Egzamin z ekonometrii 30.01.2009
1 semestr

Pytania teoretyczne

1. Pokazać, że w modelu ze stałą suma reszt jest równa zeru.
2. Wyprowadzić postać macierzy wariancji kowariancji $\boldsymbol{b}$ i podać interpretację jej elementów.
3. Jaką postać ma statystyka służąca do testowania hipotezy o tym, że $\beta_k = \beta_k^*$?
4. W jakim szczególnym przypadku można uzyskać prawidłowe oszacowania parametrów mimo, że w modelu pominięto istotne zmienne?

ZADANIE 1 Przyjmijmy, że y to logarytm dochodu respondenta, a x to zmienna opisująca płeć respondenta, przyjmująca wartość 2 dla mężczyzny oraz 1 dla kobiety. Czyli badamy wpływ płci na dochód, ale badacz zapomniał rozkodować zmienną dyskretną na zmienne zerojedynkowe. Badacz szacuje model:

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \tag{*}$$

podczas gdy chce oszacować model:

$$y_i = \beta_0 + \beta_1 plec\_m + \varepsilon_i \tag{**}$$

gdzie *plec_m* przyjmuje wartość 1 dla mężczyzny a 0 dla kobiety.

1. Zakładając, że są spełnione założenia KMRL, pokaż, że szacując MNK model (*) uzyskamy nieobciążony estymator $\widehat{\alpha}_1$ dla parametru β_1, ale obciążony estymator $\widehat{\alpha}_0$ parametru dla stałej β_0.
2. Porównaj *RSS* i R^2 w regresjach (*) i (**) – odpowiedź dokładnie uzasadnij.

Rozwiązanie:

1. Zauważmy, że $x_i = plec\_m_i + 1$. Model (**) można przekształcić w sposób następujący:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 plec\_m_i + \varepsilon_i \\ &= (\beta_0 - \beta_1) + \beta_1 (plec\_m_i + 1) + \varepsilon_i \\ &= (\beta_0 - \beta_1) + \beta_1 x_i + \varepsilon_i \\ &= \alpha_0 + \alpha_1 x_i + \varepsilon_i \end{aligned}$$

   gdzie $\alpha_0 = \beta_0 - \beta_1$, $\beta_1 = \alpha_1$. Przy spełnionych założeniach KMRL estymator $\widehat{\alpha}_1$ jest niobciążonym estymatorem parametru β_1 a estymator $\widehat{\alpha}_0$ nieobciążonym estymatorem parametru $\alpha_0 = \beta_0 - \beta_1$ ale obciążonym estymatorem parametru β_0. Obciążenie estymatora $\widehat{\alpha}_0$ jako estymatora parametru β_0 wynosi $\beta_0 - \mathrm{E}(\widehat{\alpha}_0) = \beta_1$.

2. TSS w modelu (*) i (**) będzie takie samo, bo zmienna zależna w obu modelach jest identyczna. Zauważmy, z zależności $plec\_m_i = x_i - 1$ wynika, że $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{A}$, gdzie $\boldsymbol{X}$ zmienne w modelu (*) a $\boldsymbol{X}^*$ zmienne w modelu (**). a $\boldsymbol{A}$ ma postać $\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$ i jest nieosobliwa (przyjmujemy, że pierwszą kolumną macierzy $\boldsymbol{X}$ jest stała). RSS w modelu (*) jest równa

$$RSS = \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\boldsymbol{M}_X\boldsymbol{y},$$

   a w modelu (**)

$$\begin{aligned} RSS^* = \boldsymbol{e}^{*\prime}\boldsymbol{e}^* = \boldsymbol{y}\boldsymbol{M}_{X^*}\boldsymbol{y} &= \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{X}^*\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\right)\boldsymbol{y} \\ &= \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{X}^*\boldsymbol{A}\left(\boldsymbol{A}'\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\boldsymbol{X}^{*\prime}\right)\boldsymbol{y} \\ &= \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{X}^*\boldsymbol{A}\boldsymbol{A}^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{A}'^{-1}\boldsymbol{A}'\boldsymbol{X}^{*\prime}\right)\boldsymbol{y} \\ &= \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{X}^*\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\right)\boldsymbol{y} \\ &= RSS \end{aligned}$$

   tak więc sumy kwadratów reszt w obu modelach są równe, ponieważ równe są też TSS więc takie same będą również $R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS}$

