Zaawansowana Ekonometria

Paweł Struski

Uniwersytet Warszawski

16 lutego 2026

Plan na dzisiaj

1. Organizacja
   - Literatura
   - Zaliczenie
   - Plan semestru
2. Powtórzenie
   - Regresja Liniowa
   - Metoda Najmniejszych Kwadratów (MNK)
   - R2
   - Testowanie Hipotez
   - Interpretacja Parametrów
3. Zadania

Kontakt i literatura

- kontakt: p.struski@uw.edu.pl

Kontakt i literatura

- kontakt: p.struski@uw.edu.pl
- literatura:
  - Introductory Econometrics: A Modern Approach, Jeffrey Wooldridge
  - Introduction to Econometrics, James H. Stock & Mark W. Watson

Zaliczenie

- Obowiązkowa obecność (max. 3 nieobecności)
- 3 short-testy
  - jeden po każdym z bloków tematycznym 1-3
  - 30 minut
  - konieczne powyżej 50% punktów średnio ze wszystkich short-testów
  - nieobecność skutkuje wynikiem 0% za dany test
  - możliwość poprawy najsłabszego testu w sesji poprawkowej
- 3 lub 4 case studies
  - praca w grupach 3 osobowych
  - zadania empiryczne
  - feedback
- Kolokwium
  - poza zajęciami
  - 3 zadania, 75 minut
- Ocena z ćwiczeń: 30% short testy, 30% case studies, 40% kolokwium

Plan semestru

- Blok 1:
  - 16 lutego: organizacja i powtórzenie
  - 23 lutego: obciążania Lovella, kryteria informacyjne (selekcja modelu)
  - 2 marca: metoda zmiennych instrumentalnych (MZI)
  - 9 marca: lab 1
- Blok 2 - szeregi czasowe:
  - 16 marca: sezonowość, stacjonarność (test 1)
  - 23 marca: DL/ADL
  - 30 marca: ARIMA
  - 13 kwietnia: kointegracja, ECM
  - 20 kwietnia: lab 2
- Blok 3 - zmienne binarne:
  - 27 kwietnia: LMP, Logit, Probit (test 2)
  - 11 maja: lab 3
- Blok 4 - dane panelowe:
  - 18 maja: dane panelowe 1 (test 3)
  - 25 maja: dane panelowe 2
  - 1 czerwca: lab 4
- kolokwium (poza zajęciami - termin zostanie ustalony wkrótce)

Regresja Liniowa

- Niech y będzie n x 1 wektorem obserwacji zmiennych zależnych
- Niech X będzie n x k macierzą zawierającą n obserwacji k zmiennych zależnych
- Niech ϵ będzie n x 1 wektorem czynników losowych
- Niech β będzie n x 1 wektorem nieznanych parametrów modelu

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

W notacji macierzowej:

$$y = X\beta + \epsilon$$

Metoda Najmniejszych Kwadratów (MNK, ang. OLS)

Chcemy uzyskać oszacowanie parametrów modelu t.j. $\hat{\beta}$. Zdefiniujmy wektor reszt:

$$e = y - X\hat{\beta}$$

Suma kwadratów reszt:

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

MNK:

$$\min_{\hat{\beta}} e'e$$

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

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

Twierdzenie Gauss'a-Markov'a

Założenia:

- Liniowa postać modelu: $y = X\beta + \epsilon$
- X jest macierzą pełnego rzędu (z ang. *full rank*). Innymi słowy, kolumny X są liniowo niezależne.
- Zerowa warunkowa wartość oczekiwana składnika losowego: $E[\epsilon|X] = 0$
  - $\implies$ $Cov(\epsilon, X) = 0$ tzn. składnik losowy nie jest skorelowany ze zmiennymi objaśniającymi.
- Wariancja składnika losowego jest sferyczna: $Var(\epsilon|X) = E[\epsilon\epsilon'|X] = \sigma^2 I$
  - $\implies$ homoskedastyczność i.e. wariancja jest stała
  - $\implies$ brak korelacji między indywidualnymi czynnikami losowymi tzn. $\forall_{i \neq j} Cov(\epsilon_i, \epsilon_j) = 0$

Przy powyższych założeniach, estymator MNK jest najlepszym liniowym nieobciążonym estymatorem (z ang. BLUE) tzn.:

- $E[\hat{\beta}] = \beta$
- wariancja estymatora $Var(\hat{\beta})$ jest najmniejsza z możliwych.

