Zadanie 1 (40%)

Jednym z najlepszych, zdaniem Państwa wykładowcy, blogów ekonometrycznych jest *Econometrics Beat: Dave Giles' Blog*[1]. Pojawił się na nim wpis o drugim najdłuższym słowie w ekonometrii[2]. Słowem tym jest *multicollinearity*, czyli współliniowość. Cały wpis dotyczy uwagi jaką poświęca się współliniowości w uczeniu ekonometrii. Autor bloga twierdzi, i nie tylko on, że współliniowości poświęca się zbyt dużo uwagi i nie jest to rzecz, którą powinniśmy się zajmować. Zdaniem autora dużo ważniejszą kwestią jest wpływ małej liczby obserwacji na jakość oszacowań niż współliniowość.

Dodatkowo, autor bloga przytacza dane o tym jak wiele uwagi poświęconych zostało współliniowości w wielu podręcznikach do ekonometrii. Państwa wykładowca pobrał udostępnione dane[3] i postanowił powtórzyć i rozszerzyć analizy przedstawione przez prof. Dave'a Gilesa.

Wydruk poniżej przedstawia oszacowania modelu regresji procentowego udziału stron poświęconych współliniowości w całkowitej liczbie stron podręcznika (zmienna `MULT` nieprzekształcona, niezlogarytmowana) na roku publikacji podręcznika (`Year`), numerze wydania (`ED` - od *edition*) oraz zmiennej zerojedynkowej `DNA` przyjmującej wartość jeden dla podręczników amerykańskich.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.55543   32.64846   1.916   0.0620 .
Year        -0.03081    0.01648  -1.870   0.0683 .
ED           0.44198    0.16676   2.650   0.0112 *
DNA          0.09333    0.63054   0.148   0.8830
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.314 on 43 degrees of freedom
Multiple R-squared:  0.1516,Adjusted R-squared:  0.0924
F-statistic: 2.561 on 3 and 43 DF,  p-value: 0.06729

> resettest(model, power=2:3, type="fitted")
RESET test
data:  model
RESET = 0.11853, df1 = 2, df2 = 41, p-value = 0.8885

> bptest(model)
studentized Breusch-Pagan test
data:  model
BP = 2.8709, df = 3, p-value = 0.412

> jarque.bera.test(model$residuals)
Jarque Bera Test
data:  model$residuals
X-squared = 1.3802, df = 2, p-value = 0.5015
```

[1] https://davegiles.blogspot.com/

[2] https://davegiles.blogspot.com/2011/03/second-longest-word-in-econometrics.html

[3] http://web.uvic.ca/~dgiles/blog/multicollinearity.xls

Przyjmij 10% poziom istotności. Na wydrukach standardowo używana jest kropka jako separator dziesiętny.

a) (9%) Do czego służy statystyka VIF? W jakim celu analizujemy współliniowość?

b) (6%) Zinterpretuj oszacowania parametrów modelu.

c) (5%) Czy forma funkcyjna modelu jest prawidłowa? Uzasadnij.

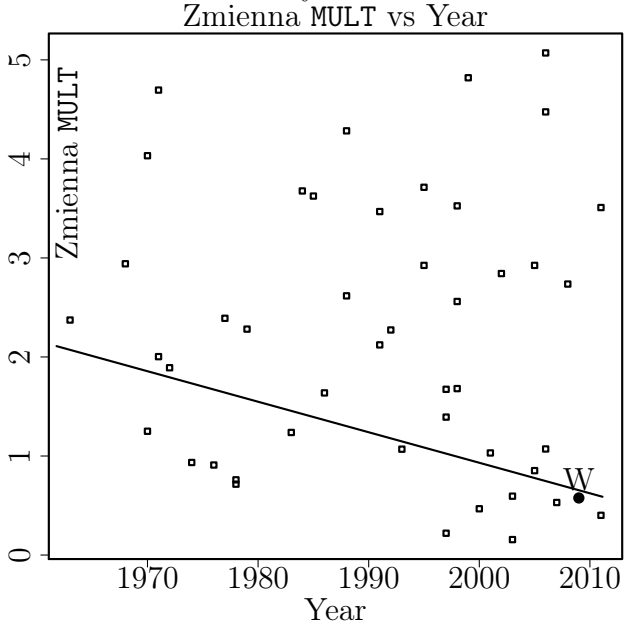
d) (5%) Czy występuje heteroskedastyczność reszt? Uzasadnij.

