

Zadanie 1

$$\text{pensja}_i = \beta_0 + \beta_1 \text{plec}_i + \beta_2 \text{start_pe}_i + \beta_3 \text{doswiad}_i + \beta_4 \text{wyzsze}_i + \varepsilon_i$$

gdzie:

pensja - miesięczna pensja inżynierów w Tajlandii (j.p.),

start_pe - początkowa pensja (j.p.)

doswiad - doświadczenie zawodowe w obecnym miejscu pracy,

wyzsze - wykształcenie (1-wyższe, 0-inne),

plec - płeć (1-mężczyzna, 0-kobieta).

Plik Edycja Testy Wykresy Dane modelu LaTeX Wykresy					
Model 1: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403					
Zmienna zależna: pensja					
	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	-5826,73	3240,86	-1,798	0,072951 *
1)	plec	2261,28	908,779	2,488	0,013245 **
3)	start_pe	1,85317	0,0616645	30,052	< 0,00001 ***
4)	doswiad	511,376	222,099	2,302	0,021824 **
5)	wyzsze	5101,85	1235,46	4,130	0,000044 ***
Srednia arytmetyczna zmiennej zależnej = 34683,4					
Odchylenie standardowe zmiennej zależnej = 17041,5					
Suma kwadratów reszt = 2,38656e+010					
Błąd standardowy reszt = 7743,62					
Nieskorygowany R-kwadrat = 0,795578					
Skorygowany R-kwadrat = 0,793524					
Statystyka F (4, 398) = 387,239 (p-value < 0,00001)					
Kryterium informacyjne Akaike (AIC) = 8366,06					
Kryterium bayesowskie Schwarz (BIC) = 8386,06					
Zamknij					

a) Przeprowadzono test RESET.

Pomocnicze równanie regresji dla testu specyfikacji RESET

Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403

Zmienna zależna: pensja

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	838,251	8465,61	0,099	0,921174
1)	plec	755,200	976,757	0,773	0,439883
3)	start_pe	1,12711	0,737952	1,527	0,127473
4)	doswiad	287,250	272,260	1,055	0,292042
5)	wyzsze	3109,21	1900,83	1,636	0,102694
6)	yhat^2	1,10240E-05	6,92509E-06	1,592	0,112208
7)	yhat^3	-8,23164E-011	3,72320E-011	-2,211	0,027613 **

Statystyka testu: F = 8,177671,

z p-value = P(F(2,396) > 8,17767) = 0,000331

Czy model ma poprawną formę funkcyjną ?

- b) W naszym przypadku populacją są inżynierowie w Tajlandii. Czy można jednak przypuszczać, że na zarobki kobiet-inżynierów i mężczyzn-inżynierów inaczej wpływa początkowa ich pensja, doświadczenie zawodowe i wykształcenie?

Czy można twierdzić, że wpływ tych zmiennych zależy od płci?

Na to pytanie można odpowiedzieć przeprowadzając **test stabilności parametrów Chowa**.

1. Szacujemy model z wyłączeniem zmiennej dzielącej populację na podgrupy (zmienna płeć):

$$\text{pensja}_i = \beta_0 + \beta_1 \text{start_pe}_i + \beta_2 \text{doswiad}_i + \beta_3 \text{wyzsze}_i + \varepsilon_i$$

Plik	Edycja	Testy	Wykresy	Dane modelu	LaTeX	Wykresy
Model 2: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403						
Zmienna zależna: pensja						
	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value	
0)	const	-6066,28	3260,44	-1,861	0,063540 *	
3)	start_pe	1,93287	0,0530357	36,445	< 0,00001 ***	
4)	doswiad	517,918	223,524	2,317	0,021006 **	
5)	wyzsze	4924,61	1241,40	3,967	0,000086 ***	
Srednia arytmetyczna zmiennej zależnej = 34683,4						
Odchylenie standardowe zmiennej zależnej = 17041,5						
Suma kwadratów reszt = 2,42368e+010						
Błąd standardowy reszt = 7793,84						
Nieskorygowany R-kwadrat = 0,792398						
Skorygowany R-kwadrat = 0,790837						
Statystyka F (3, 399) = 507,649 (p-value < 0,00001)						
Kryterium informacyjne Akaike (AIC) = 8370,28						
Kryterium bayesowskie Schwarz (BIC) = 8386,28						

Obliczamy RSS (RSS tego modelu).

