Egzamin z ekonometrii - wersja IiE, MSEMAT

27-01-2012

Pytania teoretyczne

1. Dlaczego w modelu nie powinno się umieszczać stałej i wszystkich zmiennych zero-jedynkowych, związanych z poziomami zmiennej dyskretnej?

2. Wyprowadzić rozkład małopróbkowy estymatora MNK. Jakie założenie, poza standardowymi założeniami KMRL, należy w tym przypadku przyjąć?

3. Udowodnić, że w modelu ze stałą $TSS = ESS + RSS$.

4. Kiedy mówimy, że zmienne w modelu są dokładnie współliniowe? Jak można rozwiązać ten problem?

Zadanie 1

Analizowano model z jedną zmienną objaśniającą i bez wyrazu wolnego:

$$y_i = \beta x_i + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2)$$

gdzie x jest nielosowe.

1. Wyznaczyć estymator MNK parametru β.
2. Obliczyć wariancję estymatora MNK parametru β.
3. Zaproponowano jako estymator parametru β wyrażenie w postaci $\frac{\bar{y}}{\bar{x}}$. Pokazać, że ten estymator jest nieobciążony.
4. Wyznaczyć wariancję estymatora parametru β postaci $\frac{\bar{y}}{\bar{x}}$.
5. Czy istnieje estymator parametru β postaci Cy, gdzie C jest pewnym wektorem a y jest wektorem zawierającym obserwacje zmiennej zależnej, który jest nieobciążony i ma mniejszą wariancję niż estymator uzyskany w punkcie 1? Odpowiedź należy uzasadnić.

Rozwiązanie Zadanie 1

1. Wyznaczyć estymator MNK parametru β.

$$X'X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}' \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \sum_{i=1}^{N} x_i^2$$

$$X'y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \sum_{i=1}^{N} x_i y_i$$

$$b = (X'X)^{-1} X'y = \left(\sum_{i=1}^{N} x_i^2\right)^{-1} \sum_{i=1}^{N} x_i y_i = \frac{\sum_{i=1}^{N} x_i y_i}{\left(\sum_{i=1}^{N} x_i^2\right)}$$

2. Obliczyć wariancję estymatora MNK parametru β.

   Zakładając, iż w modelu nie występuje autokorelacja składnika losowego:

   $Var(y_i) = Var(\beta x_i + \varepsilon_i)$ wiedząc, że $\{\varepsilon_i \sim N(0, \sigma^2)\}$ i x jest nielosowe otrzymujemy

   $Var(y_i) = \sigma^2$. Wobec tego:

   $$Var(b) = Var\left(\frac{\sum_{i=1}^N x_i y_i}{(\sum_{i=1}^N x_i^2)}\right) = \left[\frac{1}{(\sum_{i=1}^N x_i^2)}\right]^2 Var\left(\sum_{i=1}^N x_i y_i\right)$$
   $$= \left[\frac{1}{(\sum_{i=1}^N x_i^2)}\right]^2 \sum_{i=1}^N x_i^2 Var(y_i) = \left[\frac{\sigma^2}{(\sum_{i=1}^N x_i^2)}\right]$$

3. Zaproponowano jako estymator parametru β wyrażenie w postaci $\frac{\bar{y}}{\bar{x}}$. Pokazać, że ten estymator jest nieobciążony.

   $$E\left[\frac{\bar{y}}{\bar{x}}\right] = \frac{1}{\bar{x}} E(\bar{y}) = \frac{1}{\bar{x}} E(\frac{1}{N} \sum_{i=1}^N y_i)$$
   $$= \frac{1}{\bar{x}} \frac{1}{N} \sum_{i=1}^N E(y_i) = \frac{1}{\bar{x}N} \sum_{i=1}^N (x_i \beta) = \beta \frac{1}{\bar{x}N} \sum_{i=1}^N x_i = \beta$$

4. Wyznaczyć wariancję estymatora parametru β postaci $\frac{\bar{y}}{\bar{x}}$.

   Zakładając, iż w modelu nie występuje autokorelacja składnika losowego:

   $$Var\left[\frac{\bar{y}}{\bar{x}}\right] = \left[\frac{1}{\bar{x}}\right]^2 Var(\bar{y}) = \left[\frac{1}{\bar{x}}\right]^2 Var(\frac{1}{N} \sum_{i=1}^N y_i)$$
   $$= \left[\frac{1}{\bar{x}N}\right]^2 \sum_{i=1}^N \sigma^2 = \left[\frac{1}{\bar{x}N}\right]^2 N \cdot \sigma^2 = \frac{\sigma^2}{\bar{x}^2 N}$$

5. Czy istnieje estymator parametru β postaci Cy, gdzie C jest pewnym wektorem a y jest wektorem zawierającym obserwacje zmiennej zależnej, który jest nieobciążony i ma mniejszą wariancję niż estymator uzyskany w punkcie 1? Odpowiedź należy uzasadnić.

