

Egzamin z ekonometrii – wersja ogólna
02-02-2015

Pytania teoretyczne

1. Udowodnić, że w modelu ze stałą $TSS=ESS+RSS$.
2. Dlaczego zmienną dyskretną rozkodowujemy na zmienne zerojedynkowe?
3. Kiedy mówimy, że zmienne są dokładnie współliniowe? Jak można rozwiązać ten problem?
4. Udowodnić, że w KMRL estymator \mathbf{b} jest nieobciążony.

Zadanie 1

Na podstawie pewnej bazy danych oszacowano wysokość rocznych dochodów w złotych (ln_dochod - logarytm rocznych dochodów). Zmiennymi niezależnymi są płeć (plec: 0 – mężczyzna, 1 – kobieta); następujące zmienne opisujące stan cywilny: stan wolny (wolny: 1 – jeśli osoba jest stanu wolnego, 0 – w pozostałych przypadkach), małżeństwo (malzenstwo: 1 – jeśli osoba jest w związku małżeńskim, 0 w pozostałych przypadkach), rozwód (rozwod: 1 – jeśli osoba jest rozwiedziona, wdowcem lub wdową, 0 w pozostałych przypadkach); interakcja między zmienną plec a malzenstwo oraz interakcja między zmienną plec a rozwod; wiek w latach (wiek).

Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając p-value.

Source	SS	df	MS	Number of obs = 1595		
Model	410.184469	6	68.3640781	F(6, 1588)	= 108.12	
Residual	1004.09783	1588	.632303418	Prob > F	= 0.0000	
Total	1414.2823	1594	.887253637	R-squared	= 0.2900	
				Adj R-squared	= 0.2873	
				Root MSE	= .79518	

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
plec	-.2373567	.0880731	-2.69	0.007	-.4101084	-.0646049
malzenstwo	.5494907	.0757085	7.26	0.000	.4009916	.6979898
rozwod	.4534236	.1228142	3.69	0.000	.2125286	.6943186
plecXmalz	-.7945073	.1005232	-7.90	0.000	-.9916794	-.5973352
plecXrozv	-.5079947	.1553932	-3.27	0.001	-.8127921	-.2031973
wiek	.014751	.0022659	6.51	0.000	.0103066	.0191954
_cons	9.4797	.089668	105.72	0.000	9.30382	9.65558

1. Wyjaśnić dlaczego wprowadzono do modelu interakcje między zmienną plec a zmiennymi opisującymi stan cywilny.
2. Ocenić, które zmienne są istotne.
3. Zinterpretować wielkość współczynnika przy interakcji pomiędzy płcią a małżeństwem.

4. Na podstawie powyższego modelu oraz Modeli 1-4 (patrz poniżej) przetestować następujące hipotezy (zapisać hipotezę zerową, podać wartość statystki testowej, podać wartość krytyczną, zweryfikować hipotezę):

- a) wpływ płci i stanu cywilnego na dochód jest addytywny;
- b) kobiety i mężczyźni w związku małżeńskim zarabiają tyle samo;
- c) premia za małżeństwo (przejście ze stanu wolnego do związku małżeńskiego) dla kobiet i mężczyzn jest taka sama.

Do testowania hipotez mogą przydać się następujące kwantyle:

$$F_{0.95}(1,1588) = 3,8473176; F_{0.95}(2,1588) = 3,0013872; F_{0.95}(3,1588) = 2,6105024$$

Model 1

x – zmienna przyjmująca wartość 1 dla kobiet nie będących w związku małżeńskim oraz wartość 0 w pozostałych przypadkach.

