KOLOKWIUM Z EKONOMETRII - ROZWIĄZANIA

STYCZEŃ 2024

*Czas pracy wynosi 90 min. Podpisz kartę z zadaniami oraz kartki z odpowiedziami. Odpowiedzi do poszczególnych zadań zapisz na oddzielnych kartkach. Możesz używać kalkulatora. Wartości krytyczne z wybranych rozkładów znajdują się na końcu arkusza. Maksymalna liczba punktów wynosi **32**. W przypadku wątpliwości - pytaj!*

Zadanie 1. *(10 p.)* Oszacowano dwa modele objaśniające liczbę lat edukacji danej osoby w zależności od dystansu do uniwersytetu (*dystans*, w dziesiątkach mil). Uwzględniono także zmienne takie jak: *czesne* (średnie czesne za studia w danym stanie), *wynik testu* (wynik z egzaminu w ostatniej klasie szkoły), *wysoki dochód* (=1 jeśli roczny dochód rodziny przekracza 25 tys. dolarów, 0 wpp.), *dom* (=1 jeśli rodzona posiada dom, 0 wpp.). W Tabeli 1 przedstawiono oszacowania dwóch modeli (w nawiasach podano błędy standardowe). Na podstawie wyników odpowiedz na pytania.

Tabela 1: Edukacja vs. dystans do uniwersytetu.

	Zmienna zależna:	
	lata edukacji	
	(1)	(2)
dystans	-0.070	-0.034
	(0.011)	(0.010)
czesne	0.160	-0.148
	(0.077)	(0.067)
wynik testu		0.083
		(0.003)
ojciec ukończył uniw.		0.511
		(0.065)
matka ukończyła uniw.		0.386
		(0.073)
wysoki dochód		0.326
		(0.054)
dom		0.099
		(0.059)
stała	13.804	9.416
	(0.072)	(0.146)
N	4,739	4,739
R^2	0.010	0.261

(a). Zinterpretuj oszacowanie parametru przy zmiennej *dystans*. *(1 p.)*
Rozwiązanie: $\hat{\beta}_{dystans} = -0.070$: Zmniejszenie dystansu do uniwersytetu o 10 mil, jest związane ze wzrostem lat edukacji średnio o 0.07 roku, przy innych czynnikach niezmienionych.

(b). Zinterpretuj R^2 dla modelu (2). *(1 p.)*
Rozwiązanie: $R^2 = 0.01$ - model (1) wyjaśnia 10% zmienności lat edukacji.

(c). Który model, (1) czy (2), jest lepiej dopasowany do danych? - użyj odpowiedniej miary. *(2 p.)*
Rozwiązanie: Należy skorzystać ze skorygowanego R^2, czyli właściwej miary do porównania dopasowania modeli.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$$
$$\bar{R}_1^2 = 1 - \frac{4739-1}{4739-2-1}(1-0.01) = 0.0095$$
$$\bar{R}_2^2 = 1 - \frac{4739-1}{4739-7-1}(1-0.261) = 0.2606$$

Model (2) charakteryzuje się wyższym skorygowanym R^2, więc jest lepiej dopasowany.

(d). Przetestuj istotność oszacowania przy zmiennej *czesne* w modelu (1) zakładając $\alpha = 0.05$. *(2 p.)*
Rozwiązanie: Hipoteza zerowa i alternatywna:

$$H_0 : \beta_{czesne} = 0$$
$$H_1 : \beta_{czesne} \neq 0$$

Statystyka testowa:

$$t = \frac{0.160}{0.077} = 2.078$$

Wartość krytyczna: $t^*_{1-\frac{0.05}{2}} = t^*_{0.975} = 1.96$. Zatem $|2.078| > |1.96|$, tak więc istnieją podstawy do odrzucenia hipotezy zerowej. Zmienna *dystans* jest istotna statystycznie przy $\alpha = 0.05$.

