

Zadanie 0

Do oszacowania modelu wykorzystano następującą bazę danych:

Nazwa zmiennej	Pełny opis zmiennej
const	stała - automatycznie generowana
bwght	waga dziecka po urodzeniu (gramy)
npvis	ilość wizyt lekarskich, w celu sprawdzenia poprawności rozwoju ciąży
mage	wiek matki (lata)
cigs	przeciętna ilość palonych dziennie papierosów przez matkę (sztuki)
male	pleć dziecka (1-chłopiec, 0-dziewczynka)
mwhite	rasa matki (1-biała, 0-inna)

Oszacowano następujący model:

Model 1: Estymacja KMNK z wykorzystaniem 1420 obserwacji 1-1574

Liczba pominiętych niekompletnych obserwacji: 154

Zmienna zależna: bwght

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	3100,19	108,648	28,534	<0,00001
npvis	11,9141	3,92531	3,035	0,00245
mage	1,71923	3,10631	0,553	0,58003
cigs	-12,7714	3,39703	-3,760	0,00018
male	82,5453	29,3790	2,810	0,00503
mwhite	125,238	45,5848	2,747	0,00608

Srednia arytmetyczna zmiennej zależnej = 3428,08

Odchylenie standardowe zmiennej zależnej = 558,719

Suma kwadratów reszt = 4,31222e+008

Błąd standardowy reszt = 552,237

Wsp. determinacji R-kwadrat = 0,0265091

Skorygowany wsp. R-kwadrat = 0,0230667

Statystyka F (5, 1414) = 7,70091 (wartość p < 0,00001)

1. Czy łącznie wszystkie zmienne w modelu są istotne na poziomie istotności 5%?
2. Jaka część zmienności wagi dziecka po urodzeniu została wyjaśniona przez model?
3. Jakie jest odchylenie standardowe błędu losowego w tym modelu?

Przykład

Zbadajmy związek między płacą a stażem pracy, wykształceniem (mierzone liczbą lat nauki) i płcią respondenta.

Definiujemy następujące zmienne:

$$staz_2 = \begin{cases} 1 & \text{gdy lata stazu ponad 12 a poniżej 31} \\ 0 & \text{w innym przypadku} \end{cases}$$

$$staz_3 = \begin{cases} 1 & \text{gdy lata stazu powyżej 30} \\ 0 & \text{w innym przypadku} \end{cases}$$

Przyjmujemy, że kategorią referencyjną jest staż pracy poniżej 13 lat. Wyniki estymacji:

$$placa_i = B_1 + B_2plec_i + B_3nauka_i + B_4staz_2_i + B_5staz_3_i + \varepsilon_i.$$

Zmienna zależna: placa

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	-525,661	143,519	-3,663	0,00026 ***
plec	384,495	48,3305	7,956	<0,00001 ***
nauka	78,7077	9,35766	8,411	<0,00001 ***
staz_2	238,453	78,7701	3,027	0,00253 ***
staz_3	257,341	83,2517	3,091	0,00205 ***

Wsp. determinacji R-kwadrat = 0,101585

Chcemy przetestować łączną istotność zmiennych zero-jedynkowych dotyczących stażu pracy. Hipoteza zerowa przyjmuje następującą postać: $H_0 : B_4 = B_5 = 0$. Szacujemy regresję z restrykcjami (pomijamy zmienną staz_2 i staz_3):

Zmienna zależna: placa

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	-253,286	116,291	-2,178	0,02962 **
plec	379,392	48,4753	7,827	<0,00001 ***
nauka	74,3087	9,20875	8,069	<0,00001 ***

Wsp. determinacji R-kwadrat = 0,0929039

$$F = \frac{(R^2 - R_R^2)/g}{(1 - R^2)/(n - K)} = \frac{(0,101585 - 0,0929039)/2}{(1 - 0,101585)/(1089 - 5)} = 5,2371746$$

Odczytujemy wartość krytyczną z tablic: $F^*(2,1084) = 3,0040265$. Ponieważ wartość statystyki testowej przewyższa wartość krytyczną, więc odrzucamy hipotezę zerową, która

zakłada, iż parametry przy zero-jedynkowych zmiennych dotyczących stażu pracy jednocześnie wynoszą zero.

