

Chapter 8: Heteroskedasticity

Introductory Econometrics: A Modern Approach

Outline

8.1 Consequences of heteroskedasticity for OLS

8.2 Heteroskedasticity-robust inference after OLS estimation

- A Heteroskedasticity-Robust LM Statistic

8.3 Testing for heteroskedasticity

- Breusch-Pagan test for heteroskedasticity
- The White test for heteroskedasticity
 - An alternative to White test

8.4 Weighted least squares estimation

- Heteroskedasticity is known up to a multiplicative constant
- Unknown heteroskedasticity function (feasible GLS)
- What if the assumed heteroskedasticity function is wrong?

Introduction

Homoskedasticity assumption states that the variance of the unobserved error, u , conditional on the explanatory variables is constant.

- In Chapters 4 and 5, we saw that homoskedasticity is needed to justify the usual **t tests**, **F tests**, and **confidence intervals** for OLS estimation of the linear regression model, even with large sample sizes

Homoskedasticity assumption fails whenever the variance of the unobserved factors changes across different segments of the population (called **Heteroskedasticity**)

In this chapter, we discuss the available remedies when heteroskedasticity occurs, and we also show how to test for its presence. We begin by briefly reviewing the consequences of heteroskedasticity for ordinary least squares estimation.

8.1 Consequences of heteroskedasticity for OLS

Consequences of heteroskedasticity

- invalidates variance formulas for OLS estimators
- **F tests** and **t tests** are not valid under heteroskedasticity
- OLS is no longer the *best linear unbiased estimator (BLUE)*; there may be more efficient linear estimators

OLS still unbiased and consistent under heteroskedasticity!

Also, interpretation of R-squared is not changed

Unconditional error variance is unaffected by heteroskedasticity (which refers to the conditional error variance)

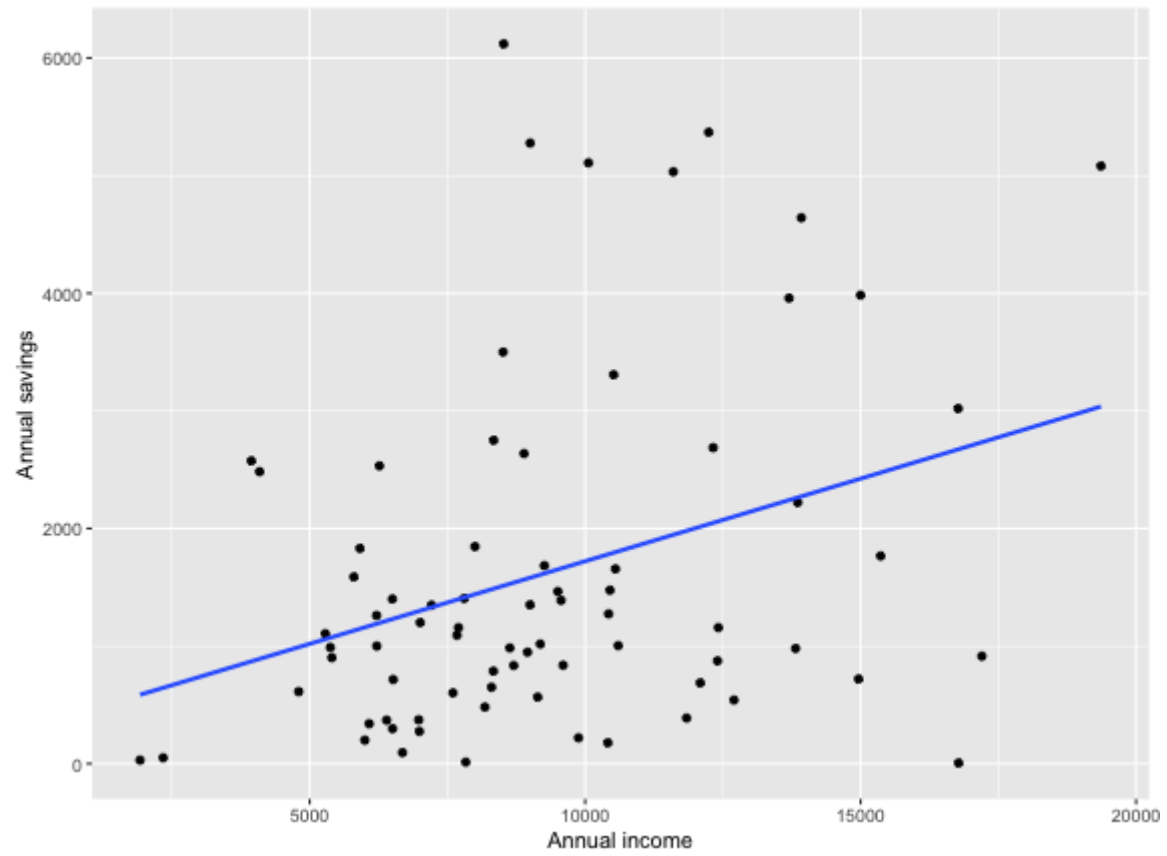
$$R^2 \approx 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

```
# Load packages
library(dplyr)
library(ggplot2)
library(wooldridge)

# Load the sample
data("saving")

# Only use positive values of saving, which are smaller than income
saving <- saving %>%
  filter(sav > 0,
         inc < 20000,
         sav < inc)

# Plot
ggplot(saving, aes(x = inc, y = sav)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Annual income", y = "Annual savings")
```