(e). Zbuduj 99% przedział ufności dla zmiennej *czesne*. Oceń jej istotność statystyczną na nowym poziomie istotności. *(1 p.)*
Rozwiązanie: $\hat{\beta}_{czesne} \pm se(\hat{\beta}_{czesne}) \times t^*_{1-\frac{0.01}{2}} = 0.160 \pm 0.077 \times 2.57 \Rightarrow \beta_{czesne} \in (-0.03789;\ 0.35789)$ przy $\alpha = 0.01$. Tak więc przy $\alpha = 0.01$, zmienna *czesne* **nie jest** istotna statystycznie.

(f). Czy zmienne dodane w modelu (2) w stosunku do modelu (1) są łącznie istotne statystycznie? Przeprowadź odpowiedni test, przyjmij $\alpha = 0.05$ *(3 p.)*
Rozwiązanie: Formułujemy hipotezy zerową i alternatywną:

$$H_0 : \begin{cases} \beta_{wyniktestu} = 0 \\ \beta_{ojciecukonczyuniw.} = 0 \\ \beta_{matkaukonczyauniw.} = 0 \\ \beta_{wysokidochd} = 0 \\ \beta_{dom} = 0 \end{cases}$$

$$H_1 : \beta_{wyniktestu} \neq 0 \vee \beta_{ojciecukonczyuniw.} \neq 0 \vee \beta_{matkaukonczyauniw.} \neq 0 \vee \beta_{wysokidochd} \neq 0 \vee \beta_{dom} \neq 0$$

Budujemy statystykę testową:

$$F = \frac{\frac{0.261-0.01}{5}}{\frac{1-0.261}{4739-7-1}} \sim F(1-0.05, 5, 4739-7-1)$$
$$F = \frac{\frac{0.261-0.01}{5}}{\frac{1-0.261}{4739-7-1}} = 321.38$$

Szukamy wartości krytycznej: $F^*_{(1-0.05,5,4739-7-1)} = F^*_{(0.95,5,\infty)} = 2.21$
Weryfikujemy hipotezę zerową: $F = 321.38 > F^*_{(0.95,5,\infty)} = 2.21$. Zatem zmienne dodane do modelu (5) są łącznie istotne statystycznie.

Zadanie 2. *(10 p.)* Oszacowano modele objaśniające logarytm rocznego wynagrodzenia, w zależności od lat edukacji (*edukacja*), doświadczenia (*doświadczenie*, w latach), płci (*kobieta*=1 jeśli dana osoba jest kobietą, 0 wpp.), członkostwa w związku zawodowym (*związek*=1 jeśli dana osoba jest członkiem zw. zaw.; 0 wpp.) oraz rodzaju pracy (*biurowy*=1 jeśli dana osoba wykonuje pracę biurową, 0 wwp. - pracuje fizycznie). Na podstawie oszacowań trzech modeli przedstawionych w Tabeli 2 odpowiedz na pytania.

Tabela 2: Model płac rocznych

	Zmienna zależna:		
	log(płaca)		
	(1)	(2)	(3)
edukacja	0.082	0.082	0.062
	(0.006)	(0.005)	(0.007)
doświadczenie	0.007	0.031	0.007
	(0.001)	(0.007)	(0.001)
doświadczenie^2		−0.0005	
		(0.0001)	
kobieta	−0.432	−0.429	−0.463
	(0.046)	(0.046)	(0.046)
związek	0.099	0.093	0.093
	(0.031)	(0.031)	
biurowy			0.148
			(0.039)
biurowy × związek			−0.097*
			(0.054)
stała	5.753	5.512	5.997
	(0.089)	(0.110)	(0.092)
N	595	595	595
R^2	0.362	0.376	0.367
Statystyka F	83.628	70.942	68.220
RESET	0.46353	0.17218	0.50669
p-value	0.6293	0.8419	0.6028

