Egzamin z ekonometrii - wersja ogólna

7-02-2013

Pytania teoretyczne

1. Dlaczego zmienną dyskretną rozkodowujemy na zmienne zerojedynkowe?

2. Wyprowadzić estymator MNK dla modelu z wieloma zmiennymi objaśniającymi. Proszę również pokazać warunek dostateczny na istnienie minimum funkcji i udowodnić, że jest to rzeczywiście minimum tej funkcji.

3. Wyjaśnić różnicę między parametrami i oszacowaniami parametrów oraz między błędami losowymi i resztami.

4. Wyjaśnić, dlaczego R^2 nie można używać do porównywania modeli.

Zadanie 1

Przeanalizowano zróżnicowania wysokości wynagrodzenia dla pracowników w mikroprzedsiębiorstwach w Polsce (*ln_placa* - logarytm wynagrodzenia). Wśród potencjalnych predyktorów uwzględniono takie zmienne jak płeć respondenta (*plec*: 0 - mężczyzna, 1 - kobieta) i wykształcenie (*wyksz*: 0 - podstawowe, 1 - zawodowe, 2 - średnie, 3 - wyższe).

1. Na podstawie poniższych informacji, zawartych w Tabelach 1 - 3, przeprowadzić wstępną analizę zmiennej płeć respondenta (*plec*: 0 - mężczyzna, 1 - kobieta). Zinterpretować przedstawione wyniki i skomentować w jakim celu przeprowadzono poszczególne testy. Czy wszystkie poniższe testy mają zastosowanie w tym przypadku? Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając *p-value*.

Tabela 1. Test Jarque-Berra dla zmiennej ln_placa według zmiennej plec

```
->plec = kobieta
                    Skewness/Kurtosis tests for Normality
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    ln_placa |    1041   0.0000         0.0000       331.80         0.0000

->plec = mężczyzna
                     Skewness/Kurtosis tests for Normality
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    ln_placa |    927      0.0000         0.0000      1346.64         0.0000
```

Tabela 2. Test t-Studenta dla zmiennej ln_placa według zmiennej plec

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     927    7.216902      .02337    .7115377    7.171038    7.262767
       1 |    1041    6.962549    .0151675    .4893736    6.932787    6.992311
---------+--------------------------------------------------------------------
combined |    1968    7.082359    .0139156    .6173274    7.055068     7.10965
---------+--------------------------------------------------------------------
    diff |            .2543534    .0272887                .2008355    .3078713
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   9.3208
Ho: diff = 0                                     degrees of freedom =     1966
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Tabela 3. Test Wilcoxona dla zmiennej ln_placa według zmiennej plec

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
        plec |      obs    rank sum    expected
-------------+---------------------------------
           0 |      927     1061375    912631.5
           1 |     1041      876121   1024864.5
-------------+---------------------------------
    combined |     1968     1937496     1937496

Ho: ln_placa(plec==0) = ln_placa(plec==1)
             z =  11.846
    Prob > |z| =   0.0000
```

2. Na podstawie poniższych informacji, zawartych w Tabelach 4 - 6, przeprowadzić wstępną analizę zmiennej wykształcenie (*wyksz*: 0 - podstawowe, 1 - zawodowe, 2 - średnie, 3 - wyższe). Zinterpretować przedstawione wyniki i skomentować w jakim celu przeprowadzono poszczególne testy. Czy wszystkie poniższe testy mają zastosowanie w tym przypadku? Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając *p-value*.

Tabela 4. Test Jarque-Berra dla zmiennej ln_placa według zmiennej wyksz

```
-> wyksz=podstawowe
                        Skewness/Kurtosis tests for Normality
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    ln_placa |    197      0.0000         0.0011        49.95         0.0000

-> wyksz=zawodowe
                        Skewness/Kurtosis tests for Normality
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    ln_placa |   1185    0.0000         0.0000      1921.86         0.0000

