Pytania teoretyczne

1. Podać wzajemne relacje między wartościami obserwowanymi zmiennej zależnej, oszacowaniami parametrów, wartościami dopasowanymi i resztami

   wektor wartości dopasowana: $\widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$, gdzie $\boldsymbol{b}$ jest oszacowaniem wektora parametrów a $\boldsymbol{X}$ macierzą obserwacji dla zmiennych zależnych

   wektor reszt: $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, gdzie $\boldsymbol{y}$ jest wektorom obserwacji dla zmiennych niezależnych

2. Dlaczego zmienną dyskretną rozkodowujemy na zmienne zerojedynkowe?

   Zmienną dyskretną rozkodowujemy w dwóch sytuacjach:

   (a) Poziomy zmiennej dyskretnej kodują charakterystyki, których nie da się logicznie uporządkować. W tym przypadku założenie liniowości wpływu zmiennej dyskretnej nie ma sensu.

   (b) Poziomy zmiennej dyskretnej kodują charakterystyki, których da się logicznie uporządkować. Rozkodowujemy zmienną dyskretną jeśli nie da się założyć, że wpływ poszczególnych poziomów zmiennej jest wprost proporcjonalny do przypisanych im numerów porządkowych.

3. Wymienić założenia Klasycznego Modelu Regresji Liniowej ($KMRL$)

   (a) Model jest liniowy

   $$y_i = x_{1i}\beta_1 + \ldots + x_{Ki}\beta_K + \varepsilon_i \quad \text{dla} \quad i = 1, \ldots, N$$

   lub

   $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

   (b) Zmienne objaśniające $x_{1i}, \ldots, x_{Ki}$ są nielosowe dla $i = 1, \ldots, N$

   (c) Wartość oczekiwana błędu losowego jest równa zeru

   $$\mathrm{E}(\varepsilon_i) = 0 \quad \text{dla} \quad i = 1, \ldots, N$$

   lub

   $$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$$

   (d) Błędy losowe są sferyczne

   $$\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$$

   lub

   i. Kowariancja między dwoma różnymi błędami losowymi jest równa zeru (brak autokorelacji)

   $$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{dla} \quad i \neq j$$

   ii. Wariancja błędu losowego jest taka sama dla wszystkich obserwacji (homoskedastyczność)

   $$\mathrm{Var}(\varepsilon_i) = \sigma^2 \quad \text{dla} \quad i = 1, \ldots, N$$

4. Jakie są konsekwencje niedokładnej współliniowości? Za pomocą jakiej statystyki można wykryć niedokładną współliniowość w modelu?

   Niedokładna współliniowość polega na korelacji między zmiennymi objaśniającymi. Niedokładna współliniowość prowadzi do wzrostu wariancji i błędów standardowych oszacowań parametrów przy skorelowanych zmiennych. W konsekwencji prowadzi do spadku statystyk t przy tych zmiennych i może spowodować, że zmienne te staną się nieistotne w modelu. Występowanie niedokładnej współliniowości w modelu można wykryć za pomocą statystyki

   $$VIF_k = \frac{1}{1 - R_k^2}$$

   gdzie R_k^2 jest współczynnikiem determinacji w regresji k-tej zmiennej objaśniającej na pozostałych zmiennych objaśniających. O silnej niedokładnej współliniowości mówi się, gdy $VIF_k > 10$.

ZADANIE 1 Rozwiązać poniższe problemy:

1. Oszacowano model: $y_i = \beta_1 + \beta_2 x_{1i} + \epsilon_i$. Obliczyć nieznane wartości reszt (e_3, e_4), jeżeli wiadomo, że $\boldsymbol{x}_1' = [2, 3, 3, 2]$, a $\boldsymbol{e'} = [2, -1, e_3, e_4]$.
2. Oszacowano model $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$ dla $\boldsymbol{y}' = [1, 3, 5, 3, 2, 4]$ i uzyskano wartości dopasowane $\widehat{\boldsymbol{y}}' = [3, 2, 6, 2, 1, \widehat{y}_6]$. Znaleźć $\widehat{y}_6$.
3. Pokazać, że w modelu $y_i = \beta + \varepsilon_i$ suma kwadratów reszt $\boldsymbol{e}'\,\boldsymbol{e} = \sum_{i=1}^{N} (y_i - \overline{y})^2$

