Kolokwium z Ekonometrii - Rozwiązania

Styczeń 2023

1. *(18 p.)* Poszczególne pytania w tym zadaniu odnoszą się do Tabeli 1 przedstawiającej oszacowania, obliczone na danych CPS (*Current Population Survey*). Dane opisują 7440 pracowników (pracują cały rok na pełen etat). Najwyższym osiągnięciem dla każdego z pracowników było albo ukończenie liceum lub dyplom licencjacki. Wiek pracowników zawiera się pomiędzy 25 a 34. Zbiór danych zawiera także informacje o regionie kraju w którym dana osoba żyje, statusie matrymonialnym oraz o liczbie dzieci:

   - AHE = średnia wynagrodzenie na godzinę (w dolarach z 2012 roku)
   - College = zm. binarna (1 jeśli dyplom lic., 0 jeśli liceum)
   - Female = zm. binarna (1 jeśli kobieta, 0 jeśli mężczyzna)
   - Age = wiek (w latach)
   - Ntheast = zm. binarna (1 jeśli region to Northeast, 0 w przeciwnym przypadku)
   - Midwest = zm. binarna (1 jeśli region to Midwest, 0 w przeciwnym przypadku)
   - South = zm. binarna (1 jeśli region to South, 0 w przeciwnym przypadku)
   - West = zm. binarna (1 jeśli region to West, 0 w przeciwnym przypadku)

Tabela 1: Zarobki w zależności od wykształcenia, płci i innych charakterystyk.

	zm. zależna: AHE			zm. zależna: log(AHE)	
	(1)	(2)	(3)	(4)	(5)
College (X1)	8.31	8.32	8.34	0.44	0.40
	(0.23)	(0.22)	(0.22)	(0.01)	(0.01)
Female (X2)	-3.85	-3.81	-3.80	-0.19	-0.24
	(0.23)	(0.22)	(0.22)	(0.01)	(0.17)
Age (X3)		0.51	0.52	0.10	0.10
		(0.04)	(0.04)	(0.04)	(0.04)
Age2 (X4)				-0.001	-0.001
				(-0.0008)	(-0.0008)
Female×College (X5)					0.09
					(0.02)
Northeast (X6)			0.18		
			(0.36)		
Midwest (X7)			-1.23		
			(0.31)		
South (X8)			-0.43		
			(0.30)		
stała	17.02	1.87	2.05	-0.79	0.80
	(0.17)	(1.18)	(1.18)	(0.67)	(0.67)
R^2	0.162	0.180	0.182	0.197	0.198
n	7440	7440	7440	7440	7440

W nawiasach () podano odchylenia standardowe.

(a) Policz skorygowany R^2 dla regresji (1). Wyjaśnij różnice między zwykłym i skorygowanym R^2.

<u>Rozwiązanie:</u> Skorygowany R^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times (1 - R^2) = 1 - \frac{7440-1}{7440-2-1} \times (1 - 0.162) = 0.1617$$

Zwykły R^2 zawsze zwiększa się, gdy do modelu zostaną dodane nowe zmienne, nawet w przypadku gdy są one dodane niepotrzebnie. Skorygowany R^2 uwzględnia to, korygując zwykły R^2 o liczbę zmiennych w modelu (stopni swobody). Dlatego nigdy nie jest wyższy niż zwykły R^2. Warto pamiętać, że skorygowany R^2 nie ma interpretacji, takiej jak zwykły R^2!

Odpowiedz, używając wyników z kolumny (1):

(b) Czy pracownicy z dyplomami licencjackimi zarabiają więcej niż pracownicy z ukończonym liceum? O ile więcej? Czy ta różnica jest statystycznie istotna?

Rozwiązanie: Pracownicy z dyplomami licencjackimi zarabiają średnio o $8,31$\$ na godzinę więcej niż pracownicy bez dyplomów licencjackich, przy innych czynnikach niezmienionych. Współczynnik przy zmiennej $College$ jest istotny statystycznie, ponieważ statystyka $t_{College}$ wynosi $\frac{8.31}{0.23} = 36.13$ i jest wyższa niż wartość krytyczna (1,96) przy $\alpha = 0.05$.