ZADANIE 2 Zakładamy, że prawdziwy jest następujący model bez stałej i z dwoma zmiennymi objaśniającymi:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

gdzie $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$.
Jako oszacowanie nieznanego parametru β_1 zaproponowano estymator postaci $b_1 = \frac{\sum_{i=1}^n y_i}{x_{n:n}}$, gdzie $x_{n:n} = \max(x_{11}, ..., x_{1n})$. Wyznaczyć obciążenie dla podanego estymatora przy założeniu, że zmienne objaśniające są nielosowe.
Rozwiązanie:

$$\mathrm{E}(y_i) = \mathrm{E}(\beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i) = \beta_1 x_{1i} + \beta_2 x_{2i}$$

$$\mathrm{E}(b_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n y_i}{x_{n:n}}\right) = \frac{1}{x_{n:n}} \mathrm{E}\left(\sum_{i=1}^n y_i\right) = \frac{1}{x_{n:n}} \sum_{i=1}^n \mathrm{E}(y_i) = \frac{1}{x_{n:n}} \sum_{i=1}^n (\beta_1 x_{1i} + \beta_2 x_{2i})$$
$$= \frac{1}{x_{n:n}} \left(\beta_1 \sum_{i=1}^n x_{1i} + \beta_2 \sum_{i=1}^n x_{2i}\right)$$

Zatem obciążenie estymatora wynosi:

$$\beta_1 - \mathrm{E}(b_1) = \beta_1 - \frac{1}{x_{n:n}} \sum_{i=1}^n (\beta_1 x_{1i} + \beta_2 x_{2i})$$

ZADANIE 3 Październikowe Badania Wynagrodzeń jest przekrojowym badaniem płace w przedsiębiorstwach zatrudniających powyżej 10 pracowników. Na podstawie uzyskanych informacji oszacowano parametry równania płacy w sektorze przedsiębiorstw dla Polski, uzyskując następujące wyniki: Zmienna $lzarobki$ oznacza logarytm płacy, $wiek$ jest wiekiem pracownika w latach, $wiek2$ to kwadrat zmiennej wiek, $plec$ przyjmuję wartość 2 dla kobiety, $dosw$ to doświadczenie zawodowe pracownika w latach, $dosw2$ to kwadrat doświadczenia, $prywatna$ oznacza, że firma jest własnością prywatną, $wielkosc$ 1 oznacza firmę do 20 pracowników, $wielkosc$ 2 oznacza firmę zatrudniającą od 20 do 100 pracowników, a $wielkosc$ 3 firmę zatrudniającą powyżej 100 pracowników.

```
      Source |       SS       df       MS              Number of obs =   48692
-------------+------------------------------           F(  8, 48683) =  788.67
       Model |  1544.52571     8  193.065714           Prob > F      =  0.0000
    Residual |  11917.5166 48683  .244798319           R-squared     =  0.1147
-------------+------------------------------           Adj R-squared =  0.1146
       Total |  13462.0423 48691  .276479068           Root MSE      =  .49477

------------------------------------------------------------------------------
    lzarobki |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .0322697   .0026325    12.26   0.000     .0271099    .0374295
       wiek2 |  -.0003021   .0000315    -9.60   0.000    -.0003638   -.0002404
     _Iplec_2 |  -.1579855   .0046649   -33.87   0.000    -.1671287   -.1488423
        dosw |   -.003419   .0012305    -2.78   0.005    -.0058308   -.0010073
       dosw2 |   .0001455   .0000294     4.94   0.000     .0000878    .0002032
    prywatna |   .1710564   .0047111    36.31   0.000     .1618226    .1802901
_Iwielkosc_2 |   .1138605   .0089159    12.77   0.000     .0963853    .1313358
_Iwielkosc_3 |   .2243529    .008288    27.07   0.000     .2081083    .2405975
       _cons |   6.645797   .0461863   143.89   0.000     6.555271    6.736322
------------------------------------------------------------------------------
Breusch-Pagan     chi2(1)            =       2.90    Prob > chi2  = 0.0886
Ramsey RESET test F(9, 48675       ) =       2.14    Prob > F =     0.0534
```