   Zakładając, iż w modelu nie występuje autokorelacja składnika losowego:

   model spełnia założenia twierdzenia Gaussa-Markowa, wobec tego estymator MNK w klasie liniowych (estymator postaci Cy) i nieobciążonych estymatorów ma najmniejszą wariancję. Nie istnieje estymator liniowy i nieobciążony, który miałby mniejszą wariancję niż estymator MNK.

Zadanie 2

Na podstawie danych BAEL z 2010 roku oszacowano długość trwania bezrobocia (*trwanie* - logarytm długości trwania bezrobocia). Zmiennymi objaśniającymi są wiek, miejsce zamieszkania (*miasto*: 0 - wieś, 1 - miasto), płeć (*plec*: 0 - mężczyzna, 1 - kobieta), wykształcenie (*educ*: 0 - podstawowe, 1 - średnie, 2 - wyższe), interakcja między płcią a wykształceniem. Oszacowania parametrów znajdują się poniżej. Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając *p-value*.

```
      Source |       SS       df       MS              Number of obs =    6048
-------------+------------------------------           F(  7,  6040) =   42.97
       Model |  313.722374     7  44.8174819           Prob > F      =  0.0000
    Residual |  6299.82817  6040  1.04301791           R-squared     =  0.0474
-------------+------------------------------           Adj R-squared =  0.0463
       Total |  6613.55054  6047  1.09369118           Root MSE      =  1.0213

------------------------------------------------------------------------------
     trwanie |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   .0148782   .0011217    13.26   0.000     .0126793    .0170772
  _Imiasto_1 |  -.0986615   .0269657    -3.66   0.000     -.151524    -.045799
    _Iplec_1 |    .062546   .0398614     1.57   0.117    -.0155966    .1406886
    _Ieduc_1 |  -.0614898    .040169    -1.53   0.126    -.1402353    .0172557
    _Ieduc_2 |  -.1784701   .0567782    -3.14   0.002    -.2897756   -.0671646
_IpleXedu_~1 |  -.1033812   .0586124    -1.76   0.078    -.2182824      .01152
_IpleXedu_~2 |  -.1639101   .0778669    -2.11   0.035    -.3165569   -.0112633
       _cons |   1.524151    .050608    30.12   0.000     1.424941    1.623361
------------------------------------------------------------------------------
RESET F(3,6037) = 11.00 [0.0000]
Jarque-Berra chi2(2) = 123.24 [0.0000]
White chi2(19) = 28.93 [0.0670]
Breusch-Pagan chi2(1) = 13.86 [0.0002]
```

1. Czy zmienne objaśniające są łącznie istotne?
2. Zinterpretować wartość współczynnika determinacji.
3. Ocenić, które zmienne są istotne.
4. Zinterpretować oszacowania parametrów przy zmiennych istotnych.
5. Zbadać, czy w modelu występuje heteroskedastyczność.
6. Zbadać, czy błąd losowy ma rozkład normalny.
7. Sprawdzić, czy forma funkcyjna modelu jest poprawna.
8. Jeżeli model nie spełnia założeń KMRL określić:

   (a) Które założenia nie sa spełnione?
   (b) Jakie to ma konsekwencje dla interpretacji modelu i wnioskowania statystycznego?
   (c) W jaki sposób można rozwiązać problemy zasygnalizowane przez wyniki testów?