8.2 Heteroskedasticity-robust inference after OLS estimation

- Formulas for OLS standard errors and related statistics have been developed that are robust to heteroskedasticity of unknown form.
- All formulas are only valid in large samples.
- Formula for heteroskedasticity-robust OLS standard error.

White/ Hubber/ Eicker standard errors: they involve the squared residuals from the regression and from a regression of x_j on all other explanatory variables

$$\widehat{\text{Var}} \left(\hat{\beta}_j \right) = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

- Using these formulas, the usual **t test** is valid asymptotically.
- The usual **F statistic** does not work under heteroskedasticity, but heteroskedasticity robust versions are available in most software.

a) Heteroskedasticity-Robust LM statistic

Not all regression packages compute F statistics that are robust to heteroskedasticity.

It is sometimes convenient to have a way of obtaining **a test of multiple exclusion restrictions** that is robust to heteroskedasticity and does not require a particular kind of econometric software--> **Heteroskedasticity Robust LM statistic**

1. Obtain the residuals \tilde{u} from the restricted model.
2. Regress each of the independent variables excluded under the null on all of the included independent variables; if there are q excluded variables, this leads to q sets of residuals $(\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_q)$
3. Find the products (multiplication) of each \tilde{r}_j and \tilde{u} (for all observations).
4. Run the regression of 1 on r_1u, r_2u, \dots, r_qu , without an intercept.
5. The heteroskedasticity-robust LM statistic is $n - SSR_1$.
 - where SSR_1 is just the usual sum of squared residuals from this final regression.
 - Under H_0 , LM is distributed approximately as χ_q^2 .

8.3 Testing for Heteroskedasticity

Breusch-Pagan test for heteroskedasticity

$$H_0 : \text{Var}(u \mid x_1, x_2, \dots, x_k) = \text{Var}(u \mid \mathbf{x}) = \sigma^2$$

$$\text{Var}(u \mid \mathbf{x}) = E(u^2 \mid \mathbf{x}) - [E(u \mid \mathbf{x})]^2 = E(u^2 \mid \mathbf{x})$$

$$\Rightarrow E(u^2 \mid x_1, \dots, x_k) = E(u^2) = \sigma^2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + u$$

Regress squared residuals on all independent variables and test whether this regression has explanatory power

$$\widehat{u}^2 = \delta_0 + \delta_1 x_1 + \dots + \delta_k x_k + \text{error}$$

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_k = 0$$

Breusch-Pagan test for heteroskedasticity

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + \text{error}$$

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$$

A large test statistic (= a high R-squared) is evidence against the null hypothesis that the expected value of u^2 is unrelated to the explanatory variables.