(c) Czy mężczyźni zarabiają więcej niż kobiety? O ile? Czy ta różnica jest istotna statystycznie?

Rozwiązanie: Mężczyźni zarabiają średnio więcej o 3.85\$ na godzinę niż kobiety, przy innych czynnikach niezmienionych. Współczynnik przy zmiennej $Female$ jest istotny statystycznie, ponieważ statystyka $| t_{Female} |=| \frac{-3.85}{0.23} |= 16.73 > 1.96$.

Odpowiedz, używając oszacowań z kolumny (2):

(d) Czy wiek jest istotną determinantą zarobków? Zbuduj 95% przedział ufności i skomentuj.

Rozwiązanie: Przedział ufności dla β_{age}:

$$
\begin{aligned}
P(\hat{\beta}_{age} - t_{1-\frac{0.05}{2}} \times se(\hat{\beta}_{age}) < \beta_{age} < \hat{\beta}_{age} + t_{1-\frac{0.05}{2}} \times se(\hat{\beta}_{age})) &= 0.95 \\
P(0.51 - 1.96 \times 0.04) < \beta_{age} < 0.51 + 1.96 \times 0.04) &= 0.95 \\
P(.4316 < \beta_{age} < .5884) &= 0.95
\end{aligned}
$$

0 nie zawiera się w 95-cio procentowym przedziale ufności, więc przy $\alpha = 0.05$ wiek jest statystycznie istotną determinantą zarobków.

Odpowiedz, używając oszacowań z kolumny (3):

(e) Czy różnice międzyregionalne są łącznie istotne? Przetestuj odpowiednią hipotezę i odpowiedz. Wartość krytyczna z odpowiedniego rozkładu dla $\alpha = 0.05$ wynosi 0.485844.

Rozwiązanie: Należy przeprowadzić test łącznej istotności parametrów $\beta_{Northeast}$, $\beta_{Midwest}$, β_{South}. W tym celu skorzystamy ze statystyki F. Hipotezy zerowa i alternatywna:

$$
\begin{aligned}
H_0 &: \beta_{Northeast} = \beta_{Midwest} = \beta_{South} = 0 \\
H_1 &: \beta_{Northeast} \neq 0 \vee \beta_{Midwest} \neq 0 \vee \beta_{South} \neq 0
\end{aligned}
$$

Następnie budujemy statystykę F; model (3) - bez restrykcji, model (2) - z restrykcjami.

$$
F = \frac{\frac{0.182-0.180}{3}}{\frac{1-0.182}{7440-6-1}} = 6.057
$$

Statystyka testowa jest wyższa od wartości krytycznej, podanej w treści zadania. Tak więc istnieją podstawy do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Zatem różnice międzyregionalne są łącznie istotne statystycznie.

(f) Dlaczego zmienna *West* została pominięta w regresji? Co by się stało, gdyby została umieszczona w równaniu?

Rozwiązanie: Zmienna *West* została pominięta ponieważ jest ona jedną z czterech zmiennych binarnych, opisujących regiony. Jedna z niuch musi być wybrana jako poziom odniesienia. Gdyby spróbowano umieścić w modelu wszystkie cztery zmienne binarne regionalne, estymator MNK nie mógłby istnieć gdyż macierz zmiennych nie miałaby pełnego rzędu kolumnowego; inaczej, istniałaby doskonała współliniowość.

Odpowiedz, używając oszacowań z kolumny (4):

(g) Jeśli wiek wzrasta z 25 do 26 przy innych czynnikach niezmienionych, jak zmienią się zarobki? A jak w przypadku gdy wiek wzrasta z 33 do 34? Użyj odpowiedniego przybliżenia.

Rozwiązanie: Możemy zastosować pochodną modelu (4) względem wieku, i w ten sposób uzyskać przybliżenie efektu jednostkowej zmiany wieku na zarobki, w danym punkcie:

$$\frac{\partial \log(AHE)}{age} = 0.10 - 2 \times 0.001 age$$

Przy $age = 25$, zwiększenie wieku o 1 (do 26), zwiększa zarobki o 5% ($0.10 - 2 \times 0.001 \times 25 = 0.05$). Przy $age = 33$, zwiększenie wieku o 1 (do 34), zwiększa zarobki o 3.4% ($0.10 - 2 \times 0.001 \times 33 = 0.034$).