-> wyksz=srednie
                     Skewness/Kurtosis tests for Normality
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    ln_placa |    326      0.0000         0.0000        65.22         0.0000

-> wyksz=wyzsze
                       Skewness/Kurtosis tests for Normality
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
    ln_placa |    259      0.0038         0.0000        29.22         0.0000
```

Tabela 5. Analiza wariancji zmiennej ln_placa według zmiennej wyksz

```
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------------
Between groups      26.8375373      3   8.94584578     24.31     0.0000
 Within groups      722.772635   1964   .368010506
------------------------------------------------------------------------------
    Total           749.610172   1967   .381093122
Bartlett's test for equal variances:  chi2(3) =  89.0791  Prob>chi2 = 0.000
```

Tabela 6. Test Kruskala-Wallisa dla zmiennej ln_placa według zmiennej wyksz

```
  +-------------------------+
  | educ |  Obs |  Rank Sum |
  |------+------+-----------|
  |    0 |  197 | 153284.00 |
  |    1 | 1186 |  1.17e+06 |
  |    2 |  326 | 305471.50 |
  |    3 |  259 | 308978.00 |
  +-------------------------+
chi-squared =    63.135 with 3 d.f.
probability =     0.0001

chi-squared with ties =    63.410 with 3 d.f.
probability =     0.0001
```

Rozwiązanie Zadanie 1

1. W celu przyjrzenia się jaki wpływ na wynagrodzenie mają poszczególne zmienne, przeprowadzono wstępną analizę danych. Sprawdzono, kto ma większe średnie wynagrodzenia, kobiety czy mężczyźni? Kobiety mają średnie zarobki ($\overline{ln\_placa} = 6.962549$) niższe niż mężczyźni ($\overline{ln\_placa} = 7.216902$). Można spróbować potwierdzić te spostrzeżenie za pomocą formalnego testu na równość średnich.

   - Na podstawie testu Jarque-Berra można wywnioskować, że zmienna *ln_placa* zarówno dla *kobiet* (*p-value*=0.000) jak i dla *mężczyzn* (*p-value*=0.000) nie ma rozkładu normalnego. Jednak próba zawiera 1041 obserwacji dla *kobiet* i 927 dla *mężczyzn* można więc przyjąć, iż w obu przypadkach jest to duża próba. Dla dużych prób rozkłady statystyk są bliskie standardowym rozkładom.
   - W związku z powyższym można interpretować wyniki testu *t-Studenta*: na podstawie *p-value*=0.000 odrzucono hipotezę zerową o równości średniego wynagrodzenia w obu podpróbach (kobiet i mężczyzn).
   - Ze względu na fakt, iż próba jest wystarczająco duża, zastosowano test t-Studenta. Nie ma potrzeby użycia testu Wilcoxona.

2. W celu przyjrzenia się jaki wpływ na wynagrodzenie mają poszczególne zmienne, przeprowadzono wstępną analizę danych. Zmienna wykształcenie (*wyksz*: 0 - podstawowe, 1 - zawodowe, 2 - średnie, 3 - wyższe) przyjmuje 4 kategorii. Porównano średnie wynagrodzenie osób posiadających różne poziomy wykształcenia.

   W celu zastosowania parametrycznej analizy wariancji ANOVA zmiennej *ln_placa* muszą być spełnione następujące warunki:

   (a) zmienna *ln_placa* musi mieć rozkład normalny w podpróbkach,

   (b) wariancja zmienej *ln_placa* musi być równa w podpróbkach,

   (c) obserwacje są niezależne.

- Wymienione założenia są niezbędne do wyznaczenia rozkładu statystyki testowej. Przy spełnieniu tych założeń statystyka testowa ma rozkład Fishera-Snedecora.
- Gdy próbki są równoliczne, test Fishera-Snedecora jest odporny na odchylenia od normalności i jednorodności wariancji.
- W przypadku gdy rozkłady mocno odbiegają od normalnego, albo wariancje znacznie się różnią, należy posłużyć się metodą nieparametryczną nazywaną testem Kruskala-Wallisa.

  Odnośnie (a): na podstawie testu Jarque-Berra można wywnioskować, że zmienna *ln_placa* zarówno dla osób z wykształceniem *podstawowym* (*p-value*=0.000), *zawodowym* (*p-value*=0.000), *średnim* (*p-value*=0.000) jak i *wyższym* (*p-value*=0.000) nie ma rozkładu normalnego. Jednak próba zawiera wystarczającą liczbę obserwacji, można więc przyjąć, iż w obu przypadkach jest to duża próba. Dla dużych prób rozkłady statystyk są bliskie standardowym rozkładom.

  Odnośnie (b): na podstawie testu Bartletta można wywnioskować, iż wariancja w podpróbkach nie jest równa (*p-value*=0.000).

  Jednocześnie warto zwrócić uwagę, że liczebności w podpróbkach nie są do siebie zbliżone.

Wniosek:
W związku z powyższym nie ma zastosowania parametryczna analiza wariancji ANOVA lecz nieparametryczny test Kruskala-Wallisa (hipotezą zerową jest równość dystrybuant rozkładów w porównywanych próbach). Na podstawie *p-value*=0.000 dla tego testu odrzucono hipotezę zerową o równości dystrybuant wynagrodzenia w podpróbach.

Zadanie 2

Na podstawie danych BAEL z 2010 roku oszacowano wysokość wynagrodzenia dla pracowników w mikroprzedsiębiorstwach (*ln_placa* - logarytm wynagrodzenia). Zmiennymi objaśniającymi są wiek (*wiek*), wiek do kwadratu (*wiek_2*), miejsce zamieszkania (*wies*: 0 - miasto, 1 - wieś), płeć (*plec*: 0 - mężczyzna, 1 - kobieta), wykształcenie (*wyksz*: 0 - podstawowe, 1 - zawodowe, 2 - średnie, 3 - wyższe), wymiar zatrudnienia (*etat*: 0 - niepełny etat, 1 - cały etat), stan cywilny (*stan*: 0 - kawaler / panna, 1 - zamężny / zamężna, 2 - wdowiec / wdowa, 3 - rozwiedziony / rozwiedziona), interakcja między płcią a wymiarem zatrudnienia. Oszacowania parametrów znajdują się poniżej. Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając *p-value*.