2. Szacujemy model dla próby mężczyzn:

$$\text{pensja}_i = \alpha_0 + \alpha_1 \text{start_pe}_i + \alpha_2 \text{doswiad}_i + \alpha_3 \text{wyzsze}_i + \varepsilon_{1i}$$

Plik	Edycja	Testy	Wykresy	Dane modelu	LaTeX	Wykresy
Model 3: Estymacja KMNK z wykorzystaniem 213 obserwacji 1-213						
Zmienna zależna: pensja						
	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value	
0)	const	-5086,06	5571,82	-0,913	0,362390	
3)	start_pe	1,86899	0,0815821	22,909	< 0,00001	***
4)	doswiad	599,774	376,226	1,594	0,112405	
5)	wyzsze	5114,80	1967,99	2,599	0,010016	**
Srednia arytmetyczna zmiennej zależnej = 42832,3						
Odchylenie standardowe zmiennej zależnej = 19153,9						
Suma kwadratów reszt = 2,05869e+010						
Błąd standardowy reszt = 9924,82						
Nieskorygowany R-kwadrat = 0,735307						
Skorygowany R-kwadrat = 0,731508						
Statystyka F (3, 209) = 193,531 (p-value < 0,00001)						
Kryterium informacyjne Akaike (AIC) = 4528,82						
Kryterium bayesowskie Schwarz (BIC) = 4542,27						

Obliczamy RSS1 (RSS tego modelu)

3. Szacujemy model dla próby kobiet:

$$\text{pensja}_i = \gamma_0 + \gamma_1 \text{start_pe}_i + \gamma_2 \text{doswiad}_i + \gamma_3 \text{wyzsze}_i + \varepsilon_{2i}$$

Plik	Edycja	Testy	Wykresy	Dane modelu	LaTeX	Wykresy
Model 4: Estymacja KMNK z wykorzystaniem 190 obserwacji 1-190						
Zmienna zależna: pensja						
	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value	
0)	const	-1344,64	3108,84	-0,433	0,665864	
3)	start_pe	1,62178	0,140852	11,514	< 0,00001	***
4)	doswiad	394,514	182,935	2,157	0,032322	**
5)	wyzsze	6213,09	1290,97	4,813	< 0,00001	***
Srednia arytmetyczna zmiennej zależnej = 25548,1						
Odchylenie standardowe zmiennej zależnej = 6888,88						
Suma kwadratów reszt = 3,21755e+009						
Błąd standardowy reszt = 4159,16						
Nieskorygowany R-kwadrat = 0,641271						
Skorygowany R-kwadrat = 0,635485						
Statystyka F (3, 186) = 110,832 (p-value < 0,00001)						
Kryterium informacyjne Akaike (AIC) = 3709,72						
Kryterium bayesowskie Schwarz (BIC) = 3722,71						

Obliczamy RSS2 (RSS tego modelu)

Przechodzimy do testowania stabilności parametrów w obu podpróbach. Hipoteza zerowa brzmi: wszystkie szacowane parametry (również stała) są takie same w obu podpróbach. Wprowadzamy następujące oznaczenia:

m – liczba wyodrębnionych podpróbek,

K – ilość szacowanych parametrów (ta sama wartość dla wszystkich modeli),

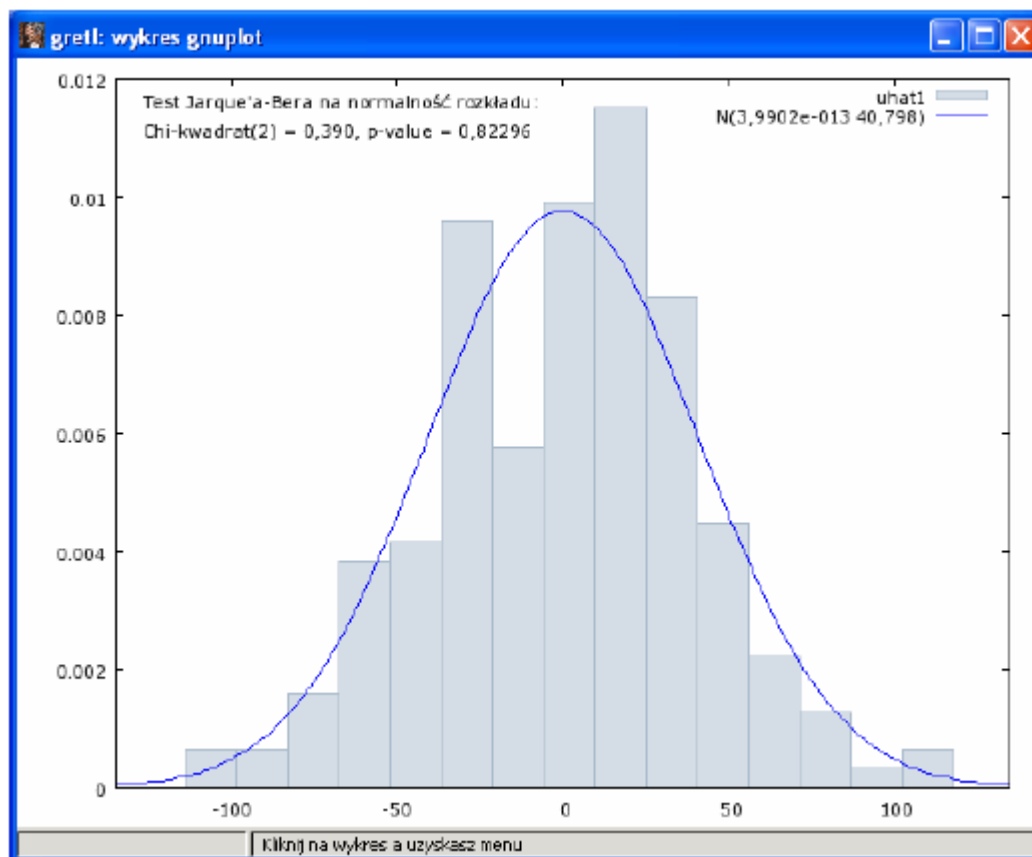
n – liczebność próbki wyjściowej.