$$F = \frac{R_{\hat{u}^2/k}^2}{1 - R_{\hat{u}^2/(n-k-1)}^2}$$

Alternative test statistic (= Lagrange multiplier statistic, LM). Again, high values of the test statistic (=high R-squared) lead to rejection of the null hypothesis.

$$LM = n \cdot R_{\hat{u}^2}^2 \sim \chi_k^2$$

Example: Heteroskedasticity in housing price equations

$$\widehat{price} = -21.77 + .0021 \text{ lotsize} + .123 \text{ sqrft} + 13.85 \text{ bdrms}$$

(29.48) (.0006) (.013) (9.01)

$$\Rightarrow R_{\hat{u}^2}^2 = .1601, p\text{-value}_F = .002, p\text{-value}_{LM} = .0028 \leftarrow \text{Heteroskedasticity}$$

$$\widehat{\log(price)} = -1.30 + .168 \log(lotsize) + .700 \log(sqrft) + .037 \text{ bdrms}$$

(.65) (.038) (.093) (.028)

$$\Rightarrow R_{\hat{u}^2}^2 = .0480, p\text{-value}_F = .245, p\text{-value}_{LM} = .2390$$

In the logarithmic specification, homoskedasticity cannot be rejected

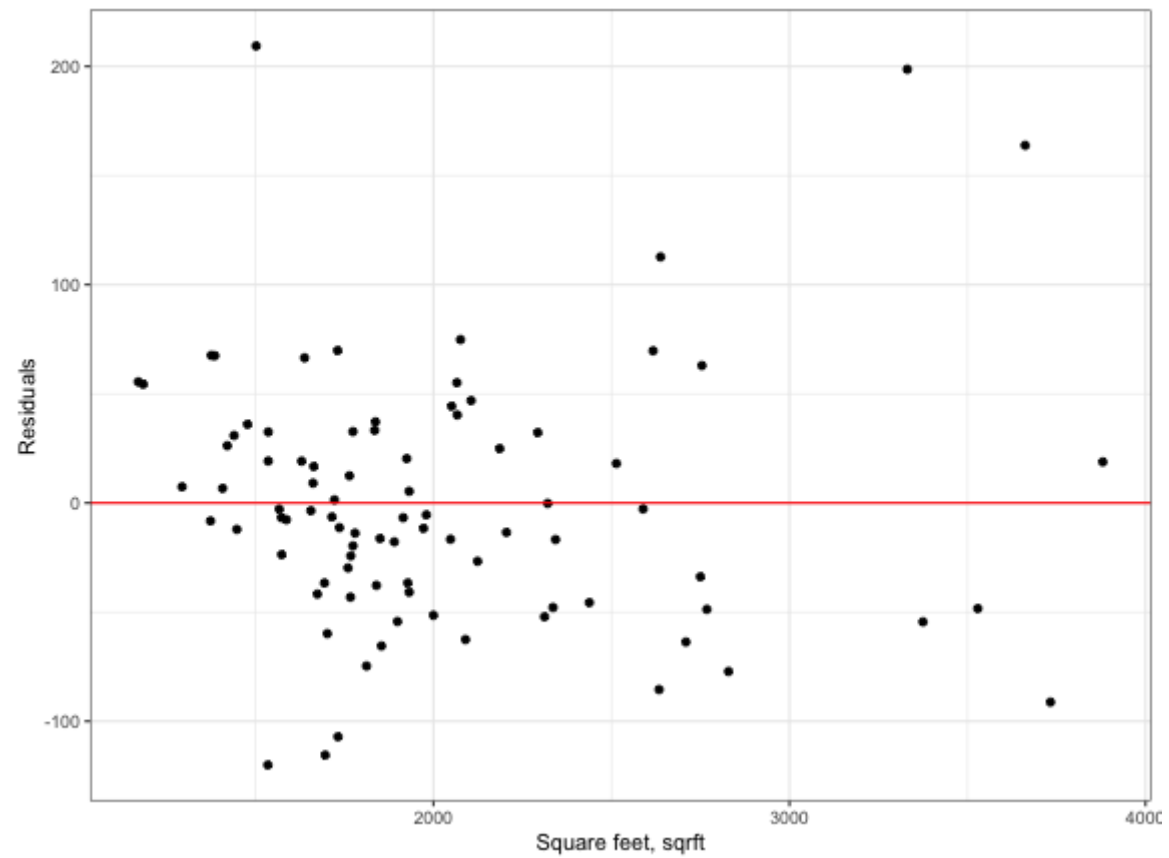
- If p-value > 0.05 then homoskedasticity
- If p-value < 0.05 then heteroskedasticity

```
data(hprice1, package='wooldridge')  
# Regression model for price  
model_a <- feols(price ~ lotsize + sqrft + bdrms, hprice1)  
#You can use clustered or robust uncertainty estimates by modifying the vcov parameter.  
#This function accepts 5 different types of input. You can use a string or a vector of strings:  
modelsummary(model_a, vcov = c("classical", "robust", "stata"), gof_omit = ".*")  
#modelsummary(models, vcov = "robust")
```