Rozwiązanie:

1. Korzystamy z tego, że w modelu ze stałą $\sum e_i = 0$ i $\boldsymbol{X}'\,\boldsymbol{e} = \boldsymbol{0} \Longrightarrow \sum e_i x_{1i} = 0$. Wektor reszt będzie postaci: $\boldsymbol{e'} = [2, -1, 1, -2]$.
2. Do rozwiązania zadania należy wykorzystać następującą właściwość hiperpłaszczyzny regresji - $\sum y_i = \sum \hat{y}_i$. Zatem $\hat{y}_6 = 4$.
3. W modelu, w którym występuje tylko stała $\boldsymbol{X} = \boldsymbol{l}$ a $\boldsymbol{b} = \left(\boldsymbol{l}'\boldsymbol{l}\right)^{-1}\boldsymbol{l}'\boldsymbol{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}$. Z kolei

$$e_i = y_i - x_i\,\boldsymbol{b} = y_i - \overline{y}$$

$$\boldsymbol{e}'\boldsymbol{e} = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

ZADANIE 2 Przeprowadzono dwie regresje wyjaśniające logarytm zarobków osiąganych (logrincome) przez zamężne kobiety za pomocą ilości lat poświęconych przez nie nauce (educ).

```
      Source |       SS       df       MS              Number of obs =     371
-------------+------------------------------           F(  1,   369) =   45.48
       Model |  10.6867044     1  10.6867044           Prob > F      =  0.0000
    Residual |  86.7135186   369  .234995985           R-squared     =  0.1097
-------------+------------------------------           Adj R-squared =  0.1073
       Total |   97.400223   370  .263243846           Root MSE      =  .48476

------------------------------------------------------------------------------
  logrincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0650681   .0096489     6.74   0.000     .0460944    .0840417
       _cons |   5.549348   .1175245    47.22   0.000     5.318247     5.78045
------------------------------------------------------------------------------
```

1. Podać interpretację oszacowań parametrów w tym modelu oraz interpretację statystyki R^2.
2. Jaki będzie prawdopodobny kierunek obciążenia oszacowania parametru przy zmiennej educ wynikły z pominięcia:
   (a) inteligencji respondentki.
   (b) liczby dzieci respondentki.
   (c) wielkości miejscowości, w której zamieszkuje respondentka
   (d) zarobków męża respondentki.

Rozwiązanie:

1. Zmienna zależna jest zlogarytmowana, zmienna niezależna nie została zlogarytmowana. Parametr przy liczbie lat poświęconych nauce jest semielastycznością, płaca kobiet wzrasta średnio o $6,5\%$ dla każdego dodatkowego roku nauki. Stałej nie interpretujemy. R^2 wskazuje, że $10,7\%$ zmienności dochodów można wyjaśnić za pomocą zmiennej educ.
2.

(a) Inteligencja jest dodatnio skorelowana z wykształceniem (ułatwia zdobycie wyższego wykształcenia) i dodatnio wpływa na zarobki. Pominięcie tej zmiennej doprowadzi więc do dodatniego obciążenia oszacowania.

(b) Kobiety mające więcej dzieci mniej czasu mogą poświęcić pracy i w związku z tym mniej zarabiają. Z drugiej strony kobiety bardziej wykształcone mają zazwyczaj mniej dzieci. Pominięcie tej zmiennej będzie więc najprawdopodobniej prowadzić do dodatniego obciążenia oszacowania.

(c) W dużych miastach zarobki są średnio wyższe niż w małych miejscowościach. W dużych miastach jest też więcej osób dobrze wykształconych. Pominięcie tej zmiennej doprowadzi więc do dodatniego obciążenia oszacowania.

(d) Wysokie zarobki męża najprawdopodobniej wpływają ujemnie na czas poświęcany pracy i tym samym zarobki uzyskiwane z pracy przez żonę. Z drugiej strony wysokie dochody osiągają dobrze wykształceni mężczyźni a tacy mają też naogół wykształcone żony. Pominięcie tego czynnika może więc doprowadzić do ujemnego obciążenia oszacowania.