Współczynnik determinacji - R^2

- Całkowita suma kwadratów $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - \bar{y})'(y - \bar{y})$
- Wyjaśniona suma kwadratów $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 = (\hat{y}_i - \bar{\hat{y}})'(\hat{y}_i - \bar{\hat{y}})$
- Suma kwadratów reszt $RSS = \sum_{i=1}^{n} e_i^2 = e'e$
- TSS = RSS + ESS

$$R^2 := \frac{ESS}{TSS}$$

Testowanie Hipotez

Aby móc testować hipotezy odnośnie parametrów w naszym modelu, potrzebujemy kolejnego założenia:

- Normalność czynnika losowego: $\epsilon|X \sim \mathcal{N}(0, \sigma^2 I)$

Dzięki temu $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$

- Hipotezy proste -> statystyka t
- Hipotezy łączne -> statystyka F

Testowanie Hipotez - t-test

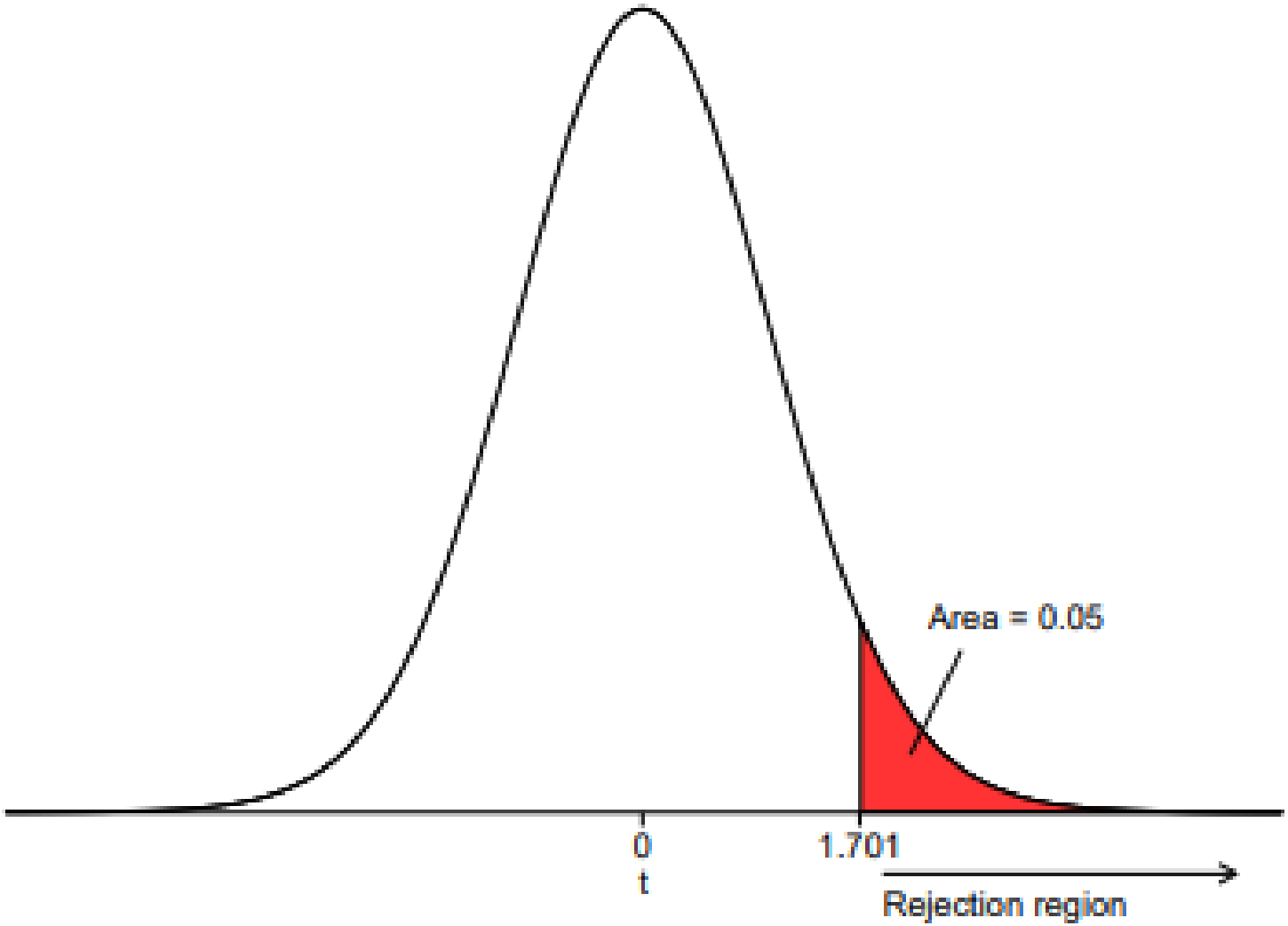
$$\begin{cases} H_0 : \beta_j = \beta_j^* \\ H_1 : \beta_j \neq \beta_j^* \end{cases}$$