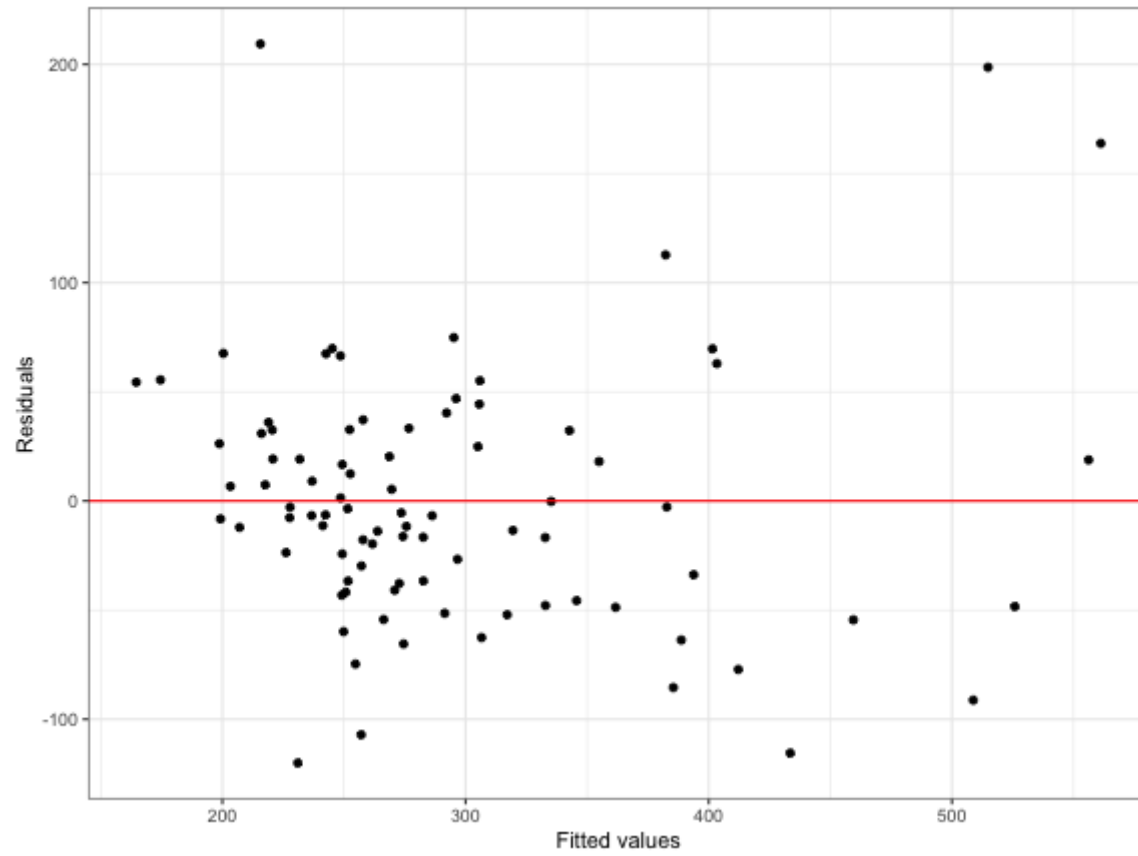
	Model 1	Model 2	Model 3
(Intercept)	-21.770	-21.770	-21.770
	(29.475)	(41.033)	(37.138)
lotsize	0.002	0.002	0.002
	(0.001)	(0.007)	(0.001)
sqrft	0.123	0.123	0.123
	(0.013)	(0.041)	(0.018)
bdrms	13.853	13.853	13.853
	(9.010)	(11.562)	(8.479)

```
data(hprice1, package='wooldridge')  
  
# Regression model for price  
model_0 <- lm(price ~ lotsize + sqrft + bdrms, hprice1)  
  
modelsummary(model_0, output = "markdown")  
  
hprice1 %<>% mutate(uhat = resid(model_0))
```

```
# Graph of residuals against independent variable  
ggplot(data = hprice1, mapping = aes(x = sqrft, y = uhat)) +  
  theme_bw() +  
  geom_point() +  
  geom_hline(yintercept = 0, col = 'red') +  
  labs(y = 'Residuals', x = 'Square feet, sqrft')
```

