Egzamin z ekonometrii 30.01.2009
1 semestr, Informatyka i Ekonometria

Pytania teoretyczne

1. Wyprowadzić estymator MNK dla modelu z wieloma zmiennymi objaśniającymi.
2. Wyprowadzić wzór na wariancję błędu prognozy.
3. Wyjaśnić, jakie korzyści i niebezpieczeństwa łączą się z narzucaniem ograniczeń na model.
4. Udowodnij, że estymator MNK jest estymatorem zgodnym. Wypisz założenia konieczne do tego dowodu.

ZADANIE 1 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 założenia KMRL, zatem uzyskane oszacowania, w myśl tw. Gaussa-Markowa, sa najlepszymi, liniowymi i nieobciążonymi estymatorami.

ZADANIE 2 Zakładamy, że analizujemy zależność między dwoma zmiennymi ciągłymi y i x oraz prawdziwy jest następujący model:

$$y_i = \beta x_i + \varepsilon_i,$$

gdzie $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 + D_i\sigma^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$. D_i jest zmienną zerojedynkową zakodowaną w następujący sposób:

$$D_i = \begin{cases} 0 & dla \quad x_i < 0 \\ 1 & dla \quad x_i \geq 0 \end{cases}.$$

1. Które z założeń KMRL nie jest spełnione dla tego modelu? Jeżeli do oszacowania nieznanego parametru β użyjemy estymatora MNK, to jakie będzie miał on własności?
2. Wiemy, że suma kwadratów wartości zmiennej x dla przypadku, gdy $x < 0$ wynosi 10 oraz to samo wyrażenie dla sytuacji gdy $x \geq 0$ przyjmuje wartość 20. Ponadto wiadomo, że $\sum x_i y_i = -20$ gdy $x < 0$ oraz $\sum x_i y_i = 30$ dla $x \geq 0$. Podaj wartość estymatora UMNK dla tego przypadku.

   Podpowiedź: Załóż dla uproszczenia, że obserwacje są posortowane według wartości zmiennej D_i.

Rozwiązanie:

1. Ponieważ $Var(\varepsilon_i) = \sigma^2 + D_i\sigma^2 = \begin{cases} \sigma^2 & dla \quad x_i < 0 \\ 2\sigma^2 & dla \quad x_i \geq 0 \end{cases}$, więc nie będzie spełnione założenie o homoscedastyczności. Ze względu na fakt, iż występuje problem heteroscedastyczność, to estymator MNK co prawda dalej jest nieobciążony, ale jest nieefektywny.

2. Zakładamy, że obserwacje w próbie są posortowane według rosnących wartości zmiennej x oraz $n$oznacza liczbę obserwacji dla których $x < 0$, a m liczbę obserwacji dla których $x \geq 0$. Macierz wariancji-kowariancji zaburzenia losowego ma następującą postać:

$$Var\left(\varepsilon\right) = \begin{bmatrix} \sigma^2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & \ddots & \sigma^2 & \ddots & & \vdots \\ \vdots & & \ddots & 2\sigma^2 & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 2\sigma^2 \end{bmatrix} = \sigma^2 \underbrace{\begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & 2 & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 2 \end{bmatrix}}_{\boldsymbol{V}}$$

Postać estymatora UMNK:

$$\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$

$$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} & x_{11} & \cdots & x_{1m} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \frac{1}{2} & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_{11} \\ \vdots \\ x_{1n} \\ x_{11} \\ \vdots \\ x_{1m} \end{bmatrix} =$$

$$= \begin{bmatrix} x_{11} & \cdots & x_{1n} & \frac{1}{2}x_{11} & \cdots & \frac{1}{2}x_{1m} \end{bmatrix} \begin{bmatrix} x_{11} \\ \vdots \\ x_{1n} \\ x_{11} \\ \vdots \\ x_{1m} \end{bmatrix} = \sum_{i=1}^{n} x_{1i}^2 + \frac{1}{2}\sum_{i=1}^{m} x_{2i}^2$$

$$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} = \begin{bmatrix} x_{11} & \cdots & x_{1n} & x_{11} & \cdots & x_{1m} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \frac{1}{2} & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ y_{11} \\ \vdots \\ y_m \end{bmatrix}$$

$$= \begin{bmatrix} x_{11} & \cdots & x_{1n} & \frac{1}{2}x_{11} & \cdots & \frac{1}{2}x_{1m} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ y_{11} \\ \vdots \\ y_m \end{bmatrix}$$

$$= \sum_{i=1}^{n} x_{1i}y_{1i} + \frac{1}{2}\sum_{i=1}^{m} x_{2i}y_{2i}$$

Ostatecznie otrzymujemy:

$$b = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} = \frac{\sum_{i=1}^{n} x_{1i}y_{1i} + \frac{1}{2}\sum_{i=1}^{m} x_{2i}y_{2i}}{\sum_{i=1}^{n} x_{1i}^2 + \frac{1}{2}\sum_{i=1}^{m} x_{2i}^2} = \frac{-20 + \frac{1}{2}\cdot 30}{10 + \frac{1}{2}\cdot 20} = -\frac{1}{4}$$

ZADANIE 3 Pokazać, że suma dźwigni dla wszystkich obserwacji wynosi K, gdzie K oznacza ilość szacowanych parametrów oraz, że wartość dźwigni jest zawsze większa od zera.
Podpowiedź:
Dżwignia $h_i = [\boldsymbol{P}]_{ii}$, gdzie $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$.
Rozwiązanie:
Korzystamy z definicji śladu jako sumy elementów diagonalnych macierzy oraz własności śladu:

$$\sum\nolimits_{i=1}^{n} h_i = \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right) = \operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{K\times K}\left(\boldsymbol{X}'\boldsymbol{X}\right)_{K\times K}\right) = \operatorname{tr}(\boldsymbol{I}_{K\times K}) = K.$$

Macierz $\boldsymbol{P}$ jest dodatnio określona, ponieważ macierz $\boldsymbol{X}'\boldsymbol{X}$ jest dodatnio określone, w konsekwencji $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ jest dodatnio określone a $\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ musi być dodatnio określone, ponieważ jeśli $\boldsymbol{v}\prime\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{v} > \boldsymbol{0}$ dla każdego $\boldsymbol{v} \neq \boldsymbol{0}$, to także $\boldsymbol{v}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{v} = \boldsymbol{v}^{*\prime}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{v}^{*} > 0$ dla $\boldsymbol{v}^{*} = \boldsymbol{v}\boldsymbol{X} \neq \boldsymbol{0}$. Zauważmy, że

$$h_i = [\boldsymbol{P}]_{ii} = \boldsymbol{\delta}_i'\boldsymbol{P}\boldsymbol{\delta}_i > \boldsymbol{0}$$

skoro $\boldsymbol{\delta}_i \neq \boldsymbol{0}$.