```
      Source |       SS       df       MS              Number of obs =    1967
-------------+------------------------------           F( 12,  1954) =  118.05
       Model |  211.507908    12   17.625659           Prob > F      =  0.0000
    Residual |  291.740412  1954  .149304203           R-squared     =  0.4203
-------------+------------------------------           Adj R-squared =  0.4167
       Total |   503.24832  1966  .255975748           Root MSE      =   .3864
------------------------------------------------------------------------------
    ln_placa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     wyksz_1 |   .2021485   .0303484     6.66   0.000     .1426299    .2616671
     wyksz_2 |   .2945919   .0362894     8.12   0.000     .2234219    .3657619
     wyksz_3 |   .4560558   .0378275    12.06   0.000     .3818694    .5302422
        wies |  -.0433189   .0180795    -2.40   0.017     -.078776   -.0078618
        wiek |   .0388162   .0056042     6.93   0.000     .0278253    .0498072
      wiek_2 |  -.0004376   .0000676    -6.47   0.000    -.0005702   -.0003049
        plec |  -.1072408   .0505662    -2.12   0.034    -.2064101   -.0080716
        etat |    .780341   .0444233    17.57   0.000      .693219    .8674629
   plecXetat |  -.1257776   .0534618    -2.35   0.019    -.2306257   -.0209296
      stan_1 |   .0951863   .0240566     3.96   0.000     .0480071    .1423656
      stan_2 |  -.1416434   .0605747    -2.34   0.019    -.2604411   -.0228456
      stan_3 |  -.0243112   .0445971    -0.55   0.586     -.111774    .0631517
       _cons |   5.470661    .106983    51.14   0.000     5.260849    5.680474
------------------------------------------------------------------------------
Breusch-Pagan LM statistic:       Chi2(1)   = 284.26  p-value = 0.000
White's general test statistic:   Chi2(70)  = 508.28  p-value = 0.000
Jarque-Berra test statistic:      Chi2(2)   = 3185.88 p-value = 0.000
Ramsey RESET test statistic:      F(3,1951) = 44.72   p-value = 0.000
```

1. Czy zmienne objaśniające są łącznie istotne?
2. Zinterpretować wartość współczynnika determinacji.
3. Ocenić, które zmienne są istotne.
4. Wykonać następujące polecenia:
