imię, nazwisko, nr indeksu:_____

Ekonometria — egzamin
wersja Informatyka i Ekonometria
26/06/08

1. Egzamin trwa 90 minut.
2. Rozwiązywanie zadań należy rozpocząć po ogłoszeniu początku egzaminu a skończyć wraz z ogłoszeniem końca egzaminu. Złamanie tej zasady skutkuje usunięciem z sali i unieważnieniem pracy.
3. W trakcie egzaminu wolno używać jedynie długopisu o innym kolorze atramentu niż czerwony oraz kalkulatora.
4. Przed przystąpieniem do pisania egzaminu należy podpisać **wszystkie kartki** pracy (na dole w przewidzianym miejscu). Złożenie podpisu pod regulaminem oznacza jego akceptację. Do egzaminu mogą przystąpić osoby, które akceptują regulamin.
5. W razie braku podpisu lub numeru zadania na kartce, kartka nie zostanie oceniona. Nie będą też oceniane rozwiązania wpisane na kartkach innych, niż te rozdawane przez prowadzących.
6. Rozwiązanie każdego zadania należy zapisać na kartce z tymże zadaniem, ewentualnie na czystych kartkach znajdujących się na końcu arkusza egzaminacyjnego lub na dodatkowych kartkach uzyskanych od prowadzących egzamin.
7. Na jednej kartce może znajdować się rozwiązanie tylko jednego zadania. Oceniane jest rozwiązanie tylko tego zadania, którego numer widnieje na kartce.
8. Egzamin składa się z czterech pytań teoretycznych i 3 zadań.
9. Posiadanie przy sobie wszelkich materiałów drukowanych (w tym książek) oraz innych np. wykonanych własnoręcznie materiałów zostanie uznane za ściąganie.
10. Rozmowy z innymi zdającymi będą traktowane identycznie jak ściąganie.
11. Każda zauważona próba ściągania skutkuje usunięciem z egzaminu.
12. Wszystkie pytania należy kierować bezpośrednio do osób pilnujących.
13. Warunkiem uzyskania oceny pozytywnej jest zdobycie conajmniej 50 % punktów z części teoretycznej egzaminu oraz min. 40 % punktów z cześci zadaniowej.

Warszawa, 26/06/2008, _____
podpis

Powodzenia :-)

1	2	$\sum$

Pytania teoretyczne

Pytanie 1.
Pokaż, że w modelu regresji ze stałą $TSS = ESS + RSS$

Całkowita suma kwadratów w modelu jest równa

$$TSS = \sum (y_i - \bar{y})^2 = (y - \bar{y})'(y - \bar{y})$$

z kolei $y = \hat{y} + e$. Odejmując $\bar{y}$ od obu stron i pamiętając, że $\bar{y} = \bar{\hat{y}}$ uzyskujemy

$$y - \bar{y} = \hat{y} - \bar{\hat{y}} + e$$

wstawiając do wzoru na całkowitą sumę kwadratów otrzymujemy

$$TSS = (y - \bar{y})'(y - \bar{y}) = [\hat{y} - \bar{\hat{y}} + e]'[\hat{y} - \bar{\hat{y}} + e]$$

$$TSS = \underbrace{[\hat{y} - \bar{\hat{y}}]'[\hat{y} - \bar{\hat{y}}]}_{ESS} + \underbrace{e'e}_{RSS} + 2\underbrace{[\hat{y} - \bar{\hat{y}}]'e}_{0}$$

Zerowość ostatniego składnika wynika z własności hiperpłaszczyzny regresji.

Pytanie 2.
Jakie są konsekwencje występowania niedokładnej współliniowości dla precyzji oszacowań parametrów

W rezultacie niedokładnej współliniowości:

1. współczynniki równania regresji będą miały duże błędy standardowe, oraz mogą być nieistotne statystycznie, nawet gdy łącznie są istotne, a współczynnik R^2 modelu jest wysoki [1]
2. niewielkie zmiany w zbiorze danych będą powodować duże zmiany w otrzymywanych oszacowaniach parametrów [0.125].
3. współczynniki równania regresji mogą mieć „złe", czyli niezgodne z teorią znaki, albo są zbyt małe lub zbyt duże [0,125].

3	4	$\sum$

Pytania teoretyczne c.d.

Pytanie 3.
Wyjaśnij różnicę między UMNK a SUMNK

Aby można było zastosować UMNK musi być znana postać macierzy wariancji-kowariancji $\mathbb{V}$. W rzeczywistości postać macierzy $\mathbb{V}$ jest znana tylko w nielicznych przypadkach. W praktyce nieznaną macierz zastępuje się jej oszacowaniem $\hat{\mathbb{V}}$ i taką metodę nazywa się SUMNK.

Pytanie 4.
Opisz dwa sposoby przeprowadzenia testu RESET.