Zadanie 1

Oszacowano model PKB następującej postaci:

$$PKB_t = \beta_1 + \beta_2 \text{konsump}_t + \beta_3 \text{inwest}_t + \beta_4 \text{wyd_rz}_t + \beta_5 \text{pop}_t + \beta_6 \text{infl}_t + \varepsilon_t$$

gdzie zmienne kolejno oznaczają konsumpcję, inwestycje, wydatki rządowe, populację i inflację.

Pierwsze trzy ze zmiennych objaśniających, podobnie jak zmienna objaśniana, wyrażone są w miliardach PLN, zmienna *pop* wyrażona jest w setkach tysięcy, zaś *infl* w procentach.

Oszacowanie w programie Gretl jest następujące:

Model 3: Estymacja KMNK z wykorzystaniem 200 obserwacji

Zmienna zależna: PKB

	Zmienna	Współczynnik	Błąd stand.
0)	const	-127,279	68,4455
2)	konsump	1,17111	0,0218024
3)	inwest	0,462004	0,0400625
4)	wyd_rz	0,597883	0,0592857
10)	pop	1,23561	0,546183
11)	infl	3,92902	0,921947

Srednia arytmetyczna zmiennej zależnej = 4577,19

Odchylenie standardowe zmiennej zależnej = 2108,93

Suma kwadratów reszt = 313924

Błąd standardowy reszt = 39,9189

Nieskorygowany R-kwadrat = 0,999651

- Oceń które zmienne są w modelu istotne na poziomie istotności 5%.
- Zinterpretuj parametry przy zmiennych istotnych.
- Czy równanie regresji jest istotne (czy zmienne objaśniające są łącznie istotne)?

Zadanie 2

Do oszacowania modelu wykorzystano następującą bazę danych:

Nazwa zmiennej	Pełny opis zmiennej
const	stała - automatycznie generowana
bwght	waga dziecka po urodzeniu (gramy)
npvis	ilość wizyt lekarskich, w celu sprawdzenia poprawności rozwoju ciąży
mage	wiek matki (lata)
cigs	przeciętna ilość palonych dziennie papierosów przez matkę (sztuki)
male	pleć dziecka (1-chłopiec, 0-dziewczynka)
mwhite	rasa matki (1-biała, 0-inna)

Oszacowano następujący model:

Model 1: Estymacja KMNK z wykorzystaniem 1420 obserwacji 1-1574
Liczba pominiętych niekompletnych obserwacji: 154
Zmienna zależna: bwght

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	3100,19	108,648	28,534	<0,00001
npvis	11,9141	3,92531	3,035	0,00245
mage	1,71923	3,10631	0,553	0,58003
cigs	-12,7714	3,39703	-3,760	0,00018
male	82,5453	29,3790	2,810	0,00503
mwhite	125,238	45,5848	2,747	0,00608

Srednia arytmetyczna zmiennej zależnej = 3428,08
Odchylenie standardowe zmiennej zależnej = 558,719
Suma kwadratów reszt = 4,31222e+008
Błąd standardowy reszt = 552,237
Wsp. determinacji R-kwadrat = 0,0265091
Skorygowany wsp. R-kwadrat = 0,0230667
Statystyka F (5, 1414) = 7,70091 (wartość p < 0,00001)

Dodatkowo oszacowano model bez zmiennych *mage* i *npvis*:

Model 2: Estymacja KMNK z wykorzystaniem 1420 obserwacji 1-1574

Liczba pominiętych niekompletnych obserwacji: 154

Zmienna zależna: bwght

Zmienna	Współczynnik	Błąd stand.	Statystyka t	Wartość p
const	3289,22	46,1112	71,332	<0,00001 ***
cigs	-13,1355	3,39998	-3,863	0,00012 ***
male	80,0612	29,4367	2,720	0,00661 ***
mwhite	127,074	45,7053	2,780	0,00550 ***