e) (5%) Czy reszty mają rozkład normalny? Uzasadnij.

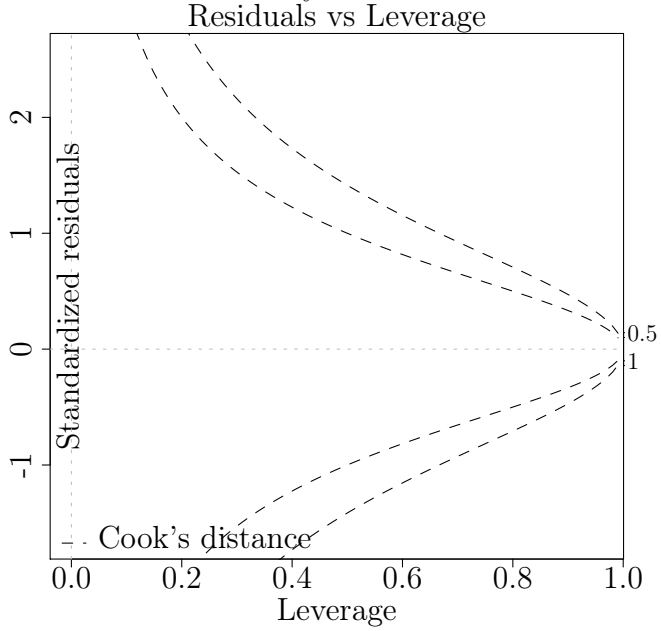
f) (5%) Jak weryfikacja założeń tego modelu (forma funkcyjna, heteroskedastyczność, rozkład normalny) przekłada się na własności modelu?

g) (5%) Na zamieszczonym poniżej wykresie 1 przedstawione zostały obserwacje. Wyróżniona obserwacja reprezentuje podręcznik J. Wooldridge'a *Introductory Econometrics*, wydanie szóste. Określ pozycję tej obserwacji na wykresie 2 *Residuals vs Leverage*. Uzasadnij.

Wykres 1
Zmienna MULT vs Year



Wykres 2
Residuals vs Leverage



Zadanie 2 (30%)

Zadanie jest kontynuacją zadania 1. Państwa wykładowca postanowił sprawdzić kilka dodatkowych specyfikacji modeli dla zmiennej `MULT`. Tablica poniżej zawiera oszacowania 4 modeli regresji. Modele (1)-(3) są modelami regresji i różnicuje je jedynie liczba regresorów. Model (3) jest jednocześnie tym samym modelem, którego wyniki i diagnostyka zostały przedstawione w zadaniu 1. Model (4) jest odpowiednikiem modelu (3) z tym, że w modelu (4) wykorzystany został estymator odporny macierzy wariancji-kowariancji White'a.

Przyjmij 10% poziom istotności. Na wydrukach standardowo używana jest kropka jako separator dziesiętny.

	Dependent variable:			
	MULT			
	OLS			*coefficient test*
	(1)	(2)	(3)	(4)
Year	−0.009	−0.031*	−0.031*	−0.031**
	(0.015)	(0.016)	(0.016)	(0.015)
ED		0.445***	0.442**	0.442***
		(0.163)	(0.167)	(0.142)
DNA			0.093	?????