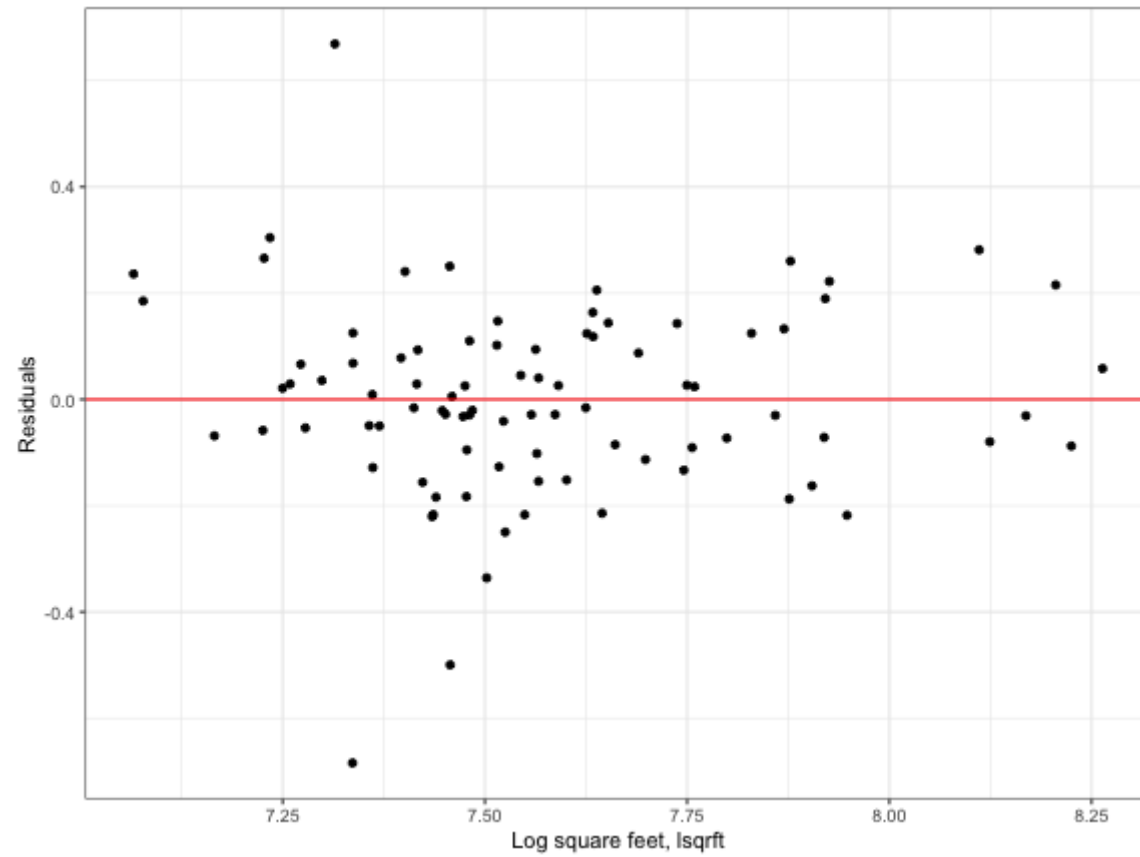



```
# Graph of residuals against fitted values
hprice1 %<>% mutate(yhat = fitted(model_0))
ggplot(data = hprice1, mapping = aes(x = yhat, y = uhat)) +
  theme_bw() +
  geom_point() +
  geom_hline(yintercept = 0, col = 'red') +
  labs(y = 'Residuals', x = 'Fitted values')
```

