

Zadanie 1 (30%)

Zadanie bazuje na wykresie przedstawionym na blogu Junk Charts¹ i oryginalnie pochodzącym z tygodnika The Economist².

Analitik miał na celu rozważyć kwestię zależności między rocznym spożyciem lodów per capita w litrach (zmienna *IceCream*) a średnimi wynikami na teście PISA (zmienna *Pisa*). W tym celu analitik zebrał dane za 2018 rok dla wielu krajów. Dokładny opis zmiennej *Pisa* to "Mean score on PISA reading scale". Jak możemy przeczytać w opisie badania PISA na stronie Wikipedii³:

"Program PISA powstał w 1997 roku. PISA bada umiejętności i wiedzę ważną z perspektywy wyzwań, przed jakimi 15-latkowie staną w swym dorosłym życiu. Punktem wyjścia jest pojęcie «alfabetyzmu» (*literacy*) odnoszące się do «zdolności stosowania wiedzy i umiejętności, analizowania, argumentowania i efektywnego komunikowania w procesie stawiania, rozwiązywania i interpretowania problemów w różnych sytuacjach» (OECD 2003:23).

[...] W badaniu PISA wyróżniono trzy podstawowe dziedziny: czytanie i interpretacja (ang. *reading literacy*), matematyka (*mathematical literacy*) i rozumowanie w naukach przyrodniczych (*scientific literacy*). Badanie realizowane jest regularnie co trzy lata, począwszy od 2000 r. W każdym z kolejnych badań szczególny nacisk jest położony na zbadanie jednej dziedziny: w 2000 roku było to czytanie ze zrozumieniem, 2003 – matematyka, w 2006 – rozumowanie w naukach przyrodniczych. W 2003 roku badano dodatkowo umiejętność rozwiązywania problemów (*problem solving*)."

Wydruk poniżej przedstawia oszacowania modelu regresji konsumpcji lodów (zmienna *IceCream* nieprzekształcona, niezlogarytmowana) na średnim wynikach z testu PISA.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-15.452777	3.462447	-4.463	5.03e-05	***
Pisa	0.042039	0.007308	5.752	6.37e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.186 on 47 degrees of freedom

Multiple R-squared: 0.4132, Adjusted R-squared: 0.4007

F-statistic: 33.09 on 1 and 47 DF, p-value: 6.365e-07

```
> resettest(model_a, power=2:3, type="fitted")
```

RESET test

data: model_a

RESET = 4.9574, df1 = 2, df2 = 45, p-value = 0.01133

```
> bptest(model_a)
```

studentized Breusch-Pagan test

data: model_a

BP = 5.4953, df = 1, p-value = 0.01907

```
> jarque.bera.test(model_a$residuals)
```

Jarque Bera Test

data: model_a\$residuals

X-squared = 18.468, df = 2, p-value = 9.768e-05

¹https://junkcharts.typepad.com/junk_charts/2019/10/the-rule-governing-which-variable-to-put-on-which-axis.html

²<https://www.economist.com/graphic-detail/2016/04/01/ice-cream-and-iq>

³[https://pl.wikipedia.org/wiki/PISA_\(badanie\)](https://pl.wikipedia.org/wiki/PISA_(badanie))

Przyjmij 1% poziom istotności. Na wydrukach standardowo używana jest kropka jako separator dziesiętny.

- a) (5%) Zinterpretuj parametr przy zmiennej *Pisa*.
- b) (5%) Czy forma funkcyjna modelu jest prawidłowa? Uzasadnij.
- c) (5%) Czy występuje heteroskedastyczność reszt? Uzasadnij.
- d) (5%) Czy reszty mają rozkład normalny? Uzasadnij.
- e) (10%) Jak weryfikacja założeń tego modelu (forma funkcyjna, heteroskedastyczność, rozkład normalny) przekłada się na własności modelu?

Zadanie 2 (30%)

Warsaw Econometric Games⁴ jest konkursem ekonometrycznym przeznaczonym dla wszystkich studentów pierwszego i drugiego stopnia wszystkich polskich uczelni. Konkurs ma na celu pokazać, że badania naukowe mogą być fajne i zachęcać studentów do kontynuowania edukacji na studiach trzeciego stopnia. Uczestnicy, startujący w 4-osobowych zespołach, mają za zadanie napisać artykuł naukowy na temat zadany przez organizatorów w ciągu 24 godzin. Wykorzystywać mogą wszystkie modele ekonometryczne i metody machine learningu.