- Przeprowadzana jest regresja
$$y = X\beta + \varepsilon$$
- oraz regresja pomocnicza
$$y = X\beta + Z\gamma + \varepsilon$$
- procedura testowa jest analogiczna do testu łącznej istotności
$$y = f(X\beta) = \gamma_0 + \gamma_1 X\beta + \gamma_2 (X\beta)^2 + \ldots + \gamma_p (X\beta)^p + \varepsilon$$
- alternatywna postać testu wykorzystuje to samo rozwinięcie w szereg Taylora ale przyjmując że $\hat{y} = X\hat{\beta}$
$$y = f(X\beta) = \gamma_0 + \gamma_1 \hat{y} + \gamma_2 \hat{y}^2 + \ldots + \gamma_p \hat{y}^p + \varepsilon$$
- test łącznej istotności. Niech $\gamma = [\gamma_1, \ldots, \gamma_k]'$
$$H_0 : \ \gamma = 0 \qquad H_A : \ \gamma \neq 0$$
$$LM = nR^2 \ \sim_a \ \chi^2(p)$$

1	2	3	4	5	$\sum$

ZADANIE 1 Na podstawie 120 elementowej próby oszacowano model regresji liniowej z czterema zmiennymi objaśniającymi. Użyta macierz regresorów zawiera stałą, $R^2 = 0,5$. Następnie w celu uzyskania większej precyzji oszacowania dokonano transformacji oryginalnych zmiennych odejmując od nich wartość średnią $y_i^* = y_i - \bar{y}$, $x_{ki}^* = x_{ki} - \bar{x_k}$ i ponownie oszacowano model.

(a) Jaka będzie relacja między oszacowaniami w pierwotnym i przekształconym modelu?

(b) Jak transformacja wpłynie na wielkość współczynnika R^2?

(c) Jak transformacja wpłynie na wielkość estymatora wariancji?

Po wykonaniu tych czynności wykryto w nieprzekształconej macierzy obserwacji *outlier*, który usunięto. W wyniku tego działania resztowa suma kwadratów zmniejszyła się o 5 %, a całkowita suma kwadratów o 2 %.

(d) Jak usunięcie obserwacji wpłynie na wielkość R^2 oraz $\bar{R}^2$?

(e) Oblicz R^2 oraz $\bar{R}^2$ dla nowego modelu

1	2	3	4	5	$\sum$

ZADANIE 2 Mamy następujący model:

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

Wiemy, że $E(x_i\varepsilon_i) = \sigma_{x\varepsilon} \neq 0$, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ oraz, że próba jest losowa. Załóżmy, że mamy w próbce takie z_i, że $E(z_i\varepsilon_i) = 0$, ale $E(z_i x_i) = \sigma_{zx} \neq 0$. Załóżmy ponadto, że średnie w próbie dla $x_i\varepsilon_i$, $z_i\varepsilon_i$, $z_i x_i$, zgodnie z prawem wielkich liczb, dążą do odpowiednich wartości oczekiwanych.

1. Proszę pokazać, że estymator MNK w tym modelu nie jest zgodny
2. Jakiego rodzaju problem pojawia się w modelu analizowanym w poprzednim punkcie?
3. Proszę pokazać, że estymator postaci:

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} z_i y_i}{\sum_{i=1}^{N} x_i z_i}$$

   jest zgodny.

4. Policz wariancję estymatora analizowanego w poprzednim punkcie.

1	2	3	4	5	$\sum$

ZADANIE 3 Na podstawie danych z czwartego kwartału 2006 z Badania Aktywności Ekonomicznej Ludności oszacowano parametry równania płac dla Polski, uzyskując następujące wyniki:

```
      Source |       SS       df       MS              Number of obs =    9957
-------------+------------------------------           F(  5,  9951) =  359.95
       Model |  387.897104     5  77.5794209           Prob > F      =  0.0000
    Residual |  2144.69627  9951  .215525703           R-squared     =  0.1532
-------------+------------------------------           Adj R-squared =  0.1527
       Total |  2532.59337  9956  .254378603           Root MSE      =  .46425

------------------------------------------------------------------------------
    lzarobki |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kobieta |  -.2123398   .0095637   -22.20   0.000    -.2310866    -.193593
        wiek |   .2723319    .012117    22.48   0.000     .2485802    .2960836
       wiek2 |  -.0190553   .0009228   -20.65   0.000    -.0208641   -.0172465
    prywatna |  -.1594339   .0101201   -15.75   0.000    -.1792713   -.1395964
    wielkosc |   .0609763   .0039596    15.40   0.000     .0532148    .0687379
       _cons |   6.102562   .0397464   153.54   0.000     6.024651    6.180473
------------------------------------------------------------------------------
Breusch-Pagan     chi2(5)      =   486.85            Prob > chi2  = 0.0000
Ramsey RESET      F(8, 9944) =     3.78            Prob > F =     0.3446
Jarque-Bera       chi2(2)     =  12567.92            Prob > chi2  = 0.0000
```

Zmienna zero-jedynkowa *kobieta* przyjmuję wartość 1 dla kobiety, *wiek* to wiek pracownika w latach, *wiek*2 to kwadrat wieku, *prywatna* to zmienna zero-jedynkowa, gdzie wartość 1 oznacza firmę prywatną, *wielkosc* oznacza wielkość firmy obliczaną jako ilość zatrudnionych pracowników.