$$t = \frac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

Odrzucamy hipotezę na poziomie istotności α (np. 5%) jeśli $t > c$, gdzie c jest wartością krytyczną t.j.
α = Pr(odrzucenia $H_0 | H_0$ jest prawdziwa) = Pr(t>c)

Testowanie Hipotez - t-test c.d.

5% rejection rule for $\mathcal{H}_1 : \beta_j > 0$ (df=28)



Testowanie Hipotez - wartość p

Problem: Przy testowaniu hipotez zawsze musimy wybrać poziom istotności. Niekiedy nasza hipoteza zostanie odrzucona dla jednego poziomu istotności (np. 5%) ale już nie dla innego (np. 10%).
Rozwiązanie: Możemy skorzystać z *wartości p* = najmniejszy poziom istotności, przy którym prawdziwa H_0 byłaby odrzucona.

$$p = Pr(T > |t||H_0)$$

gdzie T jest zmienną losową o znanym rozkładzie, a t jest wartością naszej statystyki testowej, którą otrzymaliśmy z danych.

Odrzucamy hipotezę H_0 jeśli $p < \alpha$, nie odrzucamy jeśli $p > \alpha$.

Testowanie Hipotez - F test

Jeśli chcemy przetestować istotność wielu zmiennych naraz.

$$\begin{cases} H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_q = 0 \\ H_1 : H_0 = falsz \end{cases}$$

Musimy mieć wspólny test. Jak taki stworzyć?

- Porównajmy RSS w modelu bez restrykcji i z restrykcjami H_0.
- Model z restrykcjami wyjaśni mniej wariancji y (będzie miał wyższą SRR).
- Czy związany z tym wzrost RRS uzasadnia odrzucenie hipotezy H_0?

$$F = \frac{(RSS_R - RSS_U)/q}{RSS_U/(n-k-1)} \sim F_{q,n-k-1}$$

gdzie q = liczba restrykcji

Interpretacja Parametrów

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

źródło: Wooldridge, Ch. 2

Zadanie 1A - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$sleep = 3,638.25 - 0.148totwrk - 11.13educ + 2.20age$

- *sleep* - liczba minut snu w tygodniu
- *totwrk* - liczba minut pracy w tygodniu
- *educ* - liczba lat edukacji
- *age* - wiek w latach

Zadanie 1A - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$sleep = 3,638.25 - 0.148totwrk - 11.13educ + 2.20age$

- *sleep* - liczba minut snu w tygodniu
- *totwrk* - liczba minut pracy w tygodniu
- *educ* - liczba lat edukacji
- *age* - wiek w latach

Odpowiedź:

- Dodatkowa minuta pracy w tygodniu powoduje spadek liczby minut snu w tygodniu o 0.148, ceteris paribus.
- Wzrost liczby lat edukacji o 1 rok powoduje spadek liczby minut snu w tygodniu o 11.13 minut, ceteris paribus.
- Wraz ze wzrostem wieku o 1 rok, liczba minut snu w tygodniu rośnie o 2.20 minuty, ceteris paribus.

Zadanie 1B - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$hours = 33 + 45.1log(wage)$

- *hours* - liczba godzin przepracowanych w tygodniu
- *wage* - stawka godzinowa

Zadanie 1B - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$hours = 33 + 45.1log(wage)$

- *hours* - liczba godzin przepracowanych w tygodniu
- *wage* - stawka godzinowa

Odpowiedź:

- Wzrost stawki godzinowej o 1% powoduje wzrost liczby godzin przepracowanych w tygodniu o 0.451 godziny (około pół godziny).