Srednia arytmetyczna zmiennej zależnej = 3428,08

Odchylenie standardowe zmiennej zależnej = 558,719

Suma kwadratów reszt = 4,34254e+008

Błąd standardowy reszt = 553,784

Wsp. determinacji R-kwadrat = 0,0196646

- Zinterpretuj oszacowania parametrów Modelu 1.
- Oceń które zmienne są istotne w Modelu 1. Wykorzystaj *p-value* lub którąś z wartości krytycznych: $t_{1420} = 1,961635$, $t_{1414} = 1,961643$, $t_{1415} = 1,961641$. Podaj wartość, którą wykorzystałeś/leś.
- Czy zmienne objaśniające są w Modelu 1 łącznie istotne? Jeśli potrzebujesz, to skorzystaj z którejś z podanych wartości krytycznych: $F(5,1414)=2,22042$, $F(6,1414)=2,105$, $F(5,1420)=2,22042$. Podaj wartość, którą wykorzystałeś/leś.
- Czy z Modelu 1 można pominąć zmienne *mage* i *npvis*? Przeprowadź odpowiedni test. Możesz wykorzystać którąś z wartości krytycznych: $F(2,1420)=3,00206$, $F(2,1414)=3,00208$, $F(5,1420)=2,22042$. Podaj wartość, którą wykorzystałeś/leś.

Zadanie 3

1. Oszacowano następujący model:

$$\text{wage}_i = \beta_0 + \beta_1 \text{marr}_i + \beta_2 \text{exper}_i + \beta_3 \text{center}_i + \beta_4 \text{points}_i + \beta_5 \text{avgmin}_i + \beta_6 \text{black}_i + \varepsilon_i$$

gdzie wage – pensja gracza NBA (tys. \$), marr - stan cywilny (1-malzenstwo, 0-inny), exper - dlugosc kariery zawodowej (lata), center - pozycja gracza (1-center, 0-inna), points - srednia ilosc zdobywanych w meczu punktow, avgmin - sredni czas gry w meczu (minuty), black - rasa (1-czarna, 0-inna). Otrzymano:

Model 1: Estymacja KMNK z wykorzystaniem 219 obserwacji 1-219
Zmienna zależna: wage

	Zmienna	Współczynnik	Błąd stand.	Statystyka t
0)	const	-397,563	187,421	-2,121
1)	marr	25,0791	104,825	0,239
3)	exper	75,2539	15,3063	4,917
4)	center	471,488	132,612	3,555
5)	points	78,4417	17,3310	4,526
6)	avgmin	22,3936	10,8767	2,059
7)	black	24,8380	133,743	0,186

Srednia arytmetyczna zmiennej zależnej = 1461,33
Odchylenie standardowe zmiennej zależnej = 1023,66
Suma kwadratów reszt = 1,04476e+008
Błąd standardowy reszt = 702,007
Nieskorygowany R-kwadrat = 0.542647

Z powodu braku istotności zmiennych black i marr, usunięto je z modelu i otrzymano:

Model 3: Estymacja KMNK z wykorzystaniem 219 obserwacji 1-219
Zmienna zależna: wage

	Zmienna	Współczynnik	Błąd stand.	Statystyka t
0)	const	-375,586	154,208	-2,436
3)	exper	76,5084	14,2700	5,361
4)	center	468,409	128,989	3,631
5)	points	78,9751	17,1532	4,604
6)	avgmin	22,3154	10,8178	2,063

Srednia arytmetyczna zmiennej zależnej = 1461,33
Odchylenie standardowe zmiennej zależnej = 1023,66
Suma kwadratów reszt = 1,04517e+008
Błąd standardowy reszt = 698,854
Nieskorygowany R-kwadrat = 0,54247

Oszacowano jeszcze jeden model:

Model 4: Estymacja KMNK z wykorzystaniem 219 obserwacji 1-219
Zmienna zależna: wage

	Zmienna	Współczynnik	Błąd stand.	Statystyka t
0)	const	29,6305	189,225	0,157
1)	marr	241,683	106,433	2,271
5)	points	80,5485	18,8071	4,283
6)	avgmin	22,3035	11,6967	1,907
7)	black	-66,6837	142,537	-0,468

Srednia arytmetyczna zmiennej zależnej = 1461,33

Odchylenie standardowe zmiennej zależnej = 1023,66

Suma kwadratów reszt = 1,24913e+008

Błąd standardowy reszt = 764,007

Nieskorygowany R-kwadrat = 0,453183

- Zinterpretuj oszacowania dla Modelu 3.
- Czy można powiedzieć, że zmienne marr i black są nieistotne w pierwszym modelu i można je pominąć?
- Czy można powiedzieć, że zmienne exper i center są nieistotne w pierwszym modelu i można je pominąć?