Source	SS	df	MS	Number of obs = 1595		
Model	129.487721	5	25.8975443	F(5, 1589)	=	32.03
Residual	1284.79458	1589	.808555428	Prob > F	=	0.0000
Total	1414.2823	1594	.887253637	R-squared	=	0.0916
				Adj R-squared	=	0.0887
				Root MSE	=	.8992

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
malzenstwo	-.0005994	.0803607	-0.01	0.994	-.1582235	.1570247
rozwod	.352757	.1387752	2.54	0.011	.0805553	.6249587
x	.2265102	.0995928	2.27	0.023	.0311631	.4218573
plecXrozv	-.5295516	.1757174	-3.01	0.003	-.8742138	-.1848893
wiek	.0202269	.0025454	7.95	0.000	.0152343	.0252195
_cons	9.30783	.1009776	92.18	0.000	9.109767	9.505894

Model 2

z – zmienna przyjmująca wartość 0 dla mężczyzn, wartość 1 dla kobiet nie będących w związku małżeńskim, wartość 2 dla kobiet w związku małżeńskim.

Source	SS	df	MS	Number of obs = 1595		
Model	404.294684	5	80.8589367	F(5, 1589)	= 127.21	
Residual	1009.98761	1589	.635612091	Prob > F	= 0.0000	
				R-squared	= 0.2859	
				Adj R-squared	= 0.2836	
				Root MSE	= .79725	
Total	1414.2823	1594	.887253637			

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
malzenstwo	.4345768	.0658536	6.60	0.000	.3054077	.5637458
rozwod	.356034	.1189062	2.99	0.003	.1228045	.5892634
z	-.496304	.0236892	-20.95	0.000	-.5427693	.4498386
plecXrozw	-.2487986	.130473	-1.91	0.057	-.5047159	.0071188
wiek	.0146238	.0022714	6.44	0.000	.0101685	.019079
_cons	9.583423	.0831949	115.19	0.000	9.42024	9.746606

Model 3

Source	SS	df	MS	Number of obs = 1595		
Model	370.341851	4	92.5854626	F(4, 1590)	= 141.01	
Residual	1043.94045	1590	.656566318	Prob > F	= 0.0000	
				R-squared	= 0.2619	
				Adj R-squared	= 0.2600	
				Root MSE	= .81029	
Total	1414.2823	1594	.887253637			

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
plec	-.8350898	.0413974	-20.17	0.000	-.916289	-.7538905
malzenstwo	.2262706	.0650395	3.48	0.001	.0986984	.3538427
rozwod	.2733851	.0908261	3.01	0.003	.0952336	.4515366
wiek	.0150732	.0023027	6.55	0.000	.0105565	.0195898
_cons	9.700266	.0858609	112.98	0.000	9.531854	9.868679

Model 4

Source	SS	df	MS			
Model	370.685209	5	74.1370419	Number of obs = 1595		
Residual	1043.59709	1589	.656763428	F(5, 1589) = 112.88		
Total	1414.2823	1594	.887253637	Prob > F = 0.0000		
				R-squared = 0.2621		
				Adj R-squared = 0.2598		
				Root MSE = .81041		

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
plec	-.845265	.0437298	-19.33	0.000	-.9310392	-.7594907
malzenstwo	.2290106	.0651595	3.51	0.000	.101203	.3568183
rozwod	.2152187	.1213398	1.77	0.076	-.0227842	.4532217
plecXrozw	.0995201	.1376389	0.72	0.470	-.1704528	.369493
wiek	.0149522	.0023091	6.48	0.000	.0104229	.0194814
_cons	9.707893	.0865192	112.21	0.000	9.538189	9.877597

Rozwiązanie Zadanie 1

1. Do modelu wprowadzono interakcję między zmienną *plec* a zmiennymi opisującymi stan cywilny aby uchwycić fakt, iż wpływ stanu cywilnego na wysokość rocznych dochodów może być różny dla kobiet i mężczyzn.
2. Istotne zmienne, to te dla których p – value jest mniejsze od przyjętego poziomu istotności wynoszącego 0.05. Czyli istotne zmienne to:
 - i. *plec* (t=-2.69, p-value=0.007)
 - ii. *malzenstwo* (t=7.26, p-value=0.000)
 - iii. *rozwod* (t=3.69, p-value=0.000)
 - iv. *plecXmalz* (t=-7.90, p-value=0.000)
 - v. *plecXrozw*(t=-3.27, p-value=0.001)
 - vi. *wiek* (t=6.51, p-value=0.000)
 - vii. *cons* (t=105.72, p-value=0.000)
3. Kobiety zameżne mają średnio o 21,34% ($100\% * [\exp(0.55) * \exp(-0.79) - 1]$) = $100\% * [\exp(-0.24) - 1]$ niższe wynagrodzenie niż kobiety stanu wolnego.