   (a) Policzyć i zinterpretować semielastyczność cząstkową dla *wieku*. Średni wiek respondentów w próbie wynosi 42 lata.
   (b) Zintpretować wielkość współczynnika przy interakcji pomiędzy *płcią a etatem*.
   (c) Zintpretować wielkość współczynnika przy zmiennej *etat*.
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 są 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 =118.05, $p-value = 0.000 < 0.05$ odrzucamy hipotezę zerową o łącznej nieistotności regresji.
2. 42.03% zmienności zmiennej zależnej zostało wyjaśnione za pomocą zmienności 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) *wyksz_1* (t = 6.66, $p-value = 0.000$)
   (b) *wyksz_2* (t = 8.12, $p-value = 0.000$)
   (c) *wyksz_3* (t = 12.06, $p-value = 0.000$)
   (d) *wies* (t = -2.4, $p-value = 0.017$)
   (e) *wiek* (t = 6.93, $p-value = 0.000$)
   (f) *wiek_2* (t = -6.47, $p-value = 0.000$)
   (g) *plec* (t = -2.12, $p-value = 0.034$)
   (h) *etat* (t = 17.57, $p-value = 0.000$)
   (i) *plecXetat* (t = -2.35, $p-value= 0.019$)
   (j) *stan_1* (t = 3.96, $p-value = 0.000$)
   (k) *stan_2* (t = -2.34, $p-value = 0.019$)
   (l) const (t = 51.14, $p-value = 0.000$)
4. Wykonać następujące polecenia:
   (a) semielastyczność cząstkowa dla wieku:

   $\frac{\partial E(ln\_placa)}{\partial wiek} = \beta_{wiek} + 2 \cdot \beta_{wiek\_2} \cdot \overline{wiek} = 0,002058$

   wzrost wieku osób 42 letnich o 1 rok powoduje wzrost wynagrodzenia średnio o 0.2% *ceteris paribus*.

   (b) kobiety pracujące na cały etat mają średnio o 92.3% $(100\% \cdot [exp(0,78) \cdot exp(-0,126) - 1]) = 100\% \cdot [exp(0,654) - 1]$ wyższe wynagrodzenie niż kobiety pracujące na niepełny etat *ceteris paribus*.
   (c) mężczyźni pracujący na cały etat mają średnio o 118% $(100\% \cdot [exp(0,78) - 1])$ wyższe wynagrodzenie niż mężczyźni pracujący na niepełny etat *ceteris paribus*.
5. Występowanie heteroskedastyczności testujemy za pomocą:
   (a) testu White'a:
      i. hipoteza zerowa: homoskedastyczność składnika losowego.

ii. wartość statystyki testowej wynosi: chi2(70)=508.28 oraz $p - value = 0.000 < 0.05$, więc odrzucamy hipotezę zerową o homoskedastyczności.

(b) testu Breuscha-Pagana:

