Egzamin z ekonometrii 25.06.2007
I semestr

Pytania teoretyczne

1. Wyprowadzić estymator MNK z zadania na minimalizację sumy kwadratów reszt. Skąd bierze się nazwa Metoda Najmniejszych Kwadratów (MNK)?
2. Podać postać estymatora dla kombinacji liniowej parametrów $\boldsymbol{\delta}'\boldsymbol{\beta}$ i udowodnić, że jest on nieobciążony.
3. Dlaczego silna współliniowość zmiennych może być problemem w badaniu statystycznym?
4. Kiedy mówimy o występowaniu autokorelacji w modelu? Jakie założenie $KMRL$ nie jest spełnione w przypadku występowania autokorelacji?

ZADANIE 1 Na podstawie danych GUS zbudowano KMRL wyjaśniający poziom wydatków na żywność za pomocą dochodu na głowę wyrażonego w złotych przeliczonego według skali oksfordzkiej, liczby osób dorosłych i dzieci w rodzinie oraz zmiennej zerojedynkowej przyjmującej wartość 1 dla gospodarstwa domowego wiejskiego. Szacowano następujący model:

$$\text{wydatki}_i=\text{stała}+\beta_1\text{dochód}_i+\beta_2\text{dorosłe}_i+\beta_3\text{dzieci}_i+\beta_4\text{wieś}_i+\varepsilon_i$$

i otrzymano wyniki:

```
      Source |       SS       df       MS        Number of obs =   29771
-------------+------------------------------     F(  4, 29766) =  .
       Model |   828538978     4   207134745     Prob > F      =  .
    Residual |   956257715 29766  32125.8387     R-squared     =  0.4642
-------------+------------------------------     Adj R-squared =  0.4641
       Total |  1.7848e+09 29770  59952.8617     Root MSE      =  179.24

------------------------------------------------
     wydatki |      Coef.   Std. Err.      t
-------------+----------------------------------
      dochod |   .1289331   .0021889       .
     dorosle |   135.2956   1.100277       .
      dzieci |    104.175   .6937367       .
        wies |   28.26728   2.287173       .
       _cons |   47.61749   3.397637       .
------------------------------------------------
```

1. Uzupełnić brakujące wielkości w tabeli.
2. Zinterpretować otrzymane wyniki.
3. Przeprowadzić testy istotności dla każdej ze zmiennych w modelu (poaza stała) i test łącznej istotności ($F_{0.95}(4, 29766) = 2,37$).
4. Ocenić dopasowanie modelu do danych empirycznych.
5. Chcemy przetestować hipotezę, czy tylko liczba osób w gospodarstwie domowym ma wpływ na wydatki żywnościowe a nie to, czy są to dorośli, czy dzieci. Zapisać tę hipotezę zerową oraz hipotezę alternatywną, a także wyjaśnić w jaki sposób można ją zweryfikować.

Przyjąć poziom istotności $\alpha = 0.05$.

Rozwiązanie:

1. Statystyki t oblicza się ze wzoru $t = \frac{\beta}{se(\beta)}$. Statystykę F oblicza się ze wzoru $F = \frac{(S_R-S)/g}{S/(N-K)}$. W modelu w którym występuje tylko stała $b = \overline{y}$ a $S_R = (\boldsymbol{y}-\overline{\boldsymbol{y}})'(\boldsymbol{y}-\overline{\boldsymbol{y}}) = TSS$. Zatem $F = \frac{(TSS-RSS)/(K-1)}{RSS/(N-K)} = \frac{N-K}{K-1}\frac{ESS}{RSS}$

```
      Source |       SS       df       MS        Number of obs =   29771
-------------+------------------------------     F(  4, 29766) = 6448.17
       Model |   828538978     4   207134745     Prob > F      =  0.0000
    Residual |   956257715 29766  32125.8387     R-squared     =  0.4642
-------------+------------------------------     Adj R-squared =  0.4641
       Total |  1.7848e+09 29770  59952.8617     Root MSE      =  179.24
------------------------------------------------
     wydatki |      Coef.   Std. Err.      t
-------------+----------------------------------
      dochod |   .1289331   .0021889    58.90
     dorosle |   135.2956   1.100277   122.97
      dzieci |    104.175   .6937367   150.17
        wies |   28.26728   2.287173    12.36
       _cons |   47.61749   3.397637    14.01
------------------------------------------------
```

2. Wzrost dochodu o złotówkę powoduje przeciętny wzrost wydatków na żywność o 12 groszy. Dodatkowa osoba dorosła w gospodarstwie domowym przeciętnie zwiększa wydatki na żywność o 135.30 zł. Dodatkowe dziecko zwiększa wydatki na żywność o 104.18 zł. Gospodarstwa domowe na wsi przeciętnie wydają na żywność o 28.27 zł więcej niż pozostałe.

3. Wszystkie zmienne są indywidualnie istotne, ponieważ wszystkie statystyki t są co do wartości bezwzględnej większe od 2, zmienne są również łącznie istotne bowiem wartość statystyki testowej jest znacznie większa od wartości krytycznej równej 2.37.