Zadanie 4

Oszacowano następujący model:

Model 1: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403
Zmienna zależna: wage

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	-6774,10	3316,17	-2,043	0,041736
1)	male	2209,19	1023,90	2,158	0,031555
4)	swage	1,90920	0,0644172	29,638	< 0,00001
5)	exper	564,774	226,432	2,494	0,013028
10)	polytech	-371,154	985,511	-0,377	0,706664

Srednia arytmetyczna zmiennej zależnej = 34683,4
Odchylenie standardowe zmiennej zależnej = 17041,5
Suma kwadratów reszt = 2,48792e+010
Błąd standardowy reszt = 7906,37
Nieskorygowany R-kwadrat = 0,786895

gdzie *wage* – stawka płacowa inżyniera w USA (w setkach dolarów), *male* – płeć (1 – mężczyzna, 0 – kobieta), *swage* – początkowa stawka płacowa (w setkach dolarów), *exper* – doświadczenia zawodowe w obecnym miejscu pracy (w latach), *polytech* – absolwent politechniki (1 – jest absolwentem, 0 – nie jest).

Przeprowadzono również kilka dodatkowych regresji. Oto one:

Model 2: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403
Zmienna zależna: wage

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	814,426	980,411	0,831	0,406638
4)	swage	1,97395	0,0522226	37,799	< 0,00001

Srednia arytmetyczna zmiennej zależnej = 34683,4
Odchylenie standardowe zmiennej zależnej = 17041,5
Suma kwadratów reszt = 2,55858e+010
Błąd standardowy reszt = 7987,8
Nieskorygowany R-kwadrat = 0,780843

Model 3: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403
Zmienna zależna: wage

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	25548,1	1067,03	23,943	< 0,00001
1)	male	17284,3	1467,70	11,776	< 0,00001

Srednia arytmetyczna zmiennej zależnej = 34683,4
Odchylenie standardowe zmiennej zależnej = 17041,5
Suma kwadratów reszt = 8,6746e+010
Błąd standardowy reszt = 14708
Nieskorygowany R-kwadrat = 0,256972

Model 4: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403
Zmienna zależna: *exper*

	Zmienna	Współczynnik	Błąd stand.	Statystyka t	p-value
0)	const	13,8928	0,232004	59,882	< 0,00001
1)	male	0,00401383	0,226377	0,018	0,985863
4)	swage	-2,33310E-05	1,41943E-05	-1,644	0,101028
10)	polytech	0,0703662	0,217862	0,323	0,746875

Srednia arytmetyczna zmiennej zależnej = 13,5136
Odchylenie standardowe zmiennej zależnej = 1,7511
Suma kwadratów reszt = 1219,21
Błąd standardowy reszt = 1,74805
Nieskorygowany R-kwadrat = 0,0109201
Skorygowany R-kwadrat = 0,0034834
Statystyka F (3, 399) = 1,46841 (p-value = 0,223)

Model 5: Estymacja KMNK z wykorzystaniem 403 obserwacji 1-403
Zmienna zależna: *swage*

	Zmienna	Współczynnik	Błąd stand.	Statystyka t
0)	const	17477,4	2424,13	
1)	male	9301,35	645,262	
5)	exper	-288,272	175,381	
10)	polytech	-4953,65	724,641	

Srednia arytmetyczna zmiennej zależnej = 17158
Odchylenie standardowe zmiennej zależnej = 7628,79
Suma kwadratów reszt = 1,50643e+010
Błąd standardowy reszt = 6144,53
Nieskorygowany R-kwadrat = 0,356109

Polecenia:

a. Zinterpretuj oszacowania Modelu 1.

b. Oceń które zmienne objaśniające są istotne w Modelu 1. Odpowiedź uzasadnij. Jak odpowiedź na to pytanie zależy od przyjętego poziomu istotności (α)?

c. Czy zmienne z Modelu 1 są w nim łącznie istotne (czy Model 1 jest istotny)? Jeśli korzystasz z jakiejś wartości krytycznej, to podaj jej wartość.

d. Czy można z Modelu 1 pominąć zmienne: *male*, *exper*, *polytech*? Co z tego wynika? Jeśli korzystasz z jakiejś wartości krytycznej, to podaj jej wartość.

f. Czy zmienne objaśniające z Modelu 4 są łącznie istotne? Odpowiedź oprzyj na *p-value* odpowiedniego testu.

g. Które zmienne objaśniające są istotne w Modelu 5? Jeśli korzystasz z jakiejś wartości krytycznej, to podaj jej wartość.