- Odczytujemy wartość krytyczną $F^*(K(m-1), N-mK)$
- Liczymy wartość statystyki testowej:

$$F = \frac{(\hat{e}'e - \hat{e}_1'e_1 - \hat{e}_2'e_2)/[K(m-1)]}{(\hat{e}_1'e_1 + \hat{e}_2'e_2)/(N-mK)}$$

Sprawdzić, czy parametry modelu są stabilne.

c) Czy reszty mają rozkład normalny ?



Zadanie 2

Rozważmy następujący model (próba przekrojowa – dane dla polskich powiatów za rok 2002):

$$\text{dochod}_i = \beta_1 + \beta_2 \text{regon}_i + \beta_3 \text{oferty}_i + \beta_4 \text{lekarze}_i + \varepsilon_i$$

Oznaczenia:

dochod – dochody budżetowe powiatów w tys. zł

regon – liczba jednostek regon na 1 tys. mieszkańców

oferty – liczba ofert pracy na 1 tys. bezrobotnych

lekarze – liczba lekarzy na 1 tys. mieszkańców

Oto oszacowania:

Model 2: Estymacja KMNK z wykorzystaniem 380 obserwacji 1-380

Zmienna zależna: dochod

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	834,065	44,0760	18,923	< 0,00001 ***
9)	regon	5,21084	0,653627	7,972	< 0,00001 ***
11)	oferty	4,80668	1,90457	2,524	0,012022 **
8)	lekarze	139,968	14,3691	9,741	< 0,00001 ***

Srednia arytmetyczna zmiennej zależnej = 1499,51

Odchylenie standardowe zmiennej zależnej = 348,283

Suma kwadratów reszt = 2,1718e+007

Błąd standardowy reszt = 240,334

Nieskorygowany R-kwadrat = 0,527594

Skorygowany R-kwadrat = 0,523825

Statystyka F (3, 376) = 139,975 (p-value < 0,00001)

Kryterium informacyjne Akaika (AIC) = 5248,72

Kryterium bayesowskie Schwarza (BIC) = 5264,48

a) Przeprowadzono test **White'a** na heteroskedastyczność. Test ten stosuje się w przypadku, gdy nie wiemy która ze zmiennych odpowiedzialna jest za powodowanie heteroskedastyczności.

W wyniku przeprowadzenia testu otrzymano:

Test White'a na heteroskedastyczność wariancji

Estymacja KMNK z wykorzystaniem 380 obserwacji 1-380

Zmienna zależna: uhat^2

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	-16417,3	70872,6	-0,232	0,816942
9)	regon	-2662,76	1727,83	-1,541	0,124147
11)	oferty	6584,55	9330,57	0,706	0,480821
8)	lekarze	145994	34274,6	4,260	0,000026 ***
14)	sq_regon	32,8983	11,5536	2,847	0,004653 ***
15)	reg_ofer	75,2968	98,1972	0,767	0,443694
16)	reg_leka	-1445,17	449,142	-3,218	0,001407 ***
17)	sq_ofert	-102,440	41,4872	-2,469	0,013993 **
18)	ofe_leka	-2394,14	1697,11	-1,411	0,159168
19)	sq_lekar	3543,72	4563,63	0,777	0,437942

Nieskorygowany R-kwadrat = 0,128862

Statystyka testu: TR^2 = 48,967748,

z p-value = P(Chi-kwadrat(9) > 48,967748) = 0,000000

b) Innym testem na występowanie heteroskedastyczności jest test **Breuscha-Pagana**. Test ten zakłada, że wiemy które ze zmiennych objaśniających mogą powodować zmiany wariancji składnika losowego. W naszym przypadku możemy podejrzewać, że wariancja dochodu może zależeć od liczby regonów dla powiatów oraz od liczby lekarzy – w powiatach o dużej liczbie regonów dochód może być dużo bardziej zróżnicowany, niż w powiatach o małej liczbie, podobnie może być dla liczby lekarzy.