4. Zmienność wydatków można w 46% wyjaśnić za pomocą modelu.

5. Liczba osób w gospodarstwie domowym to suma dorosłych oraz dzieci. Zatem hipotezy dotycząca braku wpływu liczby osób na wydatki żywnościowe wygląda następująco: $H_0 : \beta_2 = \beta_2$ przy hipotezie alternatywnej $H_0 : \beta_2 \neq \beta_3$. Szacowane są dwa modele regresji, jeden z narzuconymi ograniczeniami (w modelu tym zmienną objaśniającą jest zmienna dzieci+dorosłe) i drugi bez ograniczeń i za pomocą testu F sprawdza się czy różnica sum kwadratów reszt w obu modeli jest statystycznie istotna.

ZADANIE 2 Na podstawie części próby PGSS z 1997 roku przeprowadzono regresję logarytmów dochodów z pracy (miesięczne wynagrodzenie) na wykształceniu liczonym w latach (educ), płci (sex: 1 - mężczyzna, 2 - kobieta) i uzyskano następujące wyniki:

```
      Source |       SS       df       MS              Number of obs =     343
-------------+------------------------------           F(  2,   340) =    5.58
       Model |  1.2493e+09     2   624635356           Prob > F      =  0.0041
    Residual |  3.8094e+10   340   112040285           R-squared     =  0.0318
-------------+------------------------------           Adj R-squared =  0.0261
       Total |  3.9343e+10   342   115037917           Root MSE      =   10585

------------------------------------------------------------------------------
     rincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |   1469.366    1161.27     1.27   0.207    -814.8126    3753.545
       educ  |   -735.175   225.1851    -3.26   0.001    -1178.106   -292.2436
       _cons |   9776.691   2636.692     3.71   0.000     4590.408    14962.97
------------------------------------------------------------------------------
```

1. Wyjaśnić dlaczego uzyskane wyniki regresji są sprzeczne z intuicją.

2. Dla wszystkich obserwacji wyliczono standaryzowane reszty, statystyki dźwigni oraz statystyki Cook'a. Obserwacje, dla których statystyki te były największe znajdują się poniżej. Wyjaśnij, które obserwacje budzą podejrzenia i dlaczego.

	rincome	sex	educ	leverage
1.	99998	1	4	.0289198
2.	1100	1	17	.0212676
3.	700	1	17	.0212676
4.	500	1	17	.0212676
5.	400	1	17	.0212676
6.	400	1	17	.0212676
7.	1800	1	17	.0212676

	rincome	sex	educ	rstandard
1.	99998	2	10	9.114934
2.	99998	2	8	9.000631
3.	99998	2	8	9.000631
4.	99998	1	4	8.931496
5.	20000	1	10	1.66571
6.	3500	1	17	.5941019
7.	2200	1	17	.4699583

	rincome	sex	educ	cook
1.	99998	1	4	.7918933
2.	99998	2	8	.3650551
3.	99998	2	8	.3650551
4.	99998	2	10	.2177331
5.	20000	1	10	.0059378
6.	3500	1	17	.0025566
7.	2200	1	17	.0015997

3. Jakie zmiany powinno się wprowadzić, by uzyskać lepsze oszacowania parametrów w modelu?

Rozwiązanie:

1. Zgodnie z intuicją parametr przy zmiennej _Isex_2 powinien być ujemny i istotny (ze względu na dyskryminację kobiet na rynku pracy), zaś przy edukacji dodatni i istotny (ze względu na wyższe wynagrodzenie osób lepiej wykształconych)

2. Największe podejrzenia budzą obserwacje, dla których rincome=99998. Dla tych obserwacji zarówno standaryzowane reszty jak i statystyki cooka są bardzo wysokie. Świadczy to o tym, że obserwacje te słabo pasują do regresji ale bardzo silnie wpływają na jej wyniki. Dziwna dla tych obserwacji jest zarówno wysokość dochodu (zbyt wysoka jak na wielkość analizowanej próby i rok 1997), jak i fakt, że wszystkie 4 dotyczą dokładnie takiego samego deklarowanego dochodu (co do złotówki 99998). Można podejrzewać, że nie są to rzeczywiste obserwacje dla dochodu ale kody braków (np. odpowiedź: nie wiem).

3. Jeśli rzeczywiście obserwacje, dla których rincome=99998 są błędne, to powinniśmy je usunąć z modelu.

ZADANIE 3 Pokaż, że jeśli $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}$ i $\boldsymbol{A}$ jest nieosobliwe, to dla znanego estymatora MNK parametrów $\boldsymbol{\beta}$, estymatorem MNK parametrów $\boldsymbol{\theta}$ jest $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}^{-1}\,\boldsymbol{b}$.

Rozwiązanie:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon = \boldsymbol{X}\boldsymbol{A}\boldsymbol{\theta} + \varepsilon = \boldsymbol{X}^{*}\boldsymbol{\theta} + \varepsilon$$

gdzie $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{A}$. Estymatorem MNK parametru $\boldsymbol{\theta}$ jest więc

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\,\boldsymbol{y} = \left(\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{A}^{-1}\left('\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{b}.\end{aligned}$$