4.

a) Hipoteza zerowa: $\beta_{plecXmalz} = \beta_{plecXrozw} = 0$.

W celu obliczenia wartości statystyki testowej korzystamy z modelu oszacowanego na początku zadania (model bez ograniczeń) i Modelu 3 (model z ograniczeniami).

Wartość statystyki testowej:

$$F = \frac{(S_R - S)/g}{S/(N - K)} = \sim F(g, N - K)$$

gdzie: S_R – suma kwadratów reszt modelu z ograniczeniami

S – suma kwadratów rest zmodel bez ograniczeń

g – liczba ograniczeń

Jeśli statystyka testowa jest większa od wartości krytycznej $F \sim (g, N - K)$, to odrzucamy hipotezę zerową.

$$F = \frac{(1043.94045 - 1004.09783)/2}{1004.09783/(1595 - 7)} = 31,51$$

Wartość krytyczna z której należy skorzystać do zweryfikowania hipotezy to $F_{0,95}(2,1588) = 3,0013872$. Ponieważ wartość statystyki testowej jest większa

od wartości krytycznej to należy odrzucić hipotezę zerową. Oznacza to, że wpływ płci i stanu cywilnego na dochód nie jest addytywny.

b) Hipoteza zerowa: $\beta_{plec} + \beta_{plec \times malz} = 0$

W celu obliczenia wartości statystyki testowej korzystamy z modelu oszacowanego na początku zadania (model bez ograniczeń) i Modelu 1 (model z ograniczeniami). Wartość statystyki testowej:

$$F = \frac{(1284.79458 - 1004.09783)/1}{1004.09783/(1595 - 7)} = 444$$

Wartość krytyczna z której należy skorzystać do zweryfikowania hipotezy to $F_{0,95}(1,1588) = 3,8473176$. Ponieważ wartość statystyki testowej jest większa od wartości krytycznej to należy odrzucić hipotezę zerową. Oznacza to, że kobiety i mężczyźni w związku małżeńskim nie zarabiają tyle samo.

c) Hipoteza zerowa: $\beta_{plc \times malz} = 0$

W celu obliczenia wartości statystyki testowej korzystamy z modelu oszacowanego na początku zadania (model bez ograniczeń) i Modelu 4 (model z ograniczeniami). Wartość statystyki testowej:

$$F = \frac{(1043.59709 - 1004.09783)/1}{1004.09783/(1595 - 7)} = 62,47$$

Wartość krytyczna z której należy skorzystać do zweryfikowania hipotezy to $F_{0,95}(1,1588) = 3,8473176$. Ponieważ wartość statystyki testowej jest większa od wartości krytycznej to należy odrzucić hipotezę zerową. Oznacza to, że premia za małżeństwo nie jest taka sama dla kobiet i mężczyzn.

Zadanie 2

Metodą Najmniejszych Kwadratów oszacowano następujące dwa modele:

$$\text{I)} \quad \hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$$

$$\text{II)} \quad \hat{y}_i^* = b_0 + b_1 x_i$$

gdzie: $\hat{y}_i^* = \hat{y}_i - x_i$

1. Udowodnij, że pomiędzy oszacowanymi parametrami zachodzą następujące zależności:

$$\text{a)} \quad \hat{\alpha}_0 = b_0$$

$$\text{b)} \quad \hat{\alpha}_1 = b_1 + 1$$

2. Metodą Najmniejszych Kwadratów szacowano model z jedną zmienną objaśniającą i stałą. Założono, że zmienne objaśniające są nielosowe.

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon \sim N(0, \sigma^2 I)$$

Wyznaczyć $\text{cov}(\hat{\alpha}, \hat{\beta})$.