Statystyka testu: LM = 78,71

z wartością $p = P(\text{Chi-kwadrat}(9) > 78,71) = 0,000000$

Zadanie 3

dochod – miesięczne zarobki netto

wiek – wiek w latach

miasto_0 – 1, jeśli wieś; 0 wpp

miasto_1 – 1, jeśli miasto do 5tyś; 0 wpp

miasto_2 – 1, jeśli miasto do 50tyś; 0 wpp

miasto_3 – 1, jeśli miasto do 100tyś; 0 wpp

miasto_4 – 1, jeśli miasto powyżej 100tyś; 0 wpp

stan – 1, jeśli kawaler, panna oraz 0 wpp

edu_0 – 1, jeśli wykształcenie podstawowe; 0 wpp

edu_1 – 1, jeśli wykształcenie średnie; 0 wpp

edu_2 – 1, jeśli wykształcenie wyższe; 0 wpp

plec – 1, jeżeli mężczyzna; 0 wpp

Model 3: Estymacja KMNK z wykorzystaniem 1044 obserwacji 1-1044

Zmienna zależna: dochod

	<i>Współczynnik</i>	<i>Błąd stand.</i>	<i>t-Student</i>	<i>wartość p</i>	
const	424,198	136,285	3,1126	0,00191	***
wiek	3,81795	2,52289	1,5133	0,13050	
plec	300,865	48,1658	6,2464	<0,00001	***
stan	-153,072	67,112	-2,2808	0,02276	**
miasto_1	-45,9128	166,553	-0,2757	0,78286	
miasto_2	31,6114	62,1547	0,5086	0,61115	
miasto_3	55,3105	170,485	0,3244	0,74568	
miasto_4	498,527	57,8194	8,6222	<0,00001	***
edu_1	201,435	74,1331	2,7172	0,00669	***
edu_2	1064,85	90,6028	11,7529	<0,00001	***

Średnia arytmetyczna zmiennej zależnej = 1231,59

Odchylenie standardowe zmiennej zależnej = 885,25

Suma kwadratów reszt = 5,91672e+008

Błąd standardowy reszt = 756,45

Wsp. determinacji $R^2 = 0,27612$

Skorygowany $R^2 = 0,26982$

Statystyka F (9, 1034) = 43,8243 (wartość p < 0,00001)

RESET: Statystyka testu: F = 25,421338, p-value = 0,0000167
Jarque-Bera: Statystyka testu: chi2=1366,77, p-value=0,0000
Chow: Statystyka testu: F=10,277, p-value=0,0000
White: Statystyka testu: F= 65,351685, p-value=0,003777
Breusch-Pagan: Statystyka testu: F= 732,096328, p-value=0,0000

Na podstawie powyższych wyników odpowiedzieć na następujące pytania. Każdą odpowiedź uzasadnić wielkościami odpowiednich statystyk testowych. Testy przeprowadzić na poziomie istotności $\alpha = 0,05$.

1. Zinterpretować parametry modelu.
2. Sprawdzić istotność oraz łączną istotność uzyskanych wyników.
3. Chcemy sprawdzić, czy wpływ miejsca zamieszkania, wykształcenia oraz wieku na dochód jest taki sam dla kobiet i mężczyzn. Proszę w tym celu przeprowadzić test stabilności parametrów Chow'a.
4. Czy spełnione jest założenie o homoscedastyczności składnika losowego?
5. Czy spełnione jest założenie o normalności składnika losowego?
6. Proszę przeprowadzić test Reset.
7. Jeśli model nie spełnia założeń KMRL określić:
 - (a) które założenie nie jest spełnione?
 - (b) jakie ma to konsekwencje dla wnioskowania statystycznego?
 - (c) jakie są metody radzenia sobie z tym problemem?