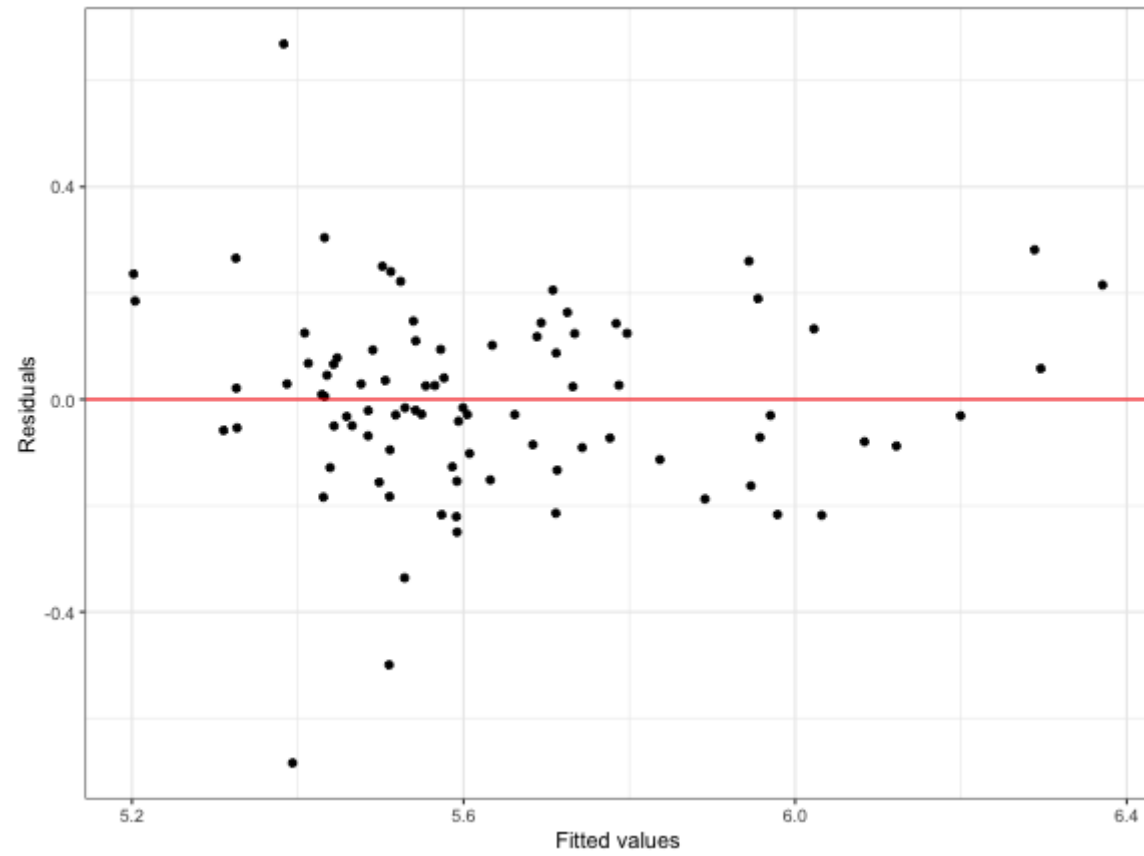


```
# Regression model for lprice  
model_1 <- lm(lprice ~ llotsize + lsqrft + bdrms, hprice1)  
hprice1 %<>% mutate(uhat1 = resid(model_1))  
  
#summary ( lm(lprice ~ llotsize + lsqrft + bdrms, hprice1))  
#modelsummary(model_1,output = "markdown")
```

```
# Graph of residuals against independent variable  
ggplot(hprice1) +  
  theme_bw() +  
  geom_point(aes(x = lsqrft, y = uhat1)) +  
  geom_hline(yintercept = 0, col = 'red') +  
  labs(y = 'Residuals', x = 'Log square feet, lsqrft')
```



```
# Graph of residuals against fitted values
hprice1 %<>% mutate(yhat1 = fitted(model_1))
ggplot(data = hprice1, mapping = aes(x = yhat1, y = uhat1)) +
  theme_bw() +
  geom_point() +
  geom_hline(yintercept = 0, col = 'red') +
  labs(y = 'Residuals', x = 'Fitted values')
```



```
data(hprice1, package='wooldridge')
model_0 <- lm(price ~ lotsize + sqrft + bdrms, hprice1)
hprice1 %<>% mutate(uhat = resid(model_0), yhat =fitted(model_0))

# Get residuals(uhat) and predicted values(yhat), and square them
hprice1 %<>% mutate(uhatsq = uhat^2, yhatsq = yhat^2)

# Regression for Breusch-Pagan test
model_BP <- lm(uhatsq ~ lotsize + sqrft + bdrms, hprice1)
#summary(model_BP)

# Number of independent variables k1
(k1 <- model_BP$rank - 1)
```

```
## [1] 3
```



```
# F-test and LM-test for heteroscedasticity  
(r2 <- summary(model_BP)$r.squared) # R-squared
```

```
## [1] 0.1601407
```

```
(n <- nobs(model_BP)) # number of observations
```

```
## [1] 88
```

```
( F_stat <- (r2/k1) / ((1-r2)/(n-k1-1)) ) # F-statistic
```

```
## [1] 5.338919
```

```
( F_pval <- pf(F_stat, k1, n-k1-1, lower.tail = FALSE) ) # p-value
```

```
## [1] 0.002047744
```

```
( LM_stat <- n * r2 ) # LM-statistic
```

```
## [1] 14.09239
```

```
( LM_pval <- pchisq(q = LM_stat, df = k1, lower.tail = FALSE)) # p-value
```

```
## [1] 0.00278206
```