Rozwiązanie Zadanie 2

1. Wiemy, że

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$$

$$\hat{y}_i^* = b_0 + b_1 x_i$$

gdzie: $\hat{y}_i^* = \hat{y}_i - x_i$

Podstawiając do równania ($\hat{y}_i^* = b_0 + b_1 x_i$) $\hat{y}_i^* = \hat{y}_i - x_i$ otrzymamy:

$$\hat{y}_i - x_i = b_0 + b_1 x_i$$

$$\hat{y}_i = b_0 + x_i(b_1 + 1) = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$$

gdzie:

$$\hat{\alpha}_0 = b_0$$

$$\hat{\alpha}_1 = b_1 + 1$$

2. Należy wyznaczyć macierz wariancji – kowariancji dla wektora oszacowanych parametrów w przypadku sferyczności zaburzenia losowego: $\sigma^2(X'X)^{-1}$

$$\text{Var} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \sigma^2(X'X)^{-1}$$

Estymator macierzy wariancji i kowariancji estymatora MNK:

$$\hat{\Sigma} = s^2(X'X)^{-1}$$

W modelu ze stałą i jedną zmienną objaśniającą:

$$X'X = \begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix}$$

Zakładając, że macierz X ma pełen rząd kolumnowy (wystarczy, że istnieją takie $i \neq j$, że $x_i \neq x_j$):

$$(X'X)^{-1} = \frac{1}{N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix}$$

Czyli

$$\hat{\Sigma} = \frac{s^2}{N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix}$$

oraz

$$\text{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{s^2 \sum_{i=1}^N x_i}{N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2}$$

Zadanie 3

Na podstawie pewnej bazy danych, zawierającej 30 tys. obserwacji, oszacowano wysokość miesięcznych wynagrodzeń w złotówkach (\ln_wyn - logarytm miesięcznych wynagrodzeń). Zmiennymi niezależnymi są wiek w latach ($wiek$); wiek do kwadratu ($wiek_2$); płeć ($plec$: 0 – mężczyzna, 1 – kobieta); miejsce zamieszkania ($miasto$: 0 – wieś, 1 – miejscowość do 5 tys. mieszkańców; 2 – miejscowość od 5 tys. do 10 tys. mieszkańców, 3 – miejscowość od 10 tys. do 100 tys. mieszkańców, 4 – miejscowość od 100 tys. do 500 tys. mieszkańców); wykształcenie ($wyksztalzenie$: 0 – podstawowe, 1 – średnie, 2 - wyższe); fakt posiadania własnej firmy ($wlasciciel$: 1 – jeśli respondent posiada własną firmę, 0 – w pozostałych przypadkach); fakt zajmowania przez respondenta stanowiska kierowniczego w firmie ($kieruje$: 0 – respondent nie zajmuje stanowiska kierowniczego, 1 - respondent kieruje personelem średniego szczebla, 2 – respondent kieruje personelem wyższego szczebla). Oszacowania modelu przedstawia $reg1$ poniżej.

Następnie starano się poprawić formę funkcyjną modelu dokonując szeregu jego modyfikacji. Najpierw zastąpiono zmienną $wieki$ $wiek_2$ następującymi zmiennymi:

$$wiek_30 = \begin{cases} 0 & \text{dla } wieku < 30 \\ wiek - 30 & \text{dla } wieku \geq 30 \end{cases}$$
$$wiek_2_30 = \begin{cases} 0 & \text{dla } wieku < 30 \\ wiek^2 - 30^2 & \text{dla } wieku \geq 30 \end{cases}$$

Wyniki oszacowania tego modelu przedstawia $reg2$ poniżej.

Następnie zastąpiono zmienną $wykszt$ zmienną $wyzsze$ (1 – jeśli respondent ma wykształcenie wyższe, 0 – w przeciwnym przypadku). Wyniki oszacowania tego modelu przedstawia $reg3$ poniżej.

W celu dalszej poprawy formy funkcyjnej wprowadzono interakcję pomiędzy zmienną $wyzsze$ a zmienną $wlasciciel$. Wyniki oszacowania tego modelu przedstawia $reg4$ poniżej.