ZADANIE 3 Na kwartalnych danych makroekonomicznych z lat 1997q1-2004q1 dotyczących Polski przeprowadzono regresję mającą na celu wyjaśnienie dynamikę inwestycji. Z przeprowadzonej regresji uzyskano następujące wyniki:

```
      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  4,    29) =   24.22
       Model |  4259.76852     4  1064.94213            Prob > F      =  0.0000
    Residual |  1274.89154    29  43.9617772            R-squared     =  0.7697
-------------+------------------------------           Adj R-squared =  0.7379
       Total |  5534.66006    33  167.716971            Root MSE      =  6.6304

------------------------------------------------------------------------------
      akumul |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stopa |  -.9236267   .5332449    -1.73   0.094    -2.014235    .1669815
         inf |   1.287234   .3542712     3.63   0.001     .5626679     2.0118
        bael |  -2.088973   .5473591    -3.82   0.001    -3.208448   -.9694982
      budz_d |  -.3574961   .5808486    -0.62   0.543    -1.545465    .8304726
       _cons |   41.48327   12.62874     3.28   0.003     15.65459    67.31195
------------------------------------------------------------------------------
```

gdzie $akumul$ jest stopą wzrostu inwestycji, $stopa$ jest wysokością stopy procentowej, inf wielkością inflacji, $bael$ stopą bezrobocia, $budz\_d$ wielkością deficytu budżetowego w relacji do PKB. Stopy wzrostu podawane są w odniesieniu do analogicznego kwartału poprzedniego roku. Wszystkie zmienne wyrażone zostały w procentach. Hipotezy testujemy na poziomie istotności $\alpha = 5\%$. Wartości krytyczne testu DW wynoszą $d_L = 1.27$ i $d_U = 1.65$. Odpowiedzi należy uzasadniać odpowiednimi statystykami.

1. Czy model dobrze objaśnia zmienną objaśnianą? Czy wszystkie zmienne w modelu są łącznie istotne?
2. Jaka jest interpretacja poszczególnych współczynników w modelu?
3. Które zmienne w modelu są istotne?
4. Z teorii ekonomii wynika, że na wielkość inwestycji powinna wpływać realna stopa procentowa. Jak można sformułować tę hipotezę w kategoriach ograniczeń nałożonych na parametry powyższego modelu?
   W omawianym powyżej modelu stworzono zmienną $rstopa = stopa - infl$ i po przeprowadzeniu regresji uzyskano następujące wyniki:

```
      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  3,    30) =   31.09
       Model |  4187.61937     3  1395.87312            Prob > F      =  0.0000
    Residual |  1347.04068    30  44.9013561            R-squared     =  0.7566
```

```
-------------+------------------------------           Adj R-squared =  0.7323
       Total |  5534.66006    33  167.716971           Root MSE      =  6.7008

------------------------------------------------------------------------------
      akumul |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rstopa |  -1.463905   .3297961    -4.44   0.000    -2.137439   -.7903721
        bael |  -2.554936   .4133782    -6.18   0.000    -3.399167   -1.710705
      budz_d |  -.4212491   .5848644    -0.72   0.477    -1.615702    .7732033
       _cons |   56.05575   5.543945    10.11   0.000      44.7335    67.37799
------------------------------------------------------------------------------
Breusch-Godfrey LM test F(  2,   28 ) = 2.825 [0.0763]
Durbin-Watson d-statistic(  4,    34) =  1.182685
Breusch-Pagan  chi2(1)      =     4.40 [0.0360]
Jarque-Bera normality test Chi(2) = 0.59 [0.6240]
Ramsey RESET test F(3, 26) =      3.71 [0.0241]
```

5. Zweryfikuj hipotezę z punktu 4 ($F_{\alpha=0.05}(1,30) = 4.17$).
6. Czy w modelu występuje autokorelacja (do testowania użyj wszystkich dostępnych statystyk)?
7. Czy forma funkcyjna modelu jest prawidłowa?
8. Czy błąd losowy w modelu ma rozkład normalny?
9. Czy w modelu występuje heteroskedastyczność (do testowania użyj wszystkich dostępnych statystyk)?
10. Jakie założenia $KMRL$ nie są spełnione w analizowanym modelu i jakie ma to konsekwencje dla wnioskowania statystycznego?
11. W jaki sposób można poradzić sobie ze zdiagnozowanymi problemami?

