HETEROSKEDASTYCZNOŚĆ - ROZWIĄZANIA.

ZAJĘCIA NR 11.

1. (W, 1 s. 297) Które z poniższych są konsekwencją heteroskedastyczności?

   (i) Estymator MNK nie jest zgodny.
   Rozwiązanie: Zgodność estymatora objawia się w sytuacji gdy jeśli rośnie liczebność próby, to rośnie również szansa, że oszacowanie uzyskane przy pomocy estymatora będzie przyjmować wartości coraz bliższe prawdziwej wartości szacowanego parametru. Nawet gdy heteroskedastyczność, czyli zmienność wariancji, jest obecna, estymator MNK pozostaje zgodny.

   (ii) Statystyka F nie ma już rozkładu F.
   Rozwiązanie: Uwaga - Tak. Aby statystyka F miała rozkład F, musi być spełnione założenie o homoskedastyczności oraz normalności składnika losowego.

   (iii) Estymator MNK nie jest już BLUE.
   Rozwiązanie: Skrót BLUE oznacza *Best Linear Unbiased Estimator* (Najlepszy Liniowy Nieobciążony Estymator). W przypadku gdy heteroskedastyczność jest obecna, estymator MNK jest nieobciążony ponieważ nieobciążoność dotyczy wartości oczekiwanej estymatora. Natomiast w przypadku heteroskedastyczności estymator nie jest już najlepszy, to znaczy że nie ma już najmilejszej wariancji w klasie liniowych, nieobciążonych estymatorów.

2. Rozważ model objaśniający zlogarytmowane płace:

$$\log(w) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \varepsilon$$

   gdzie $educ$ to liczba lat edukacji, $exper$ to liczba lat zawodowego doświadczenia, $female$ to zm. zerojedynkowa oznaczająca kobiety, oraz ε to składnik losowy.

   (i) Czy spodziewasz się heteroskedastyczności składnika losowego w powyższym modelu?
   Rozwiązanie: Heteroskedastyczność dotyczny sytuacji gdy wariancja estymatora MNK nie jest stała. W przypadku rozważanego modelu, wariancja (dokładniej macierz wariancji-kowariancji) ma postać:

$$\mathrm{Var}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\intercal} \mid \boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[\varepsilon_1^2 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_1\varepsilon_2 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_1\varepsilon_3 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_1\varepsilon_4 \mid \boldsymbol{X}] \\ \mathbb{E}[\varepsilon_2\varepsilon_1 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_2^2 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_2\varepsilon_3 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_2\varepsilon_4 \mid \boldsymbol{X}] \\ \mathbb{E}[\varepsilon_3\varepsilon_1 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_3\varepsilon_2 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_3^2 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_3\varepsilon_4 \mid \boldsymbol{X}] \\ \mathbb{E}[\varepsilon_4\varepsilon_1 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_4\varepsilon_2 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_4\varepsilon_3 \mid \boldsymbol{X}] & \mathbb{E}[\varepsilon_4^2 \mid \boldsymbol{X}] \end{bmatrix}$$

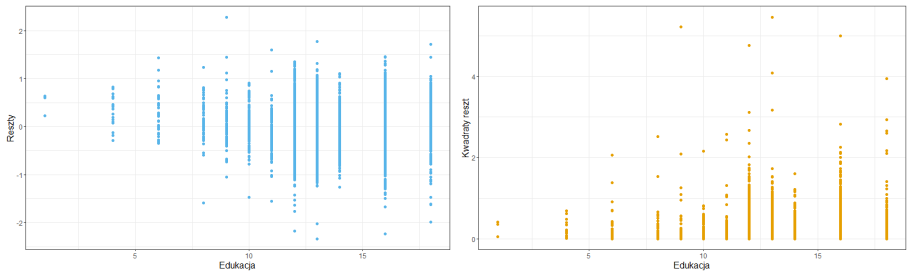
   W przypadku spełnienia założenia o sferyczności wariancji, elementy poza diagonala są równe zero, natomiast te na diagonali są równe. W przypadku heteroskedastyczności, elementy na diagonali macierzy wariacji-kowariancji estymatora MNK są różne. Elementy te oznaczają warunkową wariancję reszt względem zmiennych objaśniających. Zatem należy zastanowić się czy wariancja reszt będzie różniła się między jednostkami, które raportują różne wartości zmiennych objaśniających. Innymi słowy, dlaczego konkretny zestaw zmiennych objaśniających lepiej prognozuje zmienną objaśniającą dla pewnych jednostek obecnych w próbie a gorzej dla innych? W rozważanym przypadku, zmienność składnika losowego może zależeć od np. poziomu edukacji, ponieważ dla niskich wartości edukacji płace raczej będą niskie i składnik losowy będzie mało zmienny; natomiast dla wyższych wartości edukacji płace będą wyższe i składnik losowy będzie bardziej zmienny, ponieważ np. czynniki specyficzne dla konkretnych zawodów mogą grać dużą rolę. Np. nauczyciele i menadżerowie, obie grupy mają ukończone studia, a możemy domyślać się