W kolejnym kroku usunięto interakcję pomiędzy zmienną $wyzsze$ a zmienną $wlasciciel$, natomiast zmienne $wyzsze$ i zmienną $wlasciciel$ zastąpiono zmienną $wykszt_wlasciciel$ (0 – jeśli respondent ma wykształcenie podstawowe lub średnie, 1 – jeśli respondent ma wykształcenie wyższe i nie posiada własnej firmy, 2 – jeśli respondent ma wykształcenie wyższe i posiada własną firmę). Wyniki oszacowania tego modelu przedstawia $reg5$ poniżej.

W ostatnim kroku w celu poprawy formy funkcyjnej wprowadzono interakcję pomiędzy zmienną $plec$ a zmienną $kieruje$. Wyniki oszacowania tego modelu przedstawia $reg6$ poniżej.

Hipotezy testować na poziomie istotności 0,05. Odpowiedzi uzasadnić podając p -value.

ln wyn	reg1	reg2	reg3	reg4	reg5	reg6
wiek	0.0423***					
wiek 2	-0.0005***					
plec	-0.3049***	-	-0.3049***	-	-	-
Imiasto 1	0.1635**	0.1707**	0.1850***	0.1777***	0.1652**	0.1622**
Imiasto 2	0.2143***	0.2164***	0.2324***	0.2233***	0.2107***	0.2095***
Imiasto 3	0.2921***	0.2925***	0.3102***	0.2980***	0.2845***	0.2846***
Imiasto 4	0.3869***	0.3869***	0.4071***	0.3976***	0.3847***	0.3839***
Iwyksztal~2	0.1055**	0.1075**				
Iwyksztal~3	0.3953***	0.4062***				
wlasciciel	0.1210***	0.1255***	0.1221***	0.0666		
Ikieruje 2	0.2467***	0.2462***	0.2563***	0.2419***	0.2453***	0.2777***
Ikieruje 3	0.3899***	0.3891***	0.3991***	0.3737***	0.3725***	0.4699***
wiek 30		0.0654***	0.0659***	0.0670***	0.0682***	0.0671***
wiek 2 30		-	-0.0008***	-	-	-
wyzsze		0.0008***	0.3054***	0.2264***		
IwyzXwla ~1				0.6231***		
Iwykszt w~1					0.2157***	0.2065***
Iwykszt w~2					0.9015***	0.8586***
IpleXkie ~2						-0.0608
IpleXkie ~3						-0.2593**
cons	5.4445***	6.1741***	6.2601***	6.2812***	6.3033***	6.2896***
RSS	264.037	264.744	265.845	260.470	261.120	259.813
R2	0.286	0.284	0.281	0.296	0.294	0.297
R2 adjusted	0.278	0.276	0.274	0.288	0.287	0.289
Statystyka F	35.710	35.376	38.065	37.420	40.515	34.804
			test RESET			
Statystyka testowa	F = 8.86	F = 8.05	F = 7.37	F = 3.21	F = 2.32	F = 1.58
p-value	Prob > F = 0.0000	Prob > F = 0.0000	Prob > F = 0.0001	Prob > F = 0.0223	Prob > F = 0.0737	Prob > F = 0.1914

*** - zmienna istotna na poziomie 1%

1. Na podstawie odpowiedniej miary dopasowania określić, który z modeli jest najlepiej dopasowany do danych.
2. Na podstawie testu RESET określić, który z modeli ma najlepszą formę funkcyjną.
3. Po oszacowaniu modelu przedstawionego w *reg6* obliczono dźwignię (*dzwignia*), standaryzowane reszty (*reszty_st*) i odległość Cook'a (*cook*). Obserwacje, dla których statystyki te były największe znajdują się w tabeli poniżej. Zakładając, że respondenci pracują na pełen etat oraz, że minimalne ustawowe wynagrodzenie wynosiło w okresie badania 1317 zł wyjaśnić na podstawie tabeli, które obserwacje budzą podejrzenia i dlaczego?