			(0.631)	(0.358)
Constant	19.728	62.604*	62.555*	62.555**
	(30.155)	(32.282)	(32.648)	(29.062)
Observations	47	47	47	
R^2	0.007	0.151	0.152	
Adjusted R^2	−0.015	0.113	0.092	
Residual Std. Error	1.390	1.300	1.314	
F Statistic	0.338	3.918**	2.561*	
Note:				*p<0.1; **p<0.05; ***p<0.01

a) (5%) Dlaczego skorygowane R^2 jest ujemne w modelu (1)?

b) (5%) Jak zweryfikować hipotezę o tym, że autorzy podręczników do ekonometrii poświęcają coraz mniejszą część miejsca współliniowości? Uzasadnij.

c) (5%) Ile wynosi oszacowanie parametru przy zmiennej `DNA` modelu (4)? Ile gwiazdek powinno być obok tego oszacowania? Uzasadnij.

d) (5%) Czy użycie macierzy odpornej w modelu (4) było uzasadnione w kontekście wyników modelu i diagnostyki modelu z zadania 1? Uzasadnij.

e) (5%) Ile powinno wynosić R^2 w modelu (4)? Uzasadnij.

f) (5%) Który model z przedstawionych w tabeli powyżej jest najlepszy? Uzasadnij.

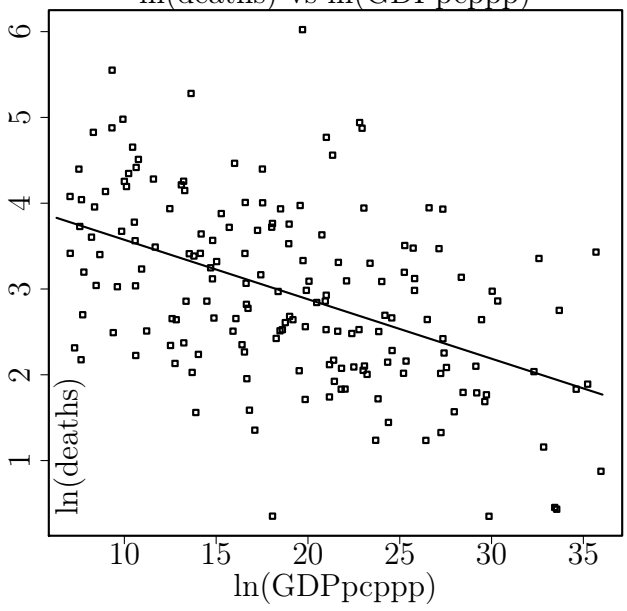
Zadanie 3 (30%)

Pomysł na to zadanie wykorzystuje hipotezę przedstawioną w artykule-grafice tygodnika The Economist[4]. Państwa wykładowca postanowił zweryfikować hipotezę o tym, że transparentne rządy demokratyczne radzą sobie z epidemiami w sposób bardziej efektywny od rządów niedemokratycznych.

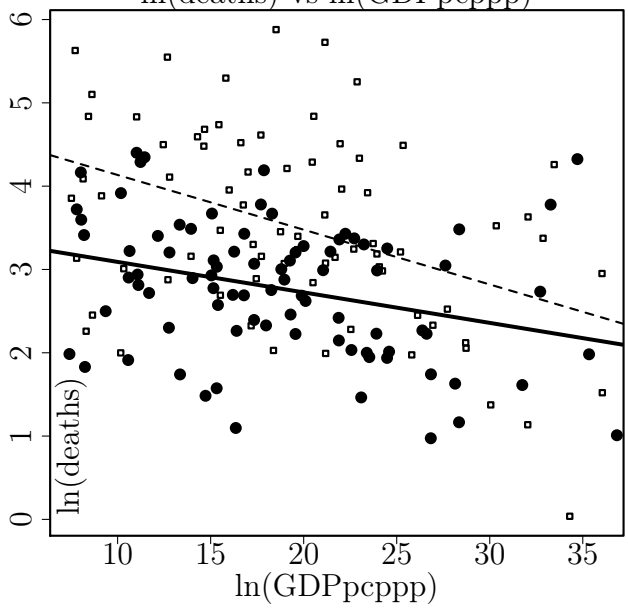
Hipoteza badawcza została zweryfikowana w oparciu o model regresji liniowej. Zmienną objaśniającą była zmienna `deaths` reprezentująca liczbę zgonów na 100 tys. mieszkańców z powodu epidemii w danym kraju w danym roku. Zmiennymi objaśniającymi były PKB per capita według parytetu siły nabywczej w cenach z 2020 roku (zmienna `GDPpcppp`) oraz zmienna zerojedynkowa `democracy` przyjmująca wartość jeden dla krajów demokratycznych i zero dla pozostałych.