że ta druga grupa odznacza się wyższymi zarobkami. Także zmienność składnika losowego może różnić się systematycznie między mężczyznami i kobietami. Heteroskedastyczność może także występować z powodu obecności obserwacji odstających.

(ii) Użyj danych `cps.dta` aby oszacować parametry modelu. Oblicz kwadraty reszt i zrób wykres kwadratów reszt z każdą zmienną objaśniającą. Czy obserwujesz heterosekdastyczność w tym przypadku? Podaj interpretację ekonomiczną.

Rozwiązanie: Zobacz kod wstaw kod!. Rysunek 1 Wykres przedstawia zależność reszt i kwadratów reszt oszacowanej regresji oraz edukacji. Na obu wykresach obserwujemy wyraźną zależność między wartościami reszt oraz edukacją. Zgodnie z przewidywaniami, zmienność reszt rośnie wraz z wartościami edukacji. Przyrost zmienności ta może być wytłumaczony specyficznymi czynnikami dla różnych zawodów, jak w podpunkcie (a).

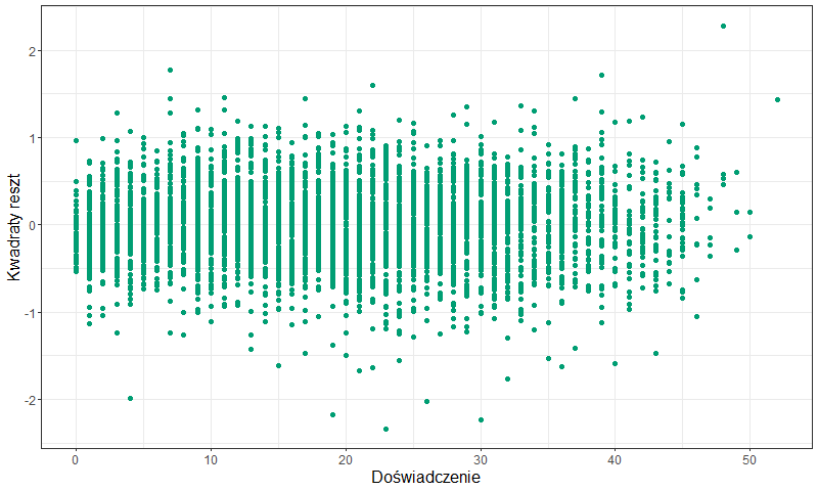
Rysunek 1: Reszty vs edukacja.



(a) Reszty. **(b)** Kwadraty reszt.

Rysunek 2 przedstawia zależność między kwadratami reszt oraz doświadczeniem (zmienna `exper`). Wariancja reszt nie zmienia się w zależności od wartości doświadczenia. Można więc podejrzewać że doświadczenie nie jest źródłem heteroskedastyczności.

Rysunek 2: Kwadraty reszt vs doświadczenie.



(iii) Użyj testu White'a do przetestowania heteroskedastyczności.

Rozwiązanie: Stosujemy procedurę testową tak jak opisano w slajdach, tzn. szacujemy regresję pomocniczą objaśniającą kwadraty reszt modelu, wyestymowanego w podpunkcie (ii). W przypadku testu White'a

regresja ta zawiera wszystkie represory z modelu z (ii) oraz ich kwadraty oraz interakcje. Wyniki z oszacowania takiego modelu są przedstawione w Tabeli 1. Mając takie wyniki, możemy od razu przeprowadzić procedurę testową. Przypomnimy, że hipoteza zerowa mówi o homokedastyczności składnika losowego, a alternatywna o heteroskedastyczności. Możemy zweryfikować hipotezę zerową na podstawie testu F. Statystyka F jest podana w ostatnim rzędzie Tabeli 1. W ynikach nie znajduje się dokładana wartość p-value, lecz widzimy że test F jest istotny na poziomie istotności 0.01. Oznacza to że mamy podstawy do odrzucenia hipotezy zerowej, na rzecz hipotezy alternatywnej. Tak więc mamy podstawy aby sądzić że rozważanym przypadku problem heteroskedaastyczności istnieje.
Co zrobić jeśli nie ma podanych wyników testu F, lub chcemy skorzystać ze statystyki mnożników Lagrange (LM)? W wynikach mamy podany R^2 oraz liczbę obserwacji (n) i zmiennych (k). Wtedy możemy samodzielnie obliczyć wartości statystyk. Wnoszą one:

$$F = \frac{0.019\frac{1}{12}}{(1-0.019\frac{1}{4733-12-1}} = 7.751316$$
$$LM = 4733 \times 0.019 = 91.46942$$