1.

wyn	plec	wiek	wlasci~1	miasto	kieruje	wyksztal~e	dzwignia	reszty_st	cook
50	kob.	37	nie	od 10 tyś do 100 tyś	kieruje prac.wyż.szczeb.	średnie	.0335269	-5.401628	.0778592

4.

wyn	plec	wiek	wlasci~1	miasto	kieruje	wyksztal~e	dzwignia	reszty_st	cook
170	męż.	22	Nie	Od 5 tyś do 10 tyś	kieruje prac.wyż.szczeb.	średnie	.0349523	-3.686431	.0378614

16.

wyn	plec	wiek	wlasci~1	miasto	kieruje	wyksztal~e	dzwignia	reszty_st	cook
500	męż.	25	Tak	wieś	kieruje prac.śre.szczeb.	wyższe	.0673235	-2.512697	.0350568

26.

wyn	plec	wiek	wlasci~1	miasto	kieruje	wyksztal~e	dzwignia	reszty_st	cook
250	kob.	45	Nie	od 5 tyś do 10 tyś	kieruje prac.śre.szczeb.	średnie	.0224564	-2.058784	.00749

1079.

dochod	plec	wiek	wlasci~1	miasto	kieruje	wyksztal~e	dzwignia	reszty_st	cook
4500	męż.	27	Tak	od 100 tyś do 500 tyś	kieruje prac.wyż.szczeb.	średnie	.0234298	2.81261	.0145996

Następnie dla modelu przedstawionego w *reg6* przeprowadzono testy diagnostyczne wyniki, których znajdują się poniżej.

Jarque-Berra test statistic: chi2 = 67.69 p-value = 0.0000
 Chow test statistic: F = 9.94 p-value = 1.708e-17
 White's general test statistic: chi2 = 61.52 p-value = 0.3510
 Breusch-Pagan test statistic: chi2 = 18.78 p-value = 0.0652

4. Zbadać, czy błąd losowy ma rozkład normalny.
5. Zbadać, czy w parametry w modelu są stabilne.

6. Zbadać, czy w modelu występuje heteroskedastyczność.
7. Jeżeli model nie spełnia założeń KMRL określić:
 - a) Które założenia nie są spełnione?
 - b) Jak to ma konsekwencje dla interpretacji modelu i wnioskowania statystycznego?
 - c) W jaki sposób można rozwiązać problemy zasygnalizowane przez wyniki testów?

Rozwiązanie Zadanie 3

1. Na podstawie R^2_{adjusted} (dopasowane R^2) można stwierdzić, iż najlepiej dopasowany do danych jest model *reg6*, dla którego R^2_{adjusted} wynosi 0.289. Jest to największa wartość R^2_{adjusted} wśród modeli *reg1-reg6*. Ze względu na różną liczbę zmiennych niezależnych w poszczególnych modelach nie możemy stosować do porównywania ich dopasowania zwykłego R^2 .
2. Na podstawie testu RESET można stwierdzić, iż model *reg6* ma najlepszą formę funkcyjną. Dla tego modelu statystyka F testu RESET ma najniższą wartość (1.58) spośród wszystkich modeli i najwyższe p-value (0.1914). Jednocześnie model ten zawiera istotne zmienne niezależne.
3. Podejrzenia budzą obserwacje numer 1, 4, 16, 26. W przypadku obserwacji numer 1 kobieta w wieku 37 lat z wykształceniem średnim, nie posiadająca firmy, kierująca pracownikami wyższego szczebla, mieszkająca w miejscowości od 10 tys. do 100 tys. mieszkańców uzyskuje wynagrodzenie miesięczne równe 50 zł. W przypadku obserwacji numer 4 mężczyzna w wieku 22 lat z wykształceniem średnim, nie posiadający firmy, kierujący pracownikami wyższego szczebla, mieszkający w miejscowości od 5 tys. do 10 tys. mieszkańców uzyskuje wynagrodzenie miesięczne równe 170 zł. W przypadku obserwacji numer 16 mężczyzna w wieku 25 lat z wykształceniem wyższym, posiadający firmę, kierujący pracownikami średniego szczebla, mieszkający na wsi, uzyskuje wynagrodzenie miesięczne równe 500 zł. W przypadku obserwacji 26 kobieta w wieku 45 lat z wykształceniem średnim, nie posiadająca firmy, kierująca pracownikami średniego szczebla, mieszkająca w miejscowości od 5 tys. do 10 tys. mieszkańców, uzyskuje wynagrodzenie miesięczne równe 250 zł. Biorąc pod uwagę fakt, iż w okresie badania minimalne ustawowe wynagrodzenie wynosiło 1317 zł a respondenci pracowali na pełen etat wyżej wymienione miesięczne wynagrodzenia, uwzględniając pozostałe charakterystyki respondentów wydają się być mało prawdopodobnymi.