W pierwszej edycji konkursu, która odbyła się w kwietniu 2021 roku, uczestnicy mieli za zadanie przeanalizować zależność między nastrojem słuchanej muzyki a warunkami atmosferycznymi wykorzystując dane panelowe.

Spotify (platforma streamingowa muzyki) ma algorytm, który klasyfikuje dla każdej piosenki jej poziom "valence", czyli jak bardzo radośnie brzmi na skali od 0 do 100. Algorytm jest uczony na ratingach radosnego brzmienia stworzonych przez ekspertów muzycznych. Algorytm przypisuje utworowi "Respect" Arethy Franklin wynik 97, natomiast ponuremu utworowi Radiohead "Creep" tylko 10. Od 2017 roku Spotify publikuje codzienne statystyki w podziale na kraje.⁵

Państwa wykładowca zdecydował się zagregować miesięczne dane udostępnione uczestnikom konkursu do danych rocznych oraz ograniczyć się do danych z jednego roku aby uzyskać próbę przekrojową.

Tablica poniżej zawiera oszacowania modeli regresji dla średniego rocznego poziomu radosnego brzmienia muzyki w danym kraju (zmienna **valence**). Zmiennymi objaśniającymi były średni roczny poziom zachmurzenia (zmienna **sky**, ciągła w zakresie 0-8, im wyższa wartość, tym większa część nieba przykryta chmurami), średnioroczna temperatura w kraju (zmienna **temperature**, w stopniach Celsjusza), liczba stopniogrzania (*heating degree days*, zmienna **HDD**, zmienna ciągła), zmienna zerojedynkowa **rain** (przyjmująca 1 dla krajów z przewagą liczby dni deszczowych nad suchymi), oraz zmienna **happiness** kodująca średni poziom szczęścia mieszkańców (zmienna ciągła w skali 0-10).

Przyjmij $\alpha = 5\%$. Na wydrukach standardowo używana jest kropka jako separator dziesiętny.

- (5%) Zinterpretuj oszacowania modelu (4).
- (5%) Czy zmienna **temperature** jest statystycznie istotna? Czy jest praktycznie istotna? Uzasadnij.
- (5%) Który z modeli (1)-(5) należy uznać za najwłaściwszy do opisanego zjawiska? Uzasadnij.
- (5%) Opisz jak należałoby prawidłowo wybrać formę funkcyjną dla modelu dla zmiennej **valence**.
- (5%) Przetestuj hipotezę o tym, że w modelu (5) oszacowanie przy zmiennej **rain** równe jest (-1), tj. $H_0 : \beta_{rain} = -1$.
- (5%) Przetestuj hipotezę o łącznej nieistotności zmiennych **rain** oraz **happiness**.

Podpowiedź:

$$F_{test} = \frac{(\mathbf{H}\boldsymbol{\beta} - \mathbf{h})' [\mathbf{H}(\mathbb{X}'\mathbb{X})^{-1} \mathbf{H}']^{-1} (\mathbf{H}\boldsymbol{\beta} - \mathbf{h})/g}{\mathbf{e}'\mathbf{e}/(N - K)}$$

$$F_{test} = \frac{(S_R - S_U)/g}{S_U/(N - K)}$$

⁴<http://coin.wne.uw.edu.pl/wec/index.html>

⁵Źródło: The Economist, February 8th 2020, *Sad songs say so much. Data from Spotify suggest that listeners are gloomiest in February*

=====					
Dependent variable:					

	valence				
	(1)	(2)	(3)	(4)	(5)