i. hipoteza zerowa: homoskedastyczność składnika losowego.

ii. wartość statystyki testowej wynosi chi2(1) = 284.26 oraz $p - value = 0.000 < 0.05$, 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) = 3185.88 oraz $p - value = 0.000 <0.05$, czyli odrzucamy hipotezę zerową o normalności zaburzenia losowego.

7. Poprawność przyjętej formy funkcyjnej modelu testujemy za pomocą:

(a) test RESET:

i. hipoteza zerowa: przyjęta postać funkcyjna modelu jest prawidłowa.

ii. wartość statystyki testowej F(3, 1951) = 44.72 i $p-value = 0.000 < 0.05$, więc odrzucamy hipotezę zerową o poprawności przyjętej formy funkcyjnej.

8. Odpowiedzi są następujące:

(a) Nie są spełnione następujące założenie:

i. o homoskedastyczności zaburzenia losowego,

ii. o sposobie „generowania danych": $y = \beta x + \varepsilon$ (czyli założenie o liniowej zaleności między zmienną zależną i zmiennymi niezależnymi),

iii. 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 1967 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

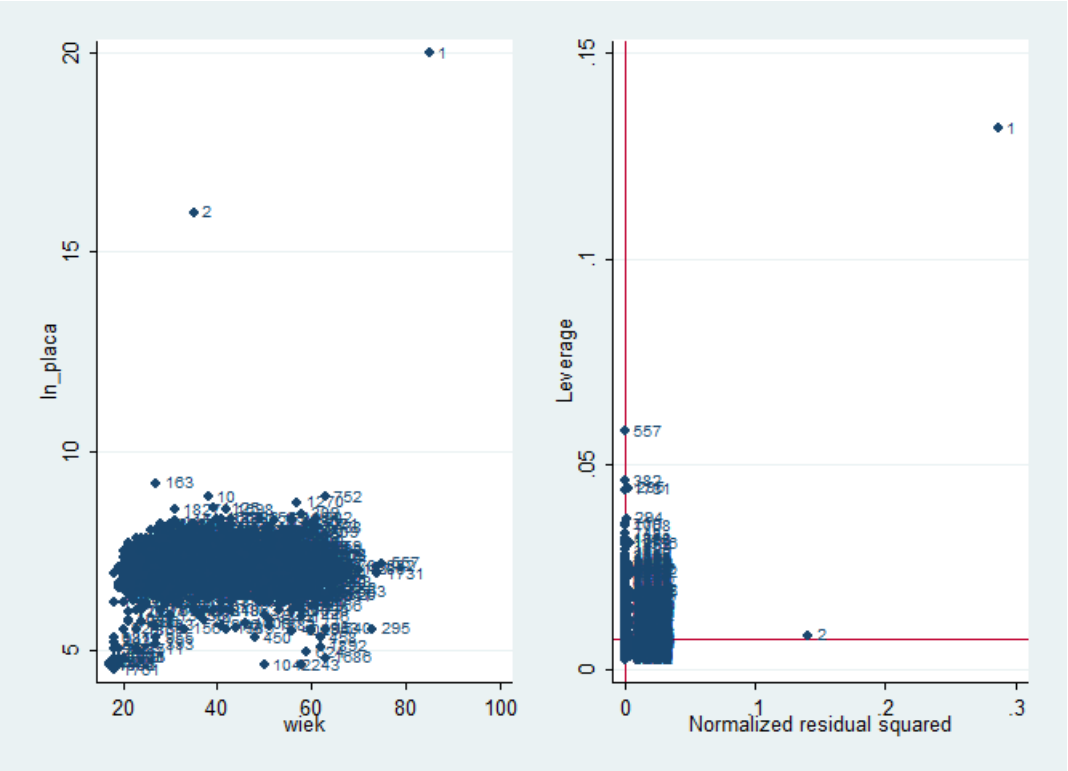
Na podstawie danych BAEL z 2010 roku oszacowano wysokość wynagrodzenia dla pracowników w mikroprzedsiębiorstwach (*ln_placa* - logarytm wynagrodzenia). Zmiennymi objaśniającymi są wiek (*wiek*), wiek do kwadratu (*wiek_2*), miejsce zamieszkania (*wies*: 0 - miasto, 1 - wieś), płeć (*plec*: 0 - mężczyzna, 1 - kobieta), wykształcenie (*wyksz*: 0 - podstawowe, 1 - zawodowe, 2 - średnie, 3 - wyższe), wymiar zatrudnienia (*etat*: 0 - niepełny etat, 1 - cały etat), stan cywilny (*stan*: 0 - kawaler / panna, 1 - zamężny / zamężna, 2 - wdowiec / wdowa, 3 - rozwiedziony / rozwiedziona), interakcja między płcią a wymiarem zatrudnienia. Oszacowania parametrów znajdują się poniżej. Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając *p-value*.

1. Przed oszacowaniem regresji uzyskano statystyki opisowe dla zmiennej wysokość wynagrodzenia dla pracowników w mikroprzedsiębiorstwach (*placa*). Wyjaśnić, które obserwacje budzą podejrzenia i dlaczego?

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       placa |      4534    57188.11    48913.66         90      99999

count if placa==99999
2567
```