a) The White Test for Heteroskedasticity

Regress squared residuals on *all independent variables, their squares, and all interaction terms*

$$\begin{aligned}\hat{u}^2 = & \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ & + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + \text{error}\end{aligned}$$

The White test detects more general deviations from heteroskedasticity than the Breusch- Pagan test

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_9 = 0$$

$$LM = n \cdot R_{\hat{u}^2}^2 \sim \chi_9^2$$

Disadvantage

- Including all squares and interactions leads to a large number of estimated parameters (e.g. k=6 leads to 27 parameters to be estimated).

```
# White test -----  
  
# Generate squares and interaction of independent variables  
hprice1 %<>% mutate(lotsizesq = lotsize^2,  
                    sqrftsq = sqrft^2,  
                    bdrmssq = bdrms^2,  
                    lotsizeXsqrft = lotsize*sqrft,  
                    lotsizeXbdrms = lotsize*bdrms,  
                    sqrftXbdrms = sqrft*bdrms)  
  
# Regression for White test  
model_White <- lm(uhatsq ~ lotsize + sqrft + bdrms + lotsizesq + sqrftsq +  
bdrmssq + lotsizeXsqrft + lotsizeXbdrms + sqrftXbdrms, hprice1)  
#summary(model_White)  
  
(k2 <- model_White$rank - 1) # Number of independent variables k2
```

```
## [1] 9
```

```
# F-test and LM-test for heteroscedasticity  
(r2 <- summary(model_White)$r.squared) # R-squared
```

```
## [1] 0.3833143
```

```
(n <- nobs(model_White)) # number of observations
```

```
# White test -----
```

```
( F_stat <- (r2/k2) / ((1-r2)/(n-k2-1)) ) # F-statistic
```

```
## [1] 5.386953
```

```
( F_pval <- pf(F_stat, k2, n-k2-1, lower.tail = FALSE) ) # p-value
```

```
## [1] 1.012939e-05
```

```
( LM_stat <- n * r2 ) # LM-statistic
```

```
## [1] 33.73166
```

```
( LM_pval <- pchisq(q = LM_stat, df = k2, lower.tail = FALSE)) # p-value
```

```
## [1] 9.95294e-05
```