(a). Czy zmienne w modelu (1) w Tabeli 2 są **łącznie** istotne statystycznie? Przyjmij $\alpha = 0.05$. *(1 p.)*
Rozwiązanie: Hipoteza zerowa mówi o łącznej istotności modelu (1). Statystyka testowa jest podana w tabeli: $F = 83.623$. Wartość krytyczna: $F^*_{(1-0.05,4,595-4-1)} = F^*_{(0.95,4,\infty)} = 2.37$. Statystyka testowa jest wyższa od wartości krytycznej, zatem istnieją podstawy do odrzucenia hipotezy zerowej. Zmienne w modelu (1) są łącznie istotne statystycznie.

(b). Zinterpretuj oszacowanie przy zmiennej *edukacja*. *(1 p.)*
Rozwiązanie: $\hat{\beta}_{edukacja} = 0.082$: wraz ze zwiększeniem lat edukacji o 1 rok, płaca rośnie średnio o 8.2% przy innych czynnikach niezmienionych.

(c). Jak średnio zmienią się zarobki przy wzroście doświadczenia o jeden rok, dla osoby z 5 latami doświadczenia? *(2 p.)*

Rozwiązanie: Korzystamy z wyników modelu (2). Należy wziąć pod uwagę, że w modelu jest uwzględniony również kwadrat zmiennej doświadczenie:

$$\frac{\partial \log(placa)}{\partial\, doswiadczenie} = 0.031 - 0.001 doswiadczenie$$
$$\frac{\partial \log(placa)}{\partial\, doswiadczenie = 25} = 0.031 - 0.001 \times 5 = 0.026$$

Przy 5-ciu latach doświadczenia, kolejny rok doświadczenia jest związany ze wzrostem rocznej płacy o średnio 2.6%, przy innych czynnikach niezmienionych,

(d). W ostatnim panelu Tabeli 2 podano statystyki testu RESET oraz związanie z nimi *p-value*. Opisz krótko ten test. Które założenie KMRL pozwala on sprawdzić? Na podstawie wyników testu RESET skomentuj, czy są argumenty aby porzucić model (1) na rzecz modelu (2) lub (3)? *(2 p.)*

Rozwiązanie: Test RESET pozwala przetestować założenie KMRL o liniowości modelu. Hipoteza zerowa testu mówi o tym, że model jest liniowy, zaś hipoteza alternatywna mówi o tym, że model nie jest liniowy. Wyniki testu RESET potwierdzają, że wszystkie modele są liniowe (*p-value* znacznie wyższe od np. 0.05). Pod względem formy funkcyjnej, nie ma powodu aby porzucać model (1).

(e). Na podstawie oszacowań modelu (3) odpowiedz: czy pracownicy biurowi, którzy nie są członkami związków zawodowych zarabiają więcej niż pracowniczy fizyczni, będący członkami związków zawodowych? *(2 p.)*

Rozwiązanie:

$$\mathbb{E}[\log(placa) \mid biurowy = 1,\ zwiazek = 0] = 0.093 \times 0 + 0.148 \times 1 - 0.097 \times 0 = 0.148$$
$$\mathbb{E}[\log(placa) \mid biurowy = 0,\ zwiazek = 1] = 0.093 \times 1 + 0.148 \times 0 - 0.097 \times 0 = 0.093$$

Pracownicy biurowi, którzy nie są członkami związków zawodowych zarabiają więcej niż pracowniczy fizyczni, będący członkami związków zawodowych o 5.5%.