Zadanie 1C - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$log(wage) = 2.78 + 0.094educ$

- *educ* - liczba lat edukacji
- *wage* - stawka godzinowa

Zadanie 1C - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$log(wage) = 2.78 + 0.094educ$

- *educ* - liczba lat edukacji
- *wage* - stawka godzinowa

Odpowiedź:

- Wzrost liczby lat edukacji o 1 powoduje wzrost otrzymywanej stawki godzinowej średnio o 9.4%.

Zadanie 1D - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$log(salary) = 4.822 + 0.257log(sales) - 0.0013cups\_of\_coffee$

- *salary* - miesięczne wynagrodzenie w \$
- *sales* - wartość sprzedanych produktów w danym miesiącu w \$
- *cups_of_coffee* - liczba wypitych kubków kawy w danym miesiącu

Zadanie 1D - Interpretacja parametrów modelu

Podaj interpretację współczynników w poniższym równaniu.
$log(salary) = 4.822 + 0.257log(sales) - 0.0013cups\_of\_coffee$

- *salary* - miesięczne wynagrodzenie w \$
- *sales* - wartość sprzedanych produktów w danym miesiącu w \$
- *cups_of_coffee* - liczba wypitych kubków kawy w danym miesiącu

Odpowiedź:

- Wzrost wartości sprzedanych produktów w miesiącu o 1% jest związany ze wzrostem miesięcznego wynagrodzenia o 0.257 %, ceteris paribus.
- Dodatkowy 1 kubek kawy wypity w danym miesiącu skutkuje spadkiem miesięcznego wynagrodzenia średnio o 0.13%, ceteris paribus.

Zadanie 2 - Wooldridge, Ch. 3, problem 2

Oszacowano następujący model:
$educ = 10.36 - 0.094sibs + 0.131meduc + 0.210feduc$
$N = 722, R2 = 0.214$
gdzie, *educ* to liczba lat edukacji, *sibs* to liczba rodzeństwa, *meduc* to liczba lat edukacji matki, *feduc* to liczba lat edukacji ojca.

1. Czy wpływ zmiennej *sibs* jest zgodny z oczekiwaniami? Uzasadnij. Dla identycznego poziomu *meduc* oraz *feduc*, jaka powinna być wartość zmiennej *sibs* aby oczekiwana liczba lat edukacji (*educ*) spadła o 1 rok? Odpowiedź może nie być liczbą całkowitą.
2. Podaj interpretację współczynnika przy zmiennej *meduc*.
3. Załóżmy, że osoba A nie ma rodzeństwa i oboje rodzice mają po 12 lat spędzonych na nauce. Osoba B również nie ma rodzeństwa, ale jej rodzice mają po 16 lat spędzonych na nauce. Jaka jest różnica w oczekiwanych latach poświęconych na naukę między osobą A i osobą B?

Zadanie 3 - Wooldridge, Ch. 7, problem 3

$$\hat{sat} = \underset{(6.29)}{1,028.10} + \underset{(3.83)}{19.30}\,hsize - \underset{(0.53)}{2.19}\,hsize^2 -$$

$$\underset{(4.29)}{45.09}\,female - \underset{(12.71)}{169.81}\,black + \underset{(18.15)}{62.31}\,female \cdot black$$

gdzie, sat to wynik egzaminu SAT, hsize to wielkość rocznika dla danej osoby w liceum (liczona w setkach osób), female to zmienna zero-jedynkowa (1=kobieta, 0=mężczyzna), black to zmienna zero-jedynkowa (1=osoba czarnoskóra, 0=inne przypadki). W nawiasach podane są błędy standardowe oszacowań. (n=4,137, R2=.0858)

1. Czy zmienna $hsize^2$ powinna być uwzględniona w modelu? Według oszacowanego równania - jaki jest optymalny rozmiar rocznika w liceum?
2. Dla identycznego poziomu *hsize*, jaka jest spodziewana różnica w wyniku testu między kobietami i mężczyznami, którzy nie są czarnoskórzy? Czy różnica jest statystycznie istotna?