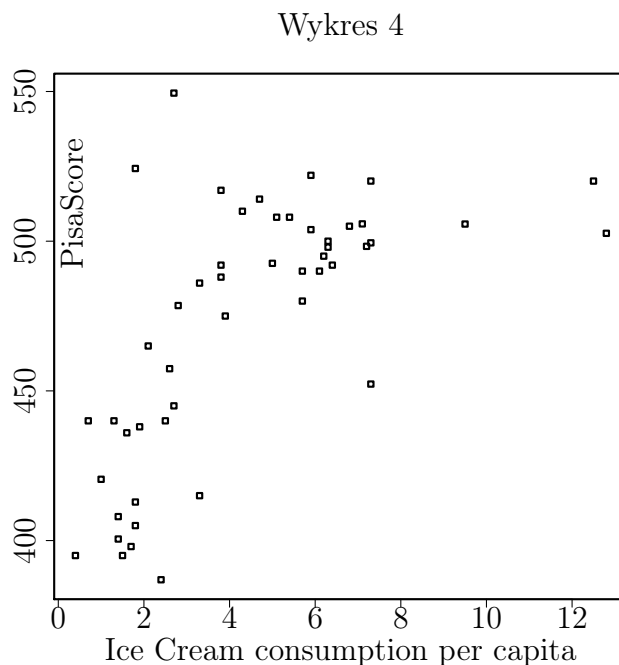
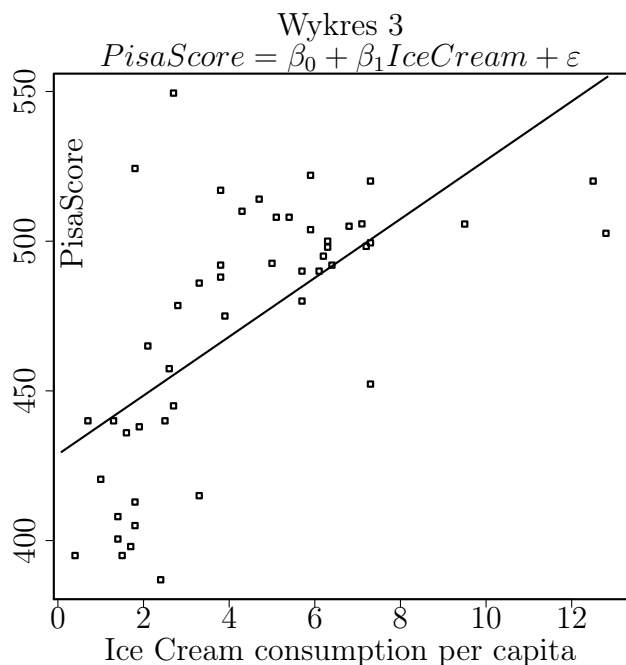
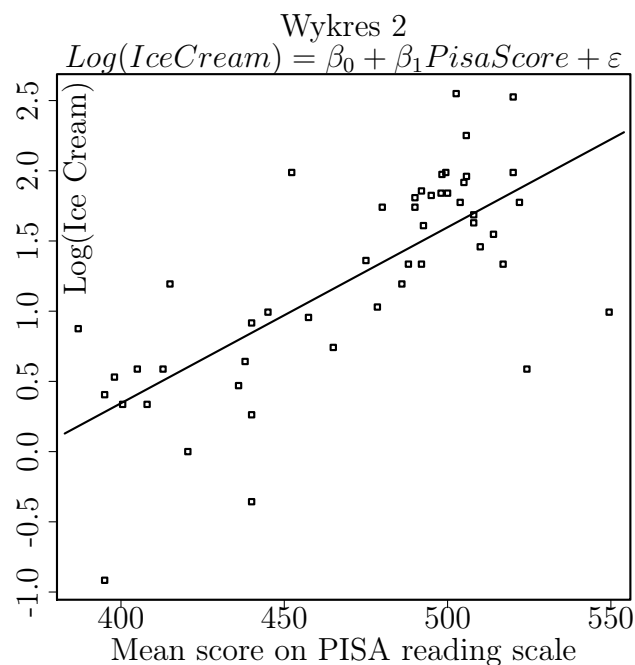
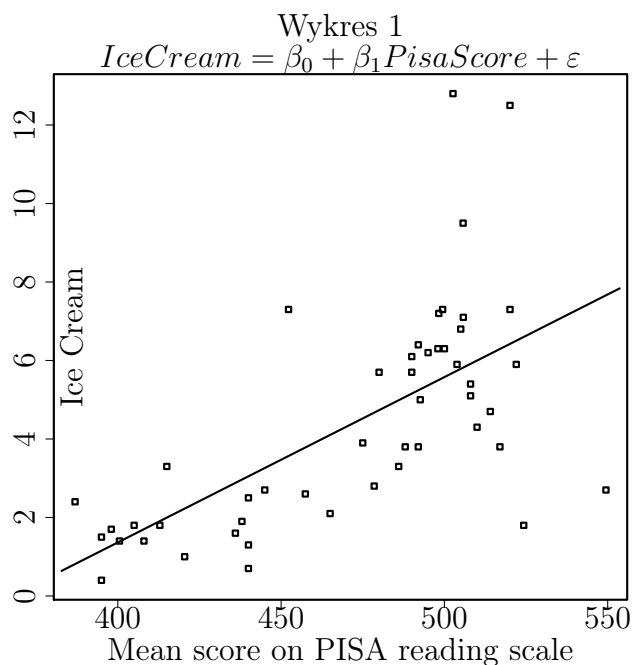
sky	0.0035 (0.0126)	0.0038 (0.0116)	0.0036 (0.0138)	0.0038 (0.0122)	
temperature	0.0004*** (0.0001)	0.0003*** (0.0001)	0.0006*** (0.0002)	0.0004*** (0.0001)	0.0004** (0.0002)
HDD		-0.0484 (2.7305)	-1.1227 (2.4707)		
rain			-1.8953*** (0.2219)	-1.7422*** (0.5089)	-1.3796*** (0.5634)
happiness				1.0773** (0.4221)	1.5239*** (0.2532)
Constant	0.0390 (0.0902)	0.1700* (0.0870)	-0.2090** (0.1030)	0.1031 (0.1192)	1.4363*** (0.1040)