(h) Przyjmij, że wpływ wieku na zarobki może być różny dla absolwentów liceum i dla absolwentów uniwersytetu. Zmodyfikuj równanie regresji w taki sposób, aby uchwyciło te różnice.

Rozwiązanie: Należy do modelu dodać interakcję między zmienną opisującą wiek i wykształcenie: $College \times age$.

Odpowiedz, używając oszacowań z kolumny (5):

(i) Co mierzy współczynnik przy interakcji? Zinterpretuj go.

Rozwiązanie: Interakcja ta uwzględnia to, że relacja między zarobkami a płcią zależy od wykształcenia. Przy interpretowaniu modelu z interakcją należy pamiętać, że trzeba uwzględniać zarówno efekty główne (β_{Female}, $\beta_{College}$) jaki i interakcję ($\beta_{Female \times College}$). Przykład: jak różnią się zarobki między kobietami z ukończonymi studiami licencjackimi i bez?:

$$(\square) \text{ jeśli } Female = 1 \text{ oraz } College = 0 = 0.40 \times 0 - 0.24 \times 1 + 0.09 \times 0 = -0.24$$
$$(\triangle) \text{ jeśli } Female = 1 \text{ oraz } College = 1 = 0.40 \times 1 - 0.24 \times 1 + 0.09 \times 1 = 0.25$$

gdy obliczymy różnicę $\triangle - \square$ otrzymamy 0.49. Oznacza to że średnio, przy innych czynnikach niezmienionych, kobiety z ukończonymi studiami licencjackimi zarabiają więcej o 49% niż kobiety bez wykształcenia licencjackiego. Analogicznie możemy wyznaczyć różnicę w zarobkach między mężczyznami i kobietami z dyplomem licencjackim:

$$(\square) \text{ jeśli } Female = 1 \text{ oraz } College = 1 = 0.40 \times 1 - 0.24 \times 1 + 0.09 \times 1 = -0.15$$
$$(\triangle) \text{ jeśli } Female = 0 \text{ oraz } College = 1 = 0.40 \times 1 - 0.24 \times 0 + 0.09 \times 0 = 0.4$$

Tak więc pośród osób z dyplomem licencjackim, mężczyźni zarabiają więcej średnio o 55%, przy innych czynnikach niezmienionych.

2. *(12 p.)* W tym zadaniu, pytania odnoszą się do oszacowań luki płacowej wysoko wykwalifikowanych imigrantów, uzyskanych na danych z amerykańskiego badania *Survey of Income and Program Participation (SIPP)*, z lat 2008-2013:

- ln_wage - logarytm zarobków
- immigr - zm. binarna (1 jeśli imigrant, 0 jeśli nie)
- age - wiek (w latach)
- age2 - wiek do kwadratu
- female - zm. binarna (1 jeśli kobieta, 0 jeśli mężczyzna)
- black - zm. binarna (1 jeśli rasa czarna, 0 w przeciwnym przypadku)
- asian - zm. binarna (1 jeśli rasa żółta, 0 w przeciwnym przypadku)

- cognitive - zm. binarna (1 jeśli pracownik umysłowy, 0 jeśli pracownik fizyczny)

Wszystkie osoby w próbie są wysoko wykwalifikowane, tzn. mają co najmniej dyplom licencjacki lub wyższy.