Rozwiązanie Zadanie 2

1. Test na łączną istotność regresji: F =42.97, $p-value = 0.000 < 0.05$ odrzucamy hipotezę zerową o łącznej nieistotności regresji.
2. 4.74% zmienności czasu trwania bezrobocia zostało wyjaśnione za pomocą zmiennych niezależnych.
3. Istotne zmienne, to te dla których $p-value$ jest mniejsze od przyjętego poziomu istotności wynoszącego 0.05. Czyli istotne zmienne to:
   (a) *wiek* (t = 13.26, $p-value = 0.000$)
   (b) *educ_2* (t = -3.14, $p-value$= 0.002)
   (c) *plecXeduc_2* (t = -2.11, $p-value = 0.035$)
   (d) *miasto_1* (t = -3.66, $p-value = 0.000$)
4. Interpretacja oszacowań parametrów:
   (a) wzrost wieku o 1 rok powoduje wzrost czasu trwania bezrobocia średnio o 1.5% *ceteris paribus* (β_{wiek}).
   (b) mężczyźni z wykształceniem wyższym mają średnio o 17.8% krótszy czas trwania bezrobocia niż mężczyźni z wykształceniem podstawowym *ceteris paribus* (β_{educ_2}).
   (c) kobiety z wykształceniem wyższym mają średnio o 34.2% (16.4%+17.8%) krótszy czas trwania bezrobocia niż kobiety z wykształceniem podstawowym *ceteris paribus* ($\beta_{educ\_2}+\beta_{plecXeduc\_2}$).
   (d) osoby mieszkające w mieście mają średnio o 9.9% krótszy czas trwania bezrobocia niż osoby mieszkające na wsi *ceteris paribus* (β_{miasto_1}).
5. Występowanie heteroskedastyczności testujemy za pomocą:
   (a) testu White'a:
      i. hipoteza zerowa: homoskedastyczność składnika losowego.
      ii. wartość statystyki testowej wynosi: chi2(19)= 28.93 oraz $p-value = 0.0670 > 0.05$, więc brak podstaw do odrzuceniu hipotezy zerowej o homoskedastyczności.
   (b) testu Breuscha-Pagana:
      i. hipoteza zerowa: homoskedastyczność składnika losowego.
      ii. wartość statystyki testowej wynosi chi2(1) = 13.86 oraz $p-value = 0.0002 < 0.05$, więc odrzucamy hipotezę zerową o homoskedastyczności.
6. Normalność zaburzenia losowego testujemy za pomocą:
   (a) testu Jarque-Bera:
      i. hipoteza zerowa: zaburzenie losowe ma rozkład normalny.
      ii. wartość statystyki testowej wynosi chi2(2) = 123.24 oraz $p-value = 0.000 <$0.05, czyli odrzucamy hipotezę zerową o normalności zaburzenia losowego.
7. Poprawnosc przyjetej formy funkcyjnej modelu testujemy za pomocą:
   (a) test RESET:
      i. hipoteza zerowa: przyjęta postać funkcyjna modelu jest prawidłowa.
      ii. wartość statystyki testowej $F(3, 6037) = 11.00$ i $p-value = 0.0000 < 0.05$, więc odrzucamy hipotezę zerową o poprawności przyjętej formy funkcyjnej.
8. Odpowiedzi są następujące:
   (a) Nie jest spełnione założenie o homoskedastyczności zaburzenia losowego oraz założenie o sposobie „generowania danych": $y = \beta x + \varepsilon$ (czyli założenie o liniowej zaleności między zmienną zależną i zmiennymi niezależnymi). Nie jest także spełnione dodatkowe założenie o normalności składnika losowego.
   (b) Konsekwencje dla interpretacji modelu i wnioskowania statystycznego są następujące:

i. W przypadku nie spełnienia założenia o homoskedastyczności zaburzenia losowego, estymator b jest co prawda nieobciążony i zgodny, ale nieefektywny. Estymator macierzy wariancji-kowariancji b jest już obciążony i niezgodny. Macierz wariancji-kowariancji jest wykorzystywana do testowania hipotez na temat istotności zmiennych, więc poprawność wnioskowania statystycznego jest podważona.

ii. Odrzucenie hipotezy o poprawności przyjętej formy funkcyjnej podważa interpretację ekonomiczną modelu (interpretacja oszacowanych parametrów). Takie własności jak nieobciążoność czy efektywność estymatora MNK są wyprowadzane przy założeniu prawdziwości przyjętej formy funkcyjnej modelu.

iii. Próba zawiera 6048 obserwacji (można przyjąć, iż jest to duża próba). Dla dużych prób rozkłady statystyk są bliskie standardowym rozkładom.

(c) Rozwiązanie problemów zasygnalizowanych przez wyniki testów:

i. **Niepoprawna forma funkcyjna**: możemy próbować poprawić formę funkcyjną modelu wprowadzając do modelu interakcje między zmiennymi, dokonać przekształceń zmiennych (np. przekształcenie Boxa-Coxa), zastosować model wielomianowy, schodkowy lub krzywej łamanej.

ii. **Problem heteroskedastyczności** można rozwiązać za pomocą Stosowalnej UMNK lub odpornego estymatora White'a macierzy wariancji kowariancji.