(f). W modelu pominięto zmienną *afam* (=1 gdy dana osoba jest Afroamerykaninem, 0 wpp.). Według danych, osoby czarnoskóre są przeciętnie gorzej wykształcone (korelacja między *afam* i *edukacja* wynosi -0.1195) oraz zarabiają gorzej od osób białych (korelacja między *afam* i *log(wage)* wynosi -0.2229). Jaki będzie kierunek obciążenia $\hat{\beta}_{edukacja}$? *(2 p.)*

Rozwiązanie: Należy skorzystać ze wzoru na obciążenie estymatora przy zmiennej *edukacja*, wywołanego pominięciem ważnej zmiennej:

$$\mathbb{E}[\hat{\beta}_{wyksztalcenie} \mid \boldsymbol{X}] - \beta_{wyksztalcenie} = \underbrace{\gamma}_{<0} \frac{\sigma_{afam}}{\sigma_{wyksztalcenie}} \underbrace{\rho_{afam,rasa}}_{<0} > 0 \Rightarrow \hat{\beta}_{wyksztalcenie} > \beta_{wyksztalcenie}$$

Wiadomo, że $\rho_{wyksztalcenie,afam} < 0$ (korelacja między rasą i wykształceniem jest ujemna) oraz że $\gamma < 0$ (γ reprezentuje relację między rasą i wykształceniem, ich korelacja jest ujemna). Zatem obciążenie będzie dodatnie, co oznacza, że prawdziwy parametr $\beta_{wyksztacenie}$ będzie niższy od oszacowania, $\hat{\beta}_{wyksztacenie}$.

Zadanie 3. *(8 p.)* W Tabeli 3 podano oszacowania modelu objaśniającego całkowity koszt w przedsiębiorstwach wytwarzających energię elektryczną. Jako zmienne objaśniające uwzględniono *kapitał* (indeks ceny kapitału), *płaca* (stawka płacy), *paliwo* (cena paliwa), oraz wolumen produkcji (zmienna *produkcja*). W Tabeli 3 podano również zwykłe błędy standardowe oraz wyniki testów diagnostycznych. Na podstawie wyników odpowiedz na pytania. Przyjmij $\alpha = 0.05$.

Tabela 3: Oszacowanie funkcji kosztu.

	Zmienna zależna: log(cost)		
	współczynniki	*bł. std*	*VIF*
log(kapitał)	0.054	(0.091)	1.040402
log(płaca)	0.243***	(0.089)	1.167211
log(paliwo)	0.663***	(0.050)	1.114998
log(produkcja)	0.391***	(0.037)	26.994142
(log(produkcja)^2	0.062***	(0.005)	26.867642
Stała	−7.044***	(0.880)	–
N	123		
R^2	0.992		
Statystyka F	2,830.934*** (df = 5; 117)		
Breusch-Pagan test	49.972	*p-value*: $1.404e^{-9}$	
White test	32.35	*p-value*: 0	
Jarque-Bera test	2.9098	*p-value* 0.2334	
Note:	*p<0.1; **p<0.05; ***p<0.01		

(a). Zinterpretuj współczynnik przy zmiennej $\log(paliwo)$. *(1 p.)*
Rozwiązanie: $\hat{\beta}_{paliwo} = 0.663$ - wzrost ceny paliwa o 1% jest związany ze wzrostem kosztów średnio o 0.663%, przy innych czynnikach niezmienionych.

(b). Czy w modelu występuje heteroskedastyczność? - odpowiedz na podstawie wyników. Czy badacz korzysta z odpowiednich błędów standardowych? Opisz zwięźle na czym polega heteroskedastyczność i powiedz, które z założeń KMRL ona łamie. *(3 p.)*
Rozwiązanie: Heteroskedastyczność to zjawisko które polega na tym, że wariancja składnika losowego modelu ekonometrycznego nie jest stała. Gdy występuje, złamane jest założenie KMRL o sferyczności wariancji składnika losowego. Obecność heteroskedastyczności w analizowanym modelu można stwierdzić na podstawie testów White'a oraz Breusha-Pagana - w obu przypadkach *p-value* jest bardzo niskie (niższe niż np. 0.05). Oznacza to że w przypadku obu testów, istnieją podstawy do odrzucenia hipotezy zerowej, mówiącej o homoskedastyczności składnika losowego. Należy uznać, że składnik losowy jest heteroskedastyczny. Badacz korzysta ze zwykłych błędów standardowych, a powinien skorzystać z odpornych błędów standardowych.