Rozwiązanie:

1. Zmienne egzogeniczne objaśniają 77% zmienności zmiennych endogenicznych. Odrzucamy hipotezę o łącznej nieistotności zmiennych ($F = 24.22$, $p - value = 0.0000 < 0.05$)
2. Wzrost stopy procentowej o 1% powoduje spadek akumulacji o 0.92%, wzrost inflacji o 1% powoduje wzrost akumulacji o 1.29%, wzrost stopy bezrobocia o 1% powoduje spadek akumulacji o 2.09%, wzrost relacji deficytu budżetowego do PKB o 1% powoduje spadek akumulacji o -0.36%. Stała nie ma interpretacji
3. W modelu istotne są zmienne inf ($t = 3.63$, $p - value = 0.001 < 0.05$), bael ($t = -3.82$, $p - value = 0.001 < 0.05$), stała ($t = 3.28$,$p - value = 0.003 < 0.05$)
4. Model ma postać

$$akumul_t = \beta_0 + \beta_1 stopa_t + \beta_2 inf_t + \beta_3 bael_t + \beta_4 budz\_d_t + \varepsilon_t$$

   Realna stopa procentowa jest równa (w przybliżeniu) różnicy między nominalną stopą procentową a inflacją. Model, w którym na akumulację wpływa jedynie realna stopa procentowa ma więc następującą postać:

$$akumul_t = \beta_0 + \beta_1 \left(stopa_t - \text{inf}_t\right) + \beta_3 bael_t + \beta_4 budz\_d_t + \varepsilon_t$$

   Model ten uzyskujemy z modelu pierwotnego jeśli prawdziwa jest hipoteza $H_0 : \beta_2 = -\beta_1$
5. Do zweryfikowania hipotezy, że $H_0 : \beta_2 = -\beta_1$ możemy wykorzystujemy wielkości sumy kwadratów reszt w modelu bez ograniczeń ($S = 1275$) i w modelu z ograniczeniami ($S_R = 1347$). Testujemy jedno ograniczenie $g = 1$, liczba obserwacji wynosi 34 a liczba szacowanych w modelu bez ograniczeń parametrów $k = 5$ i wzór

$$F = \frac{\frac{S_R - S}{g}}{\frac{S}{n-k}} = \frac{1275}{30} \approx 1.70 < 4.17$$

6. Test Breuscha-Godfreya ($F = 2.825$, $p-value = 0.0763 > 0.05$) nie daje podstaw do odrzucenia hipotezy zerowej o braku autokorelacji, z drugiej strony test DW ($1.18 < d_L = 1.27$) odrzuca hipotezę zerową o braku autokorelacji pierwszego rzędu

7. Test RESET ($F = 3.71$, $p - value = 0.0241 < 0.05$) odrzuca hipotezę zerową o prawidłowości formy funkcyjnej

8. Test Jarque-Berra ($F = 0.59$, $p-value = 0.6240 > 0.05$) nie daje podstaw do odrzucenia hipotezy zerowej o normalności rozkładu błędu losowego

9. Test Breuscha-Pagana ($F = 4.40$, $p - value = 0.0360 < 0.05$) skłania nas do odrzucenia hipotezy zerowej o braku heteroskedastyczności

10. Wielkość statystyk testowych skłania nas do wniosku, że w analizowanym modelu nie jest spełnione założenie $KMRL$, że $\mathrm{Var}(\varepsilon) = \sigma^2 \boldsymbol{I}$, czyli błędy losowe wykazują autokorelację i homoskedastyczność. Dodatkowo test RESET sugeruje, że błędne jest założenie, że $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, czyli że zależność między zmienną objaśnianą a zmiennymi objaśniającymi ma charakter liniowy. Autokorelacja i heteroskedastyczność błędu losowego nie powodują obciążenia estymatora parametrów ale mogą spowodować błędne wyniki testów statystycznych. Nieliniowość analizowanej zależności między zmiennymi może spowodować obciążenie i brak zbieżności oszacowań parametrów.

11. Problem nieliniowości zależności między zmiennymi można rozwiązać analizując inne potencjalnie poprawne formy modelu (np. model na logarytmach). Problemy wynikłe z autokorelacji i heteroskedastyczności można rozwiązać stosując odporną macierz wariancji kowariancji bądź Uogólnioną Metodę Najmniejszych Kwadratów.