Przyjmując poziom istotności $\alpha = 0{,}05$ odpowiedz na poniższe pytania:

1. Oceń dopasowanie modelu do danych empirycznych.

2. Określ które zmienne można uznać za statystycznie istotne?

3. Dokonaj interpretacji wartości parametrów modelu.

4. Czy forma funkcyjna oszacowanego powyżej modelu jest liniowa?

5. Czy składnik losowy oszacowanego modelu jest homoscedastyczny

6. Zaproponuj sposób weryfikacji hipotezy o braku wpływu doświadczenia zawodowego na wysokość uzyskiwanych zarobków.

7. Z jakiego powodu nie przeprowadzono testu autokorelację składnika losowego?

8. Jeżeli model nie spełnia założeń KMRL opisz jakie to ma konsekwencje dla poprawności wnioskowania statystycznego, oraz jakie są metody radzenia sobie z tym problemem.

   Każdą odpowiedź uzasadnij wynikiem odpowiednich testów diagnostycznych zapisując po wartość statystyki testowej lub jej *p-value,* oraz interpretację wyniku.

Rozwiązanie:

1. Zmienne objaśniające wyjaśniają wariancję zarobków w 11 %, o czym świadczy wielkość statystyki R^2. Parametry modelu są łącznie istotne, gdyż $p - value$ statystyki F wynosi 0.000

2. Statystycznie istotne są wszystkie zmienne objaśniające, poza zmienną płeć, gdyż statystyki t są większe od 2, a ich *p-value* wynosi 0.

3. Zmienna zależna jest zlogarytmowana więc współczynniki należy interpretować jako semielastyczności. Kobiety zarabiają przeciętnie 15 % mniej niż mężczyźni, pracownicy firm prywatnych zarabiają przeciętnie 17 % więcej niż pracownicy firm państwowych w sektorze przedsiębiorstw, pracownicy form średniej wielkości zarabiają przeciętnie o 11% więcej od pracowników małych firm, pracownicy dużych firm zarabiają przeciętnie o 22 % więcej od pracowników małych firm. Wiek oraz doświadczenie zawodowe wpływa na uzyskiwane zarobki w sposób nieliniowy.

$$\frac{\partial lzarobki}{\partial wiek} = \beta_{wiek} + 2\beta_{wiek2}\bar{wiek}$$

$$\frac{\partial lzarobki}{\partial dosw} = \beta_{dosw} + 2\beta_{dosw2}\bar{dosw}$$

   Bez informacji o średnim wieku i doświadczeniu zawodowym w próbie nie można dokonać ilościowej interpretacji.

4. Założenie o poprawności formy funkcyjnej można weryfikować testem RESET. Wartość statystyki testowej wynosi 2.14, a jej *p-value* $Prob > F = 0.0534 > \alpha = 0.05$ wobec tego brak jest podstaw do odrzucenia hipotezy o poprawności formy funkcyjnej.

5. Na podstawie wyniku testu Breuscha-Pagana można stwierdzić, że brak jest podstaw do odrzucenia hipotezy o homoscedastyczności składnika losowego, gdyż *p-value* statystyki testowej wynosi $0.0886 > \alpha = 0.05$

6. Aby przetestować hipotezę o braku wpływu doświadczenia na wysokość zarobków należy sprawdzić czy współczynniki przy zmiennych $dosw$ oraz $dosw2$ wynosi zero. Mając oszacowania modelu bez ograniczeń można oszacować model z narzuconymi ograniczeniami, a następnie na podstawie sumy kwadratów reszt obu modeli zbudować statystykę $F = \frac{(RSS_R - RSS_U)}{RSS_U} / \frac{N-k}{J}$ i porównać z wartością krytyczną z rozkładu F.

7. Nie testowano występowania autokorelacji, ponieważ zjawisko autokorelacji jest związane z czasowym wymiarem danych, a parametry modelu zostały uzyskane na podstawie próby przekrojowej.

8. Na podstawie przeprowadzonych testów można uznać, że model spełnia załozenia KMRL, zatem uzyskane oszacowania, w myśl tw. Gaussa-Markowa, sa najlepszymi, liniowymi i nieobciążonymi estymatorami.