Zadanie 3

W badaniu różnic wynagrodzeń pomiędzy pracownikami rodzimymi i obcokrajowcami o podobnym wieku i wykształceniu oszacowano następujące równanie:

$$W_i = \alpha + \beta D_i + \varepsilon_i,\ i = 1, 2, ..., n$$

gdzie:

- W_i to wynagrodzenie pracownika i-tego,
- D_i to zmienna zero-jedynkowa przyjmująca wartość 1 jeżeli pracownik jest obcokrajowcem i zero w przeciwnym wypadku.
- Przez $\bar{W}_{nat}$ i $\bar{W}_{non}$ oznaczono odpowiednio średnią płacę rodzimego pracownika i średnią płacę obcokrajowca.
- Przez n_{nat} i n_{non} oznaczono odpowiednio liczbę pracowników rodzimych i liczbę obcokrajowców.
- Przez $\bar{W}$ i $\bar{D}$ oznaczono odpowiednio średnie z W_i i D_i .

1. Pokazać, że zachodzą następujące relacje:

   (a) $\bar{W} = \frac{n_{non}\bar{W}_{non} + n_{nat}\bar{W}_{nat}}{n}$.

   (b) $\bar{D} = \frac{n_{non}}{n_{non} + n_{nat}}$

   (c) $\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{n_{non} \cdot n_{nat}}{n_{non} + n_{nat}}$

2. Pokazać, że estymator MNK zastosowany do modelu daje następujące oszacowania:

$$\hat{\alpha} = \bar{W}_{nat}\ i\ \hat{\beta} = \bar{W}_{non} - \bar{W}_{nat}$$

Rozwiązanie Zadanie 3

1. Wyprowadzenie:

   (a) $W_{nat} = \alpha + \varepsilon$, n_{nat} - liczba pracowników rodzimych.

   $W_{non} = \alpha + \beta + \varepsilon$, n_{non} - liczba obcokrajowców.

   $\hat{W}_{nat} = \hat{\alpha}$ i korzystając z własności hiperpłaszczyzny regresji otrzymujemy

   $\bar{\hat{W}}_{nat} = \bar{W}_{nat} = \hat{\alpha}$

   $\hat{W}_{non} = \hat{\alpha} + \hat{\beta}$ i korzystając z własności hiperpłaszczyzny regresji otrzymujemy

   $\bar{\hat{W}}_{non} = \bar{W}_{non} = \hat{\alpha} + \hat{\beta}$

   Wobec tego:

   $$\bar{W} = \bar{\hat{W}} = \frac{\sum_{i=1}^{n} \hat{W}_i}{n} = \frac{\sum_{i=1}^{n} (\hat{\alpha} + \hat{\beta} D_i)}{n} = \frac{n_{non}\left(\hat{\alpha} + \hat{\beta}\right) + n_{nat}\hat{\alpha}}{n} = \frac{n_{non}\bar{W}_{non} + n_{nat}\bar{W}_{nat}}{n}$$

   (b) $\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{n_{non}}{n_{non}+n_{nat}}$

   (c) $\sum_{i=1}^{n}(D_i - \bar{D})^2 = (1-\bar{D})^2 \cdot n_{non} + (0-\bar{D})^2 \cdot n_{nat} = (1 - \frac{n_{non}}{n_{non}+n_{nat}})^2 \cdot n_{non} + (\frac{n_{non}}{n_{non}+n_{nat}})^2 \cdot n_{nat} =$

   $= (\frac{n_{non}+n_{nat}-n_{non}}{n_{non}+n_{nat}})^2 \cdot n_{non} + (\frac{n_{non}}{n_{non}+n_{nat}})^2 \cdot n_{nat} = (\frac{n_{nat}}{n_{non}+n_{nat}})^2 \cdot n_{non} + (\frac{n_{non}}{n_{non}+n_{nat}})^2 \cdot n_{nat} =$

   $= \frac{n_{nat}^2 n_{non} + n_{non}^2 n_{nat}}{(n_{non}+n_{nat})^2} = \frac{n_{nat} n_{non}(n_{nat}+n_{non})}{(n_{non}+n_{nat})^2} = \frac{n_{non} \cdot n_{nat}}{n_{non}+n_{nat}}$

2. $W_{nat} = \alpha + \varepsilon$, n_{nat} - liczba pracowników rodzimych.

   $W_{non} = \alpha + \beta + \varepsilon$, n_{non} - liczba obcokrajowców.

   $\hat{W}_{nat} = \hat{\alpha}$ i korzystając z własności hiperpłaszczyzny regresji otrzymujemy $\bar{\hat{W}}_{nat} = \bar{W}_{nat} = \hat{\alpha}$

   $\hat{W}_{non} = \hat{\alpha} + \hat{\beta}$ i korzystając z własności hiperpłaszczyzny regresji otrzymujemy

   $\bar{\hat{W}}_{non} = \bar{W}_{non} = \hat{\alpha} + \hat{\beta}$

   Wobec tego: $\hat{\alpha} = \bar{W}_{nat}$ *i* $\hat{\beta} = \bar{W}_{non} - \bar{W}_{nat}$