Observations	109	109	109	109	109
R2	0.1058	0.1131	0.1894	0.2315	0.2256
Adjusted R2	0.0956	0.0923	0.1671	0.2267	0.2250
RSS	16.4512	16.0251	12.9964	6.6273	6.8035
=====					
Note:	*p<0.1; **p<0.05; ***p<0.01				

Zadanie 3 (30%)

Ten sam analityk co w zadaniu (1) postanowił przeanalizować dane pod innym kątem. Sporządził 4 wykresy i oszacował 3 modele ekonometryczne. Wykresy (1) i (2) przedstawiają zależność między średnimi wynikami z testu PISA a konsumpcją lodów, przy czym na wykresie (1) mamy poziom zmiennej a na wykresie (2) - jej logarytm. Wykres (3) prezentuje zależność między poziomem konsumpcji lodów a średnimi wynikami na teście PISA, odwrotnie niż na wykresie (1).

Na wydrukach standardowo używana jest kropka jako separator dziesiętny.



- a) (10%) Który z modeli (1)-(2) dla konsumpcji lodów należy uznać za lepszy? Uzasadnij.
- b) (10%) Rozważmy model dla średniego wyniku na teście PISA z wykresu (3). Zaproponuj transformację którejś ze zmiennych tak aby prosta $y = ax + b$ miała szansę być dobrze dopasowana do danych. Uzasadnij.
- c) (10%) Wykres (4) przedstawia te same punkty co wykres (3) z tym, że pozbawiony jest prostej regresji. Zaproponuj model ekonometryczny, który będzie składał się z liniowych splajnów a zależność będzie ciągła we wszystkich punktach dziedziny. Uzasadnij. **Narysuj na wykresie oczekiwaną formę zależności wymaganej postaci.**
- d) (10%) Skomentuj kierunek zależności między konsumpcją lodów a wynikami z badania PISA. Określ kierunek zależności o ile istnieje. Odpowiedź uzasadnij.

Tabela 1: Wyniki regresji przedstawionych na wykresach (1)-(3)

	<i>Dependent variable:</i>		
	IceCream	Log(IceCream)	Pisa
	(1)	(2)	(3)
Pisa	0.042*** (0.007)	0.013*** (0.002)	
IceCream			9.828*** (1.709)
Constant	-15.453*** (3.462)	-4.667*** (0.818)	428.778*** (8.882)
Observations	49	49	49
R ²	0.413	0.528	0.413
Adjusted R ²	0.401	0.518	0.401
Residual Std. Error (df = 47)	2.186	0.516	33.420
F Statistic (df = 1; 47)	33.091***	52.649***	33.091***
<i>Note:</i>		*p<0.1; **p<0.05; ***p<0.01	