Korzystając ze wspominanych danych musimy odleźć jeszcze wartości krytyczne, z którymi porównamy nasze statystyki. Wynoszą one (przyjmujemy $\alpha = 0.05$):

$$F^{kryt}_{\alpha=0.05,12,4733-12-1} = 1.75422$$
$$\chi^{2,kryt}_{\alpha=0.05,12} = 21.02607$$

Statystyki testowe są wyższe niż wartości krytyczne więc mamy podstawy aby odrzucić hipotezę zerową o homoskedastyczności, na rzecz hipotezy alternatywnej.
Oczywiście, wartości krytyczne oraz związane z nimi p-value można wyliczyć w R. Procedura ta została opisana w kodzie.

(iv) Użyj testu BP do przetestowania heteroskedastyczności.
Rozwiązanie: Robimy to analogicznie jak w podpunkcie (iii). Test Breusch'a-Pagana przeprowadza się tak ja test White'a, jedną różnicą jest postać regresji pomocniczej. Szacując ją, uwzględniamy jedynie poziomy zmiennych z modelu z podpunktu (ii). Wyniki są przedstawione w Tabeli 1 w panelu (2). Statystyka F jest istotna na poziomie 0.01, więc mamy podstawy aby hipotezę zerową o homoskedastyczności odrzucić na rzecz hipotezy alternatywnej. Do takiej samej konkluzji dojdziemy, obliczając potrzebne statystyki samodzielnie:

$$F = \frac{0.012\frac{1}{4}}{(1-0.012\frac{1}{4733-4-1}} = 13.81687$$
$$LM = 4733 \times 0.012 = 54.68668$$

Korzystając ze wspominanych danych musimy odleźć jeszcze wartości krytyczne, z którymi porównamy nasze statystyki. Wynoszą one (przyjmujemy $\alpha = 0.05$):

$$F^{kryt}_{\alpha=0.05,4,4733-4-1} = 2.373811$$
$$\chi^{2,kryt}_{\alpha=0.05,4} = 9.487729$$

Wartości statystyk testowych są wyższe niż wartości krytyczne.

Tabela 1: Wyniki regresji pomocniczych

	Zmienna zależma: residsq	
	(1)	(2)
educ	0.012	0.008***
	(0.016)	(0.002)
exper	0.013	0.004**
	(0.011)	(0.001)
exper2	0.000	−0.000
	(0.001)	(0.000)
female	−0.125*	−0.008
	(0.064)	(0.010)
I(educ^2)	0.000	
	(0.001)	
I(exper^2)		
I(exper2^2)	0.000	
	(0.000)	
I(female^2)		
educ:exper	−0.001	
	(0.001)	
educ:exper2	−0.000	
	(0.000)	
educ:female	0.010**	
	(0.004)	
exper:female	0.000	
	(0.003)	
exper:exper2	−0.000	
	(0.000)	
exper2:female	−0.000	
	(0.000)	
Constant	−0.078	0.039
	(0.134)	(0.032)
Observations	4,733	4,733
R^2	0.019	0.012
Adjusted R^2	0.017	0.011
Residual Std. Error	0.339 (df = 4720)	0.340 (df = 4728)
F Statistic	7.751*** (df = 12; 4720)	13.817*** (df = 4; 4728)

Notes: ***Significant at the 1 percent level, **Significant at the 5 percent level, *Significant at the 10 percent level.

(v) Oszacuj model używając odpornych błędów standardowych. Omów różnice.
Rozwiązanie: Heteroskedastyczność psuje własności wariancji estymatora MNK. Na podstawie wariancji wyliczamy błędy standardowe naszych oszacowań, z których później korzystamy do konstrukcji statystyk testowych. Jeśli wiemy że występuje problem z heteroskedastycznością, to należy zastosować odporne błędy standardowe. W `R` możemy uzyskac je korzystając z komendy `coeftest`. Oszacowania, zwykłe i odporne błędy standardowe przedstawiono w Tabeli 2. W przypadku doświadczenia oraz indykatora kobiet, odporne błędy standardowe są nieco niższe niż zwykłe. Odwrotna sytuacja zachodzi w przypadku kwadratu doświadczenia oraz edukacji. W rozważanym przypadku, różnice te nie są duże, jednak czasami, kiedy te różnice są wyższe, zastosowanie odpornych błędów standardowych może prowadzić np.

do zmiany oceny istotności statystycznej danego oszacowania parametru.