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

1. Oceń dopasowanie danych empirycznych do modelu. Określ które zmienne można uznać za statystycznie istotne?
2. Czy forma funkcyjna oszacowanego powyżej modelu jest liniowa?
3. Czy powyższy model spełnia założenia KMRL (wymień, które spełnia, a których nie spełnia)
4. Zaproponuj sposób weryfikacji hipotezy o braku wpływu wieku pracownika na wysokość uzyskiwanych zarobków.
5. 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 wartość statystyki testowej lub jej *p-value,* oraz jej interpretację.

Rozwiązania zadań

Zadanie 1

ad. (a)

$$y_i = x_i b + e_i$$

Odjęcie wartości średniej nie wpływa na wielkość, więc

$$y_i - \bar{y_i} = x_i b + e_i - \bar{y_i} = x_i b - \bar{x_i} b + e_i$$

$$y_i^* = x_i^* b + e_i$$

ad. (b) Ponieważ wektor współczynników się nie zmieni, to reszt się nie zmieni, suma kwadratów reszt pozostanie taka sama, więć R^2 też się nie zmieni.

ad. (c) Ponieważ wektor reszt się nie zmieni, to estymator wariancji również pozostanie bez zmian.

ad. (d) $R^2 = 1 - \frac{RSS}{TSS}$. Resztowa suma kwadratów spadnie o 5 %, całkowita suma o 2%, więc R^2 zwiększy się.

ad. (e) $R^2 = 1 - \frac{475}{980} = 0.5153$, $\bar{R^2} = 1 - \frac{119}{116}(1 - 0.5153) = 0.5028$

Zadanie 2

1. Estymator MNK w tym modelu ma granicę:

$$\begin{aligned} \text{plim} b &= \text{plim} \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} = \frac{\beta \text{plim} \frac{1}{N} \sum_{i=1}^N x_i^2 + \text{plim} \frac{1}{N} \sum_{i=1}^N x_i \varepsilon_i}{\text{plim} \frac{1}{N} \sum_{i=1}^N x_i^2} \\ &= \beta + \frac{\sigma_{x\varepsilon}}{\sigma_x^2} \end{aligned}$$

2. Problem równoczesności

3. Drugi rozważany estymator ma granicę

$$\begin{aligned} \text{plim} \widehat{\beta} &= \text{plim} \frac{\sum_{i=1}^N z_i y_i}{\sum_{i=1}^N x_i z_i} = \frac{\beta \text{plim} \frac{1}{N} \sum_{i=1}^N x_i z_i + \text{plim} \frac{1}{N} \sum_{i=1}^N z_i \varepsilon_i}{\text{plim} \frac{1}{N} \sum_{i=1}^N x_i z_i} \\ &= \beta \end{aligned}$$

4.

$$\begin{aligned} Var\left(\widehat{\beta}\right) &= Var\left(\frac{\sum_{i=1}^N z_i y_i}{\sum_{i=1}^N x_i z_i}\right) = \frac{1}{\left(\sum_{i=1}^N x_i z_i\right)^2} Var\left(\sum_{i=1}^N x_i z_i + \sum_{i=1}^N z_i \varepsilon_i\right) \\ &= \frac{\sum_{i=1}^N z_i^2}{\left(\sum_{i=1}^N x_i z_i\right)^2} \sigma^2 \end{aligned}$$

Zadanie 3

1. Zmienne objaśniające wyjaśniają wariancję zarobków w 11 %, o czym świadczy wielkość statystyki R^2. Statystycznie istotne są wszystkie zmienne objaśniające, poza zmienną płeć, gdyż statystyki t są większe od 2, a ich *p-value* wynosi 0.

2. Założenie o poprawności formy funkcyjnej można weryfikować testem RESET. Do modelu dodawany jest zbiór dodatkowych zmiennych Z i sprawdzana jest statystyczna istotność współczynników przy dodatkowych zmiennych (oznaczmy je γ) $H_0: \gamma = 0$ - forma funkcyjna jest poprawna, $H_A: \gamma \neq 0$ - forma funkcyjna jest niepoprawna. Wartość statystyki testowej wynosi 3.78, a jej *p-value* $Prob > F = 0.3446$ wobec tego należy odrzucić hipotezę o poprawności formy funkcyjnej.

3. Model spełnia założenie o liniowości (test RESET), nie są spełnione założenie o normalności (test JB), o homoscedastyczności (White,Breusch-Pagan). Założenie o nielosowości macierzy obserwacji jest nietestowalne.

4. Aby przetestować hipotezę o braku wpływu wieku na wysokość zarobków należy sprawdzić czy suma współczynników przy zmiennych *wiek* oraz *wiek*2 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.

5. Składnik losowy modelu jest heteroscedastyczny, więc uzyskane oszacowania MNK są nieobciążone ale nieefektywne. Można zastosować odporną na występowanie heteroscedastyczności metodę estymacji, np. macierz White'a. Dodatkowym problemem mogę być nieprawidłowo obliczone statystyki t-studenta i błędne wnioski odnośnie istotności oszacowań.