Wykres 3 przedstawia zależność między logarytmem zmiennej `deaths` a logarytmem zmiennej `GDPpcppp` wraz z prostą regresji. Wykres 4 zawiera te same obserwacje ale w podziale na państwa demokratyczne (obserwacje i prosta wyboldowana) oraz niedemokratyczne (kwadraty i przerywana linia).

Wykres 3
ln(deaths) vs ln(GDPpcppp)



Wykres 4
ln(deaths) vs ln(GDPpcppp)



Przyjmij 10% poziom istotności. Na wydrukach standardowo używana jest kropka jako separator dziesiętny.

a) (5%) Jak można wytłumaczyć istnienie zależności między poziomem dochodu w kraju a stopą zgonów z powodu epidemii?

b) (5%) Jaką dodatkową zmienną możnaby uwzględnić w modelowaniu stopy zgonów z powodu epidemii? Uzasadnij.

c) (5%) Państwa wykładowca podejrzewał, że parametry modelu przedstawionego na wykresie 3 mogą być niestabilne względem zmiennej `democracy`. Czy wykres 4 potwierdza to przypuszczenie?

d) (5%) Zaproponować postać teoretycznego modelu regresji za pomocą którego możnaby wykonać test stabilności parametrów w przypadku modelowania stopy zgonów względem PKB per capita według parytetu siły nabywczej w logarytmach.

e) (5%) Przeprowadzić test Chowa na stabilność parametrów względem zmiennej `democracy` wykorzystując wyniki modeli przedstawionych poniżej.

[4] `https://www.economist.com/graphic-detail/2020/06/06/democracies-contain-epidemics-most-effectively`

f) (5%) Państwa wykładowca podejrzewał także, że w przypadku modeli przedstawionych na wykresie 4 możemy mieć doczynienia z heteroskedastycznością. Mianowicie, dla państw demokratycznych obserwujemy mniejszą wariację reszt niż dla krajów niedemokratycznych. Opisać jak wykorzystać *heteroskedasticity partition* w tym kontekście.

Podpowiedź:

$$F = \frac{(S_R - \sum S_j)/[K(m-1)]}{\sum S_j/(N-mK)} \sim F(K(m-1), N-mK) \tag{1}$$

Tabela poniżej przedstawia oszacowania modeli regresji dla modeli postaci:

$$ln(deaths) = \beta_0 + \beta_1 ln(GDPpcppp) + \varepsilon. \tag{2}$$

W kolumnie (1) znajdują się wyniki dla wszystkich krajów, w kolumnie (2) - dla krajów demokratycznych, w kolumnie (3) - krajów niedemokratycznych.

	Dependent variable:		
	lnDeaths	[dem == 1]	[dem == 0]
	(1)	(2)	(3)
lnGDP	−0.048***		
	(0.011)		
lnGDP[dem == 1]		−0.037***	
		(0.012)	
lnGDP[dem == 0]			−0.066***
			(0.017)
Constant	4.025***	3.456***	4.792***
	(0.226)	(0.245)	(0.348)
Observations	161	85	76
R^2	0.104	0.097	0.176
Adjusted R^2	0.099	0.086	0.165
Residual Std. Error	1.003	0.786	1.059
RSS	159.920	51.249	82.933
F Statistic	18.502***	8.934***	15.767***
Note:		*p<0.1; **p<0.05; ***p<0.01	