2. Po oszacowanej regresji badacz uzyskał wykres rozrzutu pomiędzy wiekiem a logarytmem płacy oraz wykres przedstawiający zależność pomiędzy dźwignią a wystandaryzowanymi resztami podniesionymi do kwadratu. Wyjaśnić, które obserwacje budzą podejrzenia i dlaczego?



3. W modelu uwzględniono dodatkowo dwie zmienne: staż w obecnym miejscu pracy (*staz*), ogólny staż pracy (*staz_ ogolny*). Następnie obliczono współczynniki korelacji *Spearmana*. Jaki problem występuje w tym modelu?

```
      Source |       SS       df       MS              Number of obs =    1967
-------------+------------------------------           F( 14,  1952) =  105.67
       Model |  216.967865    14  15.4977047           Prob > F      =  0.0000
    Residual |  286.280455  1952  .146660069           R-squared     =  0.4311
-------------+------------------------------           Adj R-squared =  0.4271
       Total |   503.24832  1966  .255975748           Root MSE      =  .38296
------------------------------------------------------------------------------
    ln_placa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     wyksz_1 |   .1889953   .0302267     6.25   0.000     .1297153    .2482753
     wyksz_2 |   .2786084   .0360692     7.72   0.000     .2078703    .3493465
     wyksz_3 |   .4458425   .0375599    11.87   0.000     .3721808    .5195043
        wies |  -.0512427   .0179803    -2.85   0.004    -.0865054     -.01598
        wiek |   .0358145   .0056632     6.32   0.000     .0247081     .046921
      wiek_2 |  -.0004565   .0000674    -6.77   0.000    -.0005887   -.0003243
        plec |  -.1157975   .0501914    -2.31   0.021    -.2142319   -.0173632
        etat |   .7589422   .0441721    17.18   0.000     .6723127    .8455717
   plecXetat |  -.1165905   .0530097    -2.20   0.028     -.220552    -.012629
      stan_1 |   .0928414   .0238605     3.89   0.000     .0460467    .1396362
      stan_2 |  -.1109334   .0602626    -1.84   0.066    -.2291193    .0072524
      stan_3 |  -.0083049   .0442809    -0.19   0.851    -.0951477    .0785379
        staz |   .0076781   .0014517     5.29   0.000     .0048311    .0105252
 staz_ogolny |   .0026003    .001987     1.31   0.191    -.0012967    .0064972
       _cons |   5.566119   .1095404    50.81   0.000     5.351291    5.780947
------------------------------------------------------------------------------
Macierz korelacji Spearmana:
             |     wiek   staz staz_ogolny  wyksz
-------------+------------------------------------
        wiek |   1.0000
        staz |   0.4875   1.0000
 staz_ogolny |   0.9201   0.5743   1.0000
       wyksz |  -0.1575   0.0251  -0.1777   1.0000
```

