2\Omega$, kde Ω je známa symetrická kladne definitná matica. Pomocou trensformácie na homoskedastický model $\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon}$ odvoďte
   (a) odhad parametra β (robili sme na cvičení) a nájdite jeho kovariančnú maticu
   (b) odhad parametra σ^2.

Metóda maximálnej vierohodnosti v regresnom modeli

1. Uvažujme lineárny regresný model $Y = X\beta + \varepsilon$. Nájdite odhady parametrov β a σ^2 metódou maximálnej vierohodnosti, ak
   (a) vektor ε má normálne rozdelenie $N(0, \sigma^2 I)$
   (b) vektor ε má normálne rozdelenie $N(0, \sigma^2\Omega)$, kde Ω je známa symetrická kladne definitná matica.
2. Uvažujme model
$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + \varepsilon_i,$$
kde ε_i sú nezávislé náhodné premenné so Studentovym rozdelením s r stupňami voľnosti.[2] Model odhadujeme metódou maximálnej vierohodnosti.Dokážte, že odhady parametrov β_i $(i = 0, \ldots, k)$ sa zhodujú s odhadmi získanými metódou najmenších štvorcov.
3. Uvažujme model
$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + \varepsilon_i,$$
kde ε_i sú nezávislé náhodné premenné s rovnomerným rozdelnením na intervale $[-a, a]$. Model odhadujeme metódou maximálnej vierohodnosti.
   (a) Dokážte, že odhady parametrov β_i $(i = 0, \ldots k)$ minimalizujú najväčšiu absolútnu hodnotu rezíduí, t.j. sú riešením úlohy
$$\max_{i=1,\ldots,n} |y_i - (\beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki})| \to \min .$$
   (b) Odhadnite parameter a.

Testovanie zhodnosti parametrov v dvoch skupinách dát

1. Uvažujme produkčnú funkciu Cobb-Douglasovho typu
$$\log Y = \alpha + \beta \log L + \gamma \log K + \varepsilon$$
a dáta z dvoch priemyselných odvetví. Navrhnite postup, ako testovať hypotézu, že obidve odvetvia majú rovnakú produkčnú funkciu.
2. Opäť uvažujme produkčnú funkciu Cobb-Douglasovho typu
$$\log Y = \alpha + \beta \log L + \gamma \log K + \varepsilon$$
a dáta z dvoch priemyselných odvetví. Teraz však predpokladáme, že výstup pri jednotkovom množstve práce a kapitálu je rovnaký v oboch odvetviach (α), ale elasticity práce a kapitálu (β_1, β_2, γ_1, γ_2) môžu byť rôzne.
   (a) Navrhnite regresiu, ktorou odhadneme parametre α, β_1, β_2, γ_1, γ_2.
   (b) Navrhnite postup, ako testovať hypotézu, že obidve odvetvia majú rovnakú produkčnú funkciu.

---

[2] Hustota náhodnej premennej so Studentovym rozdelením s r stupňami voľnosti je
$$f(x) = \frac{1}{\sqrt{r\pi}} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{\frac{r+1}{2}} .$$