Zadanie 3 - Wooldridge, Ch. 7, problem 3 c.d.

$$\hat{sat} = \underset{(6.29)}{1,028.10} + \underset{(3.83)}{19.30}\,hsize - \underset{(0.53)}{2.19}\,hsize^2 -$$

$$\underset{(4.29)}{45.09}\,female - \underset{(12.71)}{169.81}\,black + \underset{(18.15)}{62.31}\,female \cdot black$$

gdzie, sat to wynik egzaminu SAT, hsize to wielkość rocznika dla danej osoby w liceum (liczona w setkach osób), female to zmienna zero-jedynkowa (1=kobieta, 0=mężczyzna), black to zmienna zero-jedynkowa (1=osoba czarnoskóra, 0=inne przypadki). W nawiasach podane są błędy standardowe oszacowań. (n=4,137, R2=.0858)

1. Jaka jest spodziewana różnica w wyniku testu między czarnoskórymi mężczyznami a mężczyznami, którzy nie są czarnoskórzy? Przetestuj hipotezę zerową, o braku różnicy w wynikach.
2. Jaka jest spodziewana różnica w wyniku testu między czarnoskórymi kobietami i kobietami, które nie są czarnoskóre? Jaki test należy przeprowadzić, żeby odpowiedzieć na pytanie czy różnica jest statystycznie istotna?

Zadanie 4

Zaznacz które zmienne w każdym z trzech modeli są istotne na poziomie 5% oraz 1%.

Results of Regressions of Average Weekly Earnings on Gender and Education Binary Variables and Other Characteristics Using 2007 Data from a Developing Country Survey

Dependent variable: log average weekly earnings (*AWE*).

Regressor	(1)	(2)	(3)
High school graduate (X_1)	0.352 (0.021)	0.373 (0.021)	0.371 (0.021)
Male (X_2)	0.458 (0.021)	0.457 (0.020)	0.451 (0.020)
Age (X_3)		0.011 (0.001)	0.011 (0.001)
North (X_4)			0.175 (0.037)
South (X_5)			0.103 (0.033)
East (X_7)			−0.102 (0.043)
Intercept	12.84 (0.018)	12.471 (0.049)	12.390 (0.057)
Summary Statistics and Joint Tests			
F-statistic for regional effects = 0			21.87
SER	1.026	1.023	1.020
R^2	0.0710	0.0761	0.0814
n	10973	10973	10973

Zadanie 4 c.d.

Korzystając z wyników w kolumnie (1): Czy na poziomie istotności 5%, możemy powiedzieć, że występuje różnica w wysokości zarobków między osobami, które ukończyły przynajmniej liceum a tymi które nie ukończyły? Skonstruuj przedział ufności na poziomie 95%

Results of Regressions of Average Weekly Earnings on Gender and Education Binary Variables and Other Characteristics Using 2007 Data from a Developing Country Survey

Dependent variable: log average weekly earnings (*AWE*).

Regressor	(1)	(2)	(3)
High school graduate (X_1)	0.352 (0.021)	0.373 (0.021)	0.371 (0.021)
Male (X_2)	0.458 (0.021)	0.457 (0.020)	0.451 (0.020)
Age (X_3)		0.011 (0.001)	0.011 (0.001)
North (X_4)			0.175 (0.037)
South (X_5)			0.103 (0.033)
East (X_7)			−0.102 (0.043)
Intercept	12.84 (0.018)	12.471 (0.049)	12.390 (0.057)
Summary Statistics and Joint Tests			
F-statistic for regional effects = 0			21.87
SER	1.026	1.023	1.020
R^2	0.0710	0.0761	0.0814
n	10973	10973	10973

Zadanie 4 c.d.

Korzystając z wyników w kolumnie (1): Czy na poziomie istotności 5%, możemy powiedzieć, że występuje różnica w wysokości zarobków między mężczyznami i kobietami? Skonstruuj przedział ufności na poziomie 95%.

Results of Regressions of Average Weekly Earnings on Gender and Education Binary Variables and Other Characteristics Using 2007 Data from a Developing Country Survey

Dependent variable: log average weekly earnings (*AWE*).

Regressor	(1)	(2)	(3)
High school graduate (X_1)	0.352 (0.021)	0.373 (0.021)	0.371 (0.021)
Male (X_2)	0.458 (0.021)	0.457 (0.020)	0.451 (0.020)
Age (X_3)		0.011 (0.001)	0.011 (0.001)
North (X_4)			0.175 (0.037)
South (X_5)			0.103 (0.033)
East (X_7)			−0.102 (0.043)
Intercept	12.84 (0.018)	12.471 (0.049)	12.390 (0.057)
Summary Statistics and Joint Tests			
F-statistic for regional effects = 0			21.87
SER	1.026	1.023	1.020
R^2	0.0710	0.0761	0.0814
n	10973	10973	10973

Zadanie 4 c.d.