4. Przeprowadzono regresję wysokości wynagrodzenia dla pracowników w mikroprzedsiębiorstwach (*ln_placa* - logarytm wynagrodzenia) na latach poświęconych na naukę (*nauka*).

```
      Source |       SS       df       MS              Number of obs =    1968
-------------+------------------------------           F(  1,  1966) =   36.69
       Model |  13.7328051     1  13.7328051            Prob > F      =  0.0000
    Residual |  735.877367  1966  .374301814            R-squared     =  0.0183
-------------+------------------------------           Adj R-squared =  0.0178
       Total |  749.610172  1967  .381093122            Root MSE      =   .6118

------------------------------------------------------------------------------
    ln_placa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nauki |   .1009973   .0166741     6.06   0.000     .0682967     .133698
       _cons |   6.948158   .0260974   266.24   0.000     6.896976    6.999339
```

(a) Jaki będzie prawdopodobny kierunek obciążenia oszacowania parametru przy zmiennej *nauka* w wyniku pominięcia zmiennej zwiazanej z wielkością miejscowości.

Rozwiązanie Zadanie 3

1. Największe podejrzenia budzą obserwacje, dla których *placa*=99999. Zwraca uwagę fakt, iż dla 2567 obserwacji spośród 4534 wysokość wynagrodzenia jest zaskakująco wysoka i dokładnie taka sama. Można podejrzewać, że nie są to rzeczywiste obserwacje dla wynagrodzenia, ale kody braków (np. odpowiedź: nie wiem). Jeśli rzeczywiście obserwacje, dla których placa=99999 są błędne, to powinniśmy je usunąć z modelu.

2. Na podstawie wykresu rozrzutu pomiędzy wiekiem a logarytmem wynagrodzenia podejrzenia budzą obserwacje **nr 1** i **nr 2**. Obserwacja **nr 1** jest nietypowa na tle pozostałych obserwacji ze względu na wysoką wartość zmiennej *wiek* i wysoką wartość *logarytmu wynagrodzenia*. Obserwacja **nr 2** jest nietypowa na tle pozostałych obserwacji ze względu na wysoką wartość *logarytmu wynagrodzenia*. Na podstawie wykresu przedstawiającego zależność pomiędzy *dźwignią* a *wystandaryzowanymi resztami podniesionymi do kwadratu* możemy potwierdzić, iż obserwacja **nr 1** charakteryzuje się zarówno wysoką wartością *dźwigni* jak i wysokimi wartościami *wystandaryzowanych reszt podniesionymi do kwadratu*. Znaczący wpływ na oszacowania ma dodanie obserwacji o **nr 1**, która jest nietypowa ze względu na zmienną objaśniającą oraz nie pasuje do linii regresji (otrzymujemy duże reszty). Jest to obserwacja nietypowa, która nie pasuje do prostej regresji. Natomiast obserwacja **nr 2** cechuje się wysoką wartością *wystandaryzowanych reszt podniesionymi do kwadratu*. Dodanie obserwacji **nr 2** do próby nie powoduje znacznych zmian w oszacowanych parametrach (jest to obserwacja o stosunkowo dużej reszcie, ale nie jest nietypowa ze względu na zmienną objaśniającą).

3. Na podstawie macierzy korelacji Spearmana można zauważyć silną zależność pomiędzy zmienną *wiek* a zmienną *staz_ogolny*. Jednocześnie zmienna *staz_ogolny* jest nieistotna w modelu (*p-value*=0.191). W związku z powyższym w modelu wystąpi problem **niedokładnej współliniowości**.

4. W dużych miastach wynagrodzenia są średnio wyższe niż w małych miejscowościach. Jednocześnie w dużych miastach jest też więcej osób dobrze wykształconych. Pominięcie tej zmiennej doprowadzi więc do dodatniego obciążenia oszacowania parametru.