Tabela 2: Zarobki w zależności od statusu imigracyjnego i innych charakterystyk

	(1) oszacowania na całej próbie					(2) oszacowania na podpróbach	
	oszacowania	bł. std.	stat. t	p-value	VIF	umysłowi	fizyczni
immigr	-0.09856	0.004359	-22.61	2×10^{-16}	1.406877	0.0262	-0.178
age	0.1112	0.001333	83.41	2×10^{-16}	89.434584	0.0944	0.09735
age2	-0.0001	0.000016	-77.12	2×10^{-16}	89.417841	-0.0009	-0.0011
female	-0.4289	0.002658	-161.39	2×10^{-16}	1.005427	-0.4077	-0.4157
black	-0.2053	0.005058	-40.59	2×10^{-16}	1.020609	-0.1456	-0.1319
asian	0.1664	0.005723	29.08	2×10^{-16}	1.415301	0.1076	0.0708
stała	6.151	0.02732	225.13	2×10^{-16}		6.645	6.166
SSR	224586					109858	84057
R^2	0.1019					0.1137	0.0882
n	358688					223083	135605
Statystyki:							
Statystyka testu White'a:		81.89	p-value	2.2×10^{-16}			
Statystyka testu RESET:		445.44	p-value	2.2×10^{-16}			
Statystyka testu Jarque-Bery:		499017	p-value	2.2×10^{-16}			

Zmienną zależną w oszacowanych modelach jest oczywiście ln_wage.

Odpowiedz na pytania korzystając z wyników z (1) części Tabeli 2

(a) Zinterpretuj współczynnik przy zmiennej $immigr$. Wyjaśnij co to oznacza w kategoriach ekonomicznych.
Rozwiązanie: Imigranci zarabiają średnio mniej o 9% od rezydentów, przy innych czynnikach niezmienionych.

(b) Czy współliniowość jest problemem w analizowanym modelu? Uzasadnij, opisz jej konsekwencje oraz podaj ewentualne rozwiązanie tego problemu.
Rozwiązanie: W tym modelu współliniowość występuje, ponieważ wartość VIF znacznie przekracza 10 w przypadku age oraz $age2$. Wyeliminowanie zmiennej $age2$ powinno zmniejszyć współliniowość. Współliniowość nie wpływa na własności estymatora, natomiast może zniekształcić odchylenia standardowe, co może doprowadzić do błędnej oceny istotności statystycznej.

(c) Czy w modelu występuje heteroskedastyczność? Uzasadnij. Jeśli tak, to co powinien zrobić ekonometryk? Jakie problemy powoduje heteroskedastyczność? Rozwiązanie: Tak, w modelu występuje problem heteroskedastyczności. Wskazuje na to niskie p-value w teście White'a (niższe niż 0.05). Tak więc istnieją podstawy do odrzucenia hipotezy zerowej o homoskedastyczności w teście White'a. W przypadku heteroskedastyczność, oszacowanie wariancji estymatora jest nieprawidłowe, zatem i odchylenia standardowe są niewłaściwe. Może to prowadzić do niewłaściwej oceny istotności statystycznej zmiennych. W takiej sytuacji warto zastosować estymator wariancji White'a (potocznie odporne błędy standardowe).

(d) Czy zastosowana forma funkcyjna jest poprawna? Uzasadnij.
Rozwiązanie: P-value dla testu RESET jest niższe od 0.05, więc są podstawy do odrzucenia hipotezy zerowej o liniowej formie funkcyjnej. Wynik testu RESET sugeruje, że forma funkcyjna modelu jest nieprawidłowa.

(e) Czy składnik losowy ma rozkład normalny w rozważanym modelu? Uzasadnij. Co się dzieje gdy tak nie jest?
Rozwiązanie: P-value dla testu Jarque-Bery jest niższe od 0.05, zatem istnieją podstawy do odrzucenia

hipotezy zerowej, mówiącej o tym że składnik losowy w modelu ma rozkład normalny. Taki wynik sugeruje niespełnienie założenia KMRL. Nie narusza to własności estymatora MNK, ale rozkłady statystyk testowych są nieznane przy założeniu że hipoteza zerowa jest prawdziwa, co utrudnia przeprowadzanie testów, szczególnie w małych próbach.

(f) W zbiorze danych zawarta jest zmienna która odróżnia pracowników umysłowych i fizycznych (cognitive). W dwóch ostatnich kolumnach, zawarto oszacowania modelu, odpowiednio dla pracowników umysłowych i fizycznych. Czy istnieje strukturalna różnica między oszacowanymi parametrami w tych dwóch regresjach? Przeprowadź odpowiedni test i odpowiedz. (*Wykorzystaj również wyniki z (2) części Tabeli 2*) Wartość krytyczna z rozkładu F wynosi 0.7633769.