Tabela 2: Porównanie błędów standardowych

	Współczynniki	Zwykłe błędy std.	Odporne błędy std.
Intercept	0.414340	0.041618	0.041395
Educ	0.109250	0.002813	0.002918
Exper	0.039256	0.001966	0.001936
Exper^2	-0.000665	0.000046	0.000048
Female	-0.249731	0.012968	0.012931

3. (W, C2 s. 299) Użyj danych `HPRICE1`

   (i) Uzyskaj odporne na heteroskedastyczność błędy standardowe dla równania

   $$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + \varepsilon$$

   Omów różnice między zwykłymi błędami standardowymi a odpornymi.
   Rozwiązanie: Dla `lotsize` oraz `sqrft` odporne błędy standardowe są wyższe niż zwykłe, a w przypadku `bdrms` przeciwnie. Przysrost błędów standardowych objawia się w zmianie poziomu istotności zmiennej `lotsize`.

   (ii) Powtórz (i) dla równania

   $$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + \varepsilon$$

   Rozwiązanie: Po zastosowaniu logarytmów odporne błędy standardowe są wyższe w przypadku wszystkich zmiennych. Jednak te różnice są niewielkie i poziomy istotności (poza stałą, co jest mało istotne) się nie zmieniają.

   (iii) Co ten przykład sugeruje na temat heteroskedastyczności oraz transformacji zastosowanej do zmiennej zależnej?
   Rozwiązanie: Heteroskedatyczność bierze się ze zróżnicowania wariancji, podczas gdy zastosowanie logarytmów zmniejsza wariancje zmiennych oraz wpływ obserwacji odstających. Tak więc transformacja logarytmami może być pomocna w przypadku heteroskeadstyczności.

4. (W, C3 s. 299) Zastosuj test White'a na heteroskedastyczność do równania (dane `HPRICE1`):

   $$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + \varepsilon$$

   Użyj statystyki χ^2, uzyskaj *p-value*. Jaka jest konkluzja?
   Rozwiązanie: Procedurę testową przeprowadzamy tak jak np. w zadaniu 2. Statystyka testowa: $LM_{\alpha=0.05,3} = 56.20563$, $p - value = 3.7972e^{-12}$. Na poziomie istotności $\alpha = 0.05$ mamy podstawy do odrzucenia hipotezy zerowej o homoskedastyczności składnika losowego.

5. (W, C4 s. 299) Użyj danych `VOTE1` do tego ćwiczenia.

   (i) Oszacuj model z $voteA$ jako zmienną zależną oraz zmiennymi $prtystrA$, $democA$, $log(expendA)$, i $log(expendB)$ jako zmiennymi niezależnymi. Uzyskaj reszty i policz regresję reszt na wszystkie zmienne niezależne. Wyjaśnij dlaczego otrzymujesz $R^2 = 0$.
   Rozwiązanie: $R^2 = 0.0525$ - to współczynnik determinacji z regresji reszt. Nie jest równy zero, ale jest

niski. Być może relacja między wariancją reszt a zestawem zmiennych objaśniających jest bardzo słaba i dlatego możemy spodziewać się

(ii) Policz test Breusch'a-Pagan'a na heteroskedastyczność. Użyj statystyki F i podaj *p-value*.
Rozwiązanie: Statystyka testowa wynosi 2.33, związana z nią $p - value = 0.05805751$. W skrócie, test BP sugeruje że składnik losowy jest homoskedastyczny.

(iii) Zastosuj test White'a na heteroskedastyczność, znów używając statystyki F. Jak silny jest dowód na istnienie heteroskedastyczności?
Rozwiązanie: Statystyka testowa wynosi 1.73, związana z nią $p - value = 0.0763$. Zatem test White'a sugeruje że składnik losowy jest homoskedastyczny. Ogólnie oba testy, przy poziomie istotności $\alpha = 0.05$ dają taki sam wynik, więc dowód na brak heteroskedastyczności jest dość przekonujący. Jednak warto zauważyć że w obu testach $p - value$ nie są bardzo wysokie, jedynie trochę przewyższają standardowy poziom istotności.