W przeciwieństwie do powyższych obserwacji obserwacja numer 1079 nie budzi podejrzeń: mężczyzna w wieku 27 lat z wykształceniem średnim, posiadający firmę, kierujący pracownikami wyższego szczebla, mieszkający w miejscowości od 100 tys. do 500 tys. mieszkańców uzyskuje wynagrodzenie miesięczne równe 4500 zł, co wydaje się wielkością prawdopodobną, uwzględniając pozostałe charakterystyki.

4. Normalność błędu losowego testujemy za pomocą:
 - a) testu Jarque-Berra:
 - i. hipoteza zerowa: zaburzenie losowe ma rozkład normalny.
 - ii. wartość statystyki testowej wynosi $\chi^2 = 67.79$ oraz $p\text{-value} = 0.000 < 0.05$, czyli odrzucamy hipotezę zerową o normalności zaburzenia losowego.
5. Stabilność parametrów testujemy za pomocą:

- a) testu Chowa:
 - i. hipoteza zerowa: parametry modelu są niestabilne.
 - ii. wartość statystyki testowej wynosi $F = 9.94$ oraz $p - \text{value} = 1.708e-17 < 0.05$, czyli odrzucamy hipotezę zerową o stabilności parametrów.

6. Występowanie heteroskedastyczności testujemy za pomocą:

- a) testu White'a:
 - i. hipoteza zerowa: homoskedastyczność składnika losowego.
 - ii. wartość statystyki testowej wynosi $\chi^2 = 61.52$ oraz $p - \text{value} = 0.3510 > 0.05$, czyli nie mamy podstaw do odrzucenia hipotezy zerowej o homoskedastyczności składnika losowego.
- b) testu Breuscha-Pagana:
 - i. hipoteza zerowa: homoskedastyczność składnika losowego.
 - ii. wartość statystyki testowej wynosi $\chi^2 = 18.78$ oraz $p - \text{value} = 0.0652 > 0.05$, czyli nie mamy podstaw do odrzucenia hipotezy zerowej o homoskedastyczności składnika losowego.

7. Odpowiedzi są następujące:

- a) Nie jest spełnione założenie o normalności rozkładu składnika losowego. Nie jest także spełnione założenie o liniowej zależności między zmienną zależną a zmiennymi niezależnymi.
- b) Konsekwencje dla interpretacji modelu i wnioskowania statystycznego są następujące:
 - i. Próba zawiera 30 000 obserwacji (można przyjąć, iż jest to duża próba). Dla dużych prób rozkłady statystyk są bliskie standardowym rozkładom.
 - ii. Niespełnienie założenia liniowej zależności między zmienną zależną a zmiennymi niezależnymi podważa interpretację ekonomiczną modelu (interpretacja oszacowanych parametrów). Takie własności jak nieobciążoność czy efektywność estymatora MNK wyprowadzane są przy założeniu liniowej zależności między zmienną zależną a zmiennymi niezależnymi.
- c) Niestabilność parametrów: można próbować oszacować model osobno dla podróbek lub wprowadzić do modelu interakcje między zmiennymi zerojedynkowymi związanymi z podziałem na podróbki i odpowiednimi zmiennymi objaśniającymi.