Zadanie 4

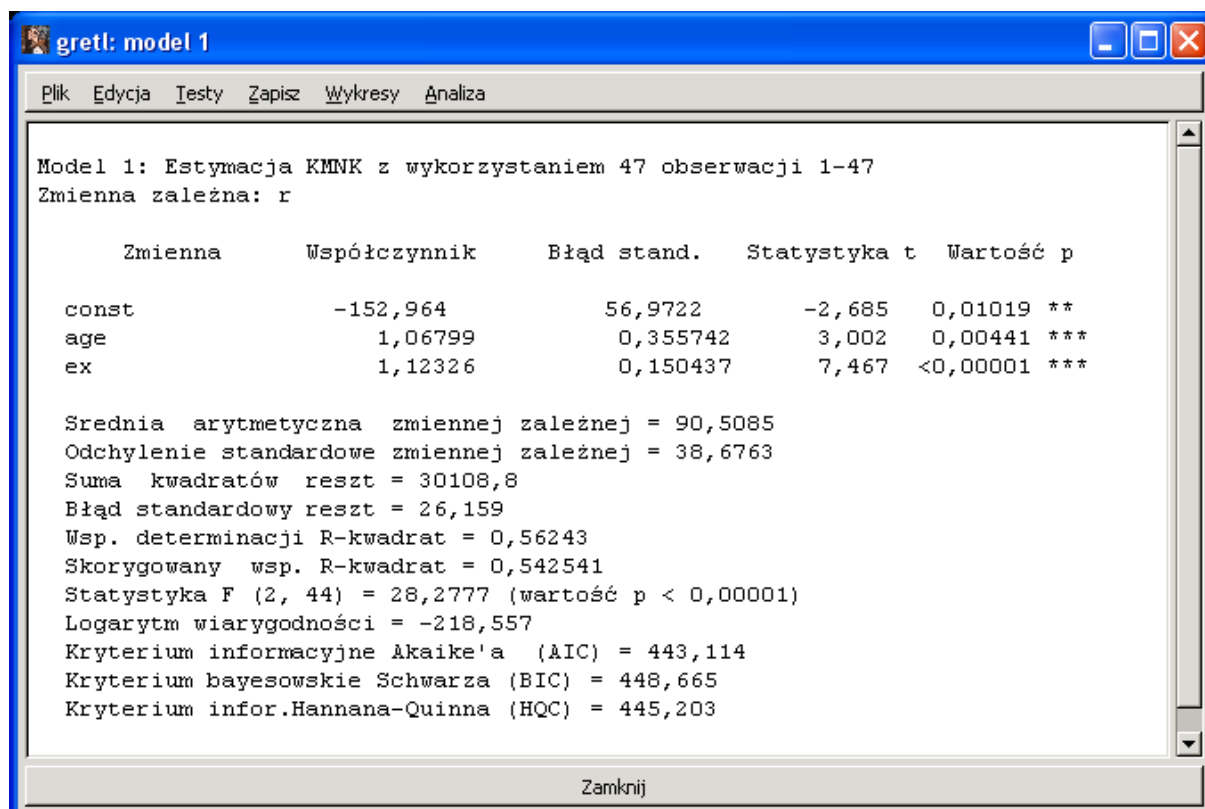
Opis zmiennych:

r - stopa przestępczości; liczba przestępstw zanotowanych przez policję na 1000000 osób

age - liczba mężczyzn w wieku 14-24 lata na 1000 mieszkańców

ex -wydatki na policję w 1959r.

Oszacujemy regresję w której zmienna zależna jest stopa przestępczości (r), natomiast zmienne niezależne to age i ex. Poniżej wyniki oszacowanej regresji:



Testy diagnostyczne:

RESET: Statystyka testu: F = 0,13228, p-value = 0,876

Jarque-Bera: Statystyka testu: chi2=0,499, p-value 0,7791

White: Statystyka testu: F= 25,035, p-value=0,000137

Breusch-Pagan: Statystyka testu: F= 32,04387, p-value=0,0000

Na podstawie powyższych wyników odpowiedzieć na następujące pytania. Każdą odpowiedź uzasadnić wielkościami odpowiednich statystyk testowych. Testy przeprowadzić na poziomie istotności $\alpha = 0,05$.

1. Określić, czy model jest dobrze dopasowany oraz czy zbiór zmiennych niezależnych istotnie objaśnia zmienna zależną.
2. Podać, które zmienne w modelu są istotne.

3. Zbadać, czy w modelu występuje heteroskedastyczność.
4. Sprawdzić, czy forma funkcyjna modelu jest prawidłowa.
5. Przetestować, czy błąd losowy w modelu ma rozkład normalny.
6. Jeśli model nie spełnia założeń KMRL określić:

(a) które założenie nie jest spełnione?

(b) jakie ma to konsekwencje dla wnioskowania statystycznego?

(c) jakie są metody radzenia sobie z tym problemem?