(c). Czy założenie o rozkładzie normalnym składnika losowego jest spełnione przez powyższy model? *(1 p.)*
Rozwiązanie: Aby przetestować to założenie, należy skorzystać z testu Jarque-Bery. Wynik tego testu wskazuje że nie ma podstaw do odrzucenia hipotezy zerowej, mówiącej o tym że składnik losowy ma rozkład normalny (*p-value*=0.2334 jest wyższe niższe niż 0.05).

(d). Czy w modelu występuje nadmierna współliniowość? Opisz na czym polega to zjawisko, oraz czym może skutkować w modelu ekonometrycznym. *(2 p.)*
Rozwiązanie: Współliniowość polega na występowaniu silnej korelacji między zmiennymi objaśniającymi w modelu ekonometrycznym. Może to skutkować np. sztucznym podwyższeniem wariancji osza-

cowań (estymatora) i w konsekwencji mylnym uznaniu zmiennych za nieistotne. Nadmierna współliniowość występuje gdy $VIF > 10$. W analizowanym modelu występuje ponieważ VIF przy zmiennych $\log(produkcja)$ oraz $\log(produkcja)^2$. Mimo nadmiernej współliniowości, jej konsekwencje się nie zmaterializowały - zmiennej są silnie istotne statystycznie.

(e). Na podstawie modelu można przetestować złożoną hipotezę o tym, że: $\beta_{placa} + \beta_{paliwo} + \beta_{kapita} = 1$, $\beta_{placa} = 0.75$ oraz $\beta_{paliwo} = 0.25$. Zapisz tę hipotezę korzystając z zapisu macierzowego. *(1 p.)*
Rozwiązanie:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{kapital} \\ \beta_{placa} \\ \beta_{paliwo} \\ \beta_{produkcja} \\ \beta_{produkcja^2} \end{bmatrix} = \begin{bmatrix} 1 \\ 0.75 \\ 0.25 \end{bmatrix}$$

Zadanie 4. *(4 p.)* Estymator parametrów modelu ekonometrycznego, który spełnia założenia Klasycznego Modelu Regresji Liniowej, jest nieobciążony.

(a). Co to znaczy że estymator jest nieobciążony? *(1 p.)*

$$\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \boldsymbol{\beta}$$

Warunkowa wartość oczekiwana estymatora jest równa prawdziwym wartościom parametrów.

(b). Udowodnij to twierdzenie. Podaj niezbędne założenia. *(3 p.)*

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ - model jest liniowy
- Zmienne losowe $\{(y_1, X_1), ..., (y_i, y_i), ..., (y_n, y_n)\}$ są niezależne oraz wylosowane z tego samego rozkładu (*independently and identically distributed - iid.*)
- $rz[\boldsymbol{X}_{n\times k}] = k$ - rząd kolumnowy $\boldsymbol{X}$ jest pełny
- $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$ - wartość oczekiwana składnika losowego jest równa 0
- Estymator uzyskany MNK $\hat{\boldsymbol{\beta}}$ wektora parametrów $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- Pokażmy, że estymator $\hat{\boldsymbol{\beta}}$ jest nieobciążony, pw. że wymienione założenia są spełnione:

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}]}_{\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{X}] + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}]}_{=0} \\ &= \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Appendix Wybrane wartości z rozkładów t i F.

Wybrane wartości rozkładu t

$t(0.95, \infty)$	1.64
$t(0.975, \infty)$	1.96
$t(0.99, \infty)$	2.32
$t(0.995, \infty)$	2.57

Wybrane wartości rozkładu F

$F(0.95, 2, \infty)$	2.99
$F(0.95, 4, \infty)$	2.37
$F(0.95, 5, \infty)$	2.21
$F(0.975, 4, \infty)$	2.78
$F(0.975, 5, \infty)$	2.57