Korzystając z wyników w kolumnie (2): Czy wiek jest statystycznie istotną determinantą zarobków?

Results of Regressions of Average Weekly Earnings on Gender and Education Binary Variables and Other Characteristics Using 2007 Data from a Developing Country Survey

Dependent variable: log average weekly earnings (*AWE*).

Regressor	(1)	(2)	(3)
High school graduate (X_1)	0.352 (0.021)	0.373 (0.021)	0.371 (0.021)
Male (X_2)	0.458 (0.021)	0.457 (0.020)	0.451 (0.020)
Age (X_3)		0.011 (0.001)	0.011 (0.001)
North (X_4)			0.175 (0.037)
South (X_5)			0.103 (0.033)
East (X_7)			−0.102 (0.043)
Intercept	12.84 (0.018)	12.471 (0.049)	12.390 (0.057)
Summary Statistics and Joint Tests			
F-statistic for regional effects = 0			21.87
SER	1.026	1.023	1.020
R^2	0.0710	0.0761	0.0814
n	10973	10973	10973

Zadanie 4 c.d.

Czy różnice między regionami są statystycznie istotne?

Results of Regressions of Average Weekly Earnings on Gender and Education Binary Variables and Other Characteristics Using 2007 Data from a Developing Country Survey

Dependent variable: log average weekly earnings (*AWE*).

Regressor	(1)	(2)	(3)
High school graduate (X_1)	0.352 (0.021)	0.373 (0.021)	0.371 (0.021)
Male (X_2)	0.458 (0.021)	0.457 (0.020)	0.451 (0.020)
Age (X_3)		0.011 (0.001)	0.011 (0.001)
North (X_4)			0.175 (0.037)
South (X_5)			0.103 (0.033)
East (X_7)			−0.102 (0.043)
Intercept	12.84 (0.018)	12.471 (0.049)	12.390 (0.057)
Summary Statistics and Joint Tests			
F-statistic for regional effects = 0			21.87
SER	1.026	1.023	1.020
R^2	0.0710	0.0761	0.0814
n	10973	10973	10973

Zadanie 5

- Podaj interpretację R2. Korzystając z informacji w tabeli, jak możemy policzyć R2?
- Które zmienne w modelu są istotne? Czy model jest łącznie istotny?

```
. reg lprice sqft_living waterfront bathrooms i.condition
```

Source	SS	df	MS	Number of obs	=	1,027
				F(7, 1019)	=	153.91
Model	151.587001	7	21.6552859	Prob > F	=	0.0000
Residual	143.378476	1,019	.14070508	R-squared	=	0.5139
				Adj R-squared	=	0.5106
Total	294.965477	1,026	.287490719	Root MSE	=	.37511

lprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
sqft_living	.0003658	.0000197	18.60	0.000	.0003273	.0004044
waterfront	.7436125	.218584	3.40	0.001	.3146862	1.172539
bathrooms	.0546636	.0240646	2.27	0.023	.0074417	.1018855
condition						
2	.5943416	.4109528	1.45	0.148	-.2120689	1.400752
3	.6375764	.3759918	1.70	0.090	-.1002303	1.375383
4	.7003995	.3763212	1.86	0.063	-.0380537	1.438853
5	.8434607	.3776874	2.23	0.026	.1023266	1.584595
_cons	11.49848	.3755286	30.62	0.000	10.76159	12.23538

Zadanie 6 - Wooldridge, Ch.6, problem 3

$$\widehat{rdintens} = \underset{(.429)}{2.613} + \underset{(.00014)}{.00030}sales - \underset{(.0000000037)}{.0000000070}sales^2$$

$n = 32, R^2 = .1484$ gdzie *rdintens* oznacza wydatki na R&D jako procent sprzedaży, sales oznacza wielkość sprzedaży w milionach dolarów. W nawiasach podane są błędy standardowe oszacowań.

1. Czy w modelu powinniśmy zostawić $sales^2$?
2. Zdefiniuj nową zmienną *salesbil*, która będzie wyrażać wielkość sprzedaży w miliardach dolarów tzn. salesbil = sales/1000. Zapisz równanie korzystając ze zmiennych *salesbil* oraz $salesbil^2$. Jakie będą wielkości błędów standardowych i R^2?
3. Który model wybierzesz? Uzasadnij.

Podsumowanie

- Powtórzyliśmy kluczowe zagadnienia z podstaw ekonometrii.