Rozwiązanie: Należy przeprowadzić test Chowa dla dwóch podprób (pracowników umysłowych i fizycznych):

$$H_0 : {}_{umyslowi} = {}_{fizyczni}$$
$$H_1 : {}_{umyslowi} \neq {}_{fizyczni}$$

hipoteza zerowa mówi o tym, że wektor parametrów oszacowany na podpróbie zawierającej dane o pracownikach umysłowych jest taki sam jak wektor parametrów oszacowany na podpróbie zawierającej dane o pracownikach fizycznych. Obliczamy statystykę testową:

$$F = \frac{\frac{224586-109858-84057}{6+1}}{\frac{109858+84057}{358688-2(6+1)}} = 8104$$

Statystyka testowa jest wyższa niż wartość krytyczna, więc mamy podstawy do odrzucenia hipotezy zerowej. Istnieje zatem różnica w parametrach oszacowanym na wyróżnionych podpróbach. W takiej sytuacji można oszacować dwa oddzielne modele lub dodać zmienną binarną do głównej regresji.

3. *(3 p.)* Ekonometryk w ramach swojego projektu chce zbadać jak na stopę zabójstw (per capita) wpływają zmiany w nakładach na policję (per capita) w powiatach w Polsce. Można się spodziewać, że przekazanie większych środków policji, powinno przyczynić się do spadku odsetka zabójstw w danym powiecie. Dodatkową zmienną którą warto uwzględnić w regresji to czy gangi są obecne w danym powiecie. Tak więc nasz Ekonometryk chciałby oszacować następujące równanie:

$$zabojstwa\_per\_capita = \beta_0 + \beta_1 fiansowanie\_policji + \beta_2 gang + \varepsilon \tag{1}$$

Jednak nasz Ekonometryk zapomniał zebrać dane o obecności gangów w powiatach, więc w praktyce może jedynie oszacować następujące równanie:

$$zabjstwa\_per\_capita = \beta_0 + \beta_1 fiansowanie\_policji + \varepsilon \tag{2}$$

(a) Ekonometryk popełnił tutaj błąd zmiennej pominiętej. Wyjaśnij krótko jakie konsekwencje niesie on dla oszacowania parametru β_1 z równania (2)?

Rozwiązanie: Estymator MNK w sytuacji zmiennej pominiętej będzie obciążony (złamane założenie o zerowej warunkowej wartości oczekiwanej składnika losowego). Objawi się to przeszacowaniem/niedoszacowaniem parametru β_1.

(b) Jeśli Ekonometryk oszacuje równanie (2), to czy $\hat{\beta_1}$ będzie przeszacowywało czy niedoszacowywało prawdziwy parametr β_1 z równianina (1)? Uzasadnij swoją odpowiedź odpowiednim rozumowaniem.

Rozwiązanie: Aby odpowiedzieć na to pytanie, należy skorzystać ze wzoru na obciążenie estymatora i określić jego znak:

$$\mathbb{E}(\hat{\beta}_1) - \beta_1 = \beta_2 \times \frac{\text{Cov}(fiansowanie\_policji, gang)}{\text{Var}(fiansowanie\_policji)}$$

Ostateczna odpowiedź zależy od tego jakie relacje zachodzą między zmiennymi w modelu. Po pierwsze przyjmijmy, że im większe finansowanie policji w powiecie, tym mniej gangów, czyli:

$$\text{Cov}(fiansowanie\_policji, gang) < 0$$

Po drugie jeżeli w danym powiecie będzie więcej gangów, tym więcej będzie zabójstw, czyli $\beta_2 > 0$. W takiej sytuacji, obciążenie będzie ujemne; oszacowanie β_1 będzie niedoszacowane; $\mathbb{E}(\hat{\beta}_1) < \beta_1$.

4. *(7 p.)* Klasyczny model regresji liniowej - podaj założenia. Opisz co mówi tw. Gaussa-Markova.
Rozwiązanie: Trywialne. Odpowiedź w np. slajdach z wykładu, albo dowolnym podręczniku.