Alternative form of the White test

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \text{error}$$

This regression indirectly tests the dependence of the squared residuals on the explanatory variables, their squares, and interactions, because the predicted values of y and its square implicitly constrain all of these terms

$$H_0 : \delta_1 = \delta_2 = 0, \quad LM = n \cdot R_{\hat{u}^2}^2 \sim \chi_2^2$$

$$R_{\hat{u}^2}^2 = .0392, \quad LM = 88(.0392) \approx 3.45, \quad p\text{-value}_{LM} = .178$$

```
# Alternative White test ----
# Regression for alternative White test
model_Alt <- lm(uhatsq ~ yhat + yhatsq, hprice1)

# Get residuals and predicted values, and square them
hprice1 %<>% mutate(uhat1 = resid(model_Alt),
                  yhat1 = fitted(model_Alt),
                  uhat1sq = uhat1^2,
                  yhat1sq = yhat1^2)

# Number of independent variables k3
(k3 <- model_Alt$rank - 1)
```

```
## [1] 2
```

```
# F-test and LM-test for heteroscedasticity  
(r2 <- summary(model_Alt)$r.squared) # R-squared
```

```
## [1] 0.1848684
```

```
(n <- nobs(model_Alt)) # number of observations
```

```
## [1] 88
```

```
( F_stat <- (r2/k3) / ((1-r2)/(n-k3-1)) ) # F-statistic
```

```
## [1] 9.638819
```

```
( F_pval <- pf(F_stat, k3, n-k3-1, lower.tail = FALSE) ) # p-value
```

```
## [1] 0.0001687248
```

```
( LM_stat <- n * r2 ) # LM-statistic
```

```
## [1] 16.26842
```

```
( LM_pval <- pchisq(q = LM_stat, df = k3, lower.tail = FALSE)) # p-value
```

```
## [1] 0.0002933311
```

```
price<- list(  
  "Model for price "= model_0,  
  "Breusch-Pagan Test (uhatsq)"=model_BP,  
  "White Test (uhatsq) "=model_White,  
  "Alternative White Test (uhatsq)"=model_Alt)  
  
#create one table  
modelsummary(price,stars = TRUE,fmt = 2,gof_omit = "R2 | R2 Within |AIC|BIC|Log.Lik.|R2 Pseudo")
```


	Model for price	Breusch-Pagan Test (uhatsq)	White Test (uhatsq)	Alternative White Test (uhatsq)
(Intercept)	-21.77	-5522.79+	15626.24	19071.59*
	(29.48)	(3259.48)	(11369.41)	(8876.23)
lotsize	0.00**	0.20**	-1.86**	
	(0.00)	(0.07)	(0.64)	
sqrft	0.12***	1.69	-2.67	
	(0.01)	(1.46)	(8.66)	
bdrms	13.85	1041.76	-1982.84	
	(9.01)	(996.38)	(5438.48)	
lotsizesq			0.00	
			(0.00)	
sqrftsq			0.00	
			(0.00)	
bdrmsq			289.75	
			(758.83)	

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

	Alternative White Test (uhatsq)
(Intercept)	19071.59*
	(8876.23)
yhat	-119.66*
	(53.32)
yhatsq	0.21**
	(0.07)
Num.Obs.	88
R2	0.185
F	9.639
RMSE	6480.06
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Heteroskedasticity tests summary for *price* regressions

	Breusch-Pagan test	White test	Alternative White test
Observations n	88	88	88
R-squared $R_{\hat{u}^2}^2$	0.160	0.383	0.185
k	3	9	2
F-stat	$(0.16/3)/((1-0.16)/(88-3-1))=5.34$	$(0.383/9)/((1-0.383)/(88-9-1))=5.39$	$(0.185/2)/((1-0.185)/(88-2-1))=9.64$
P-value for F-test	0.002	0.00001	0.0002
LM-stat	$88*0.16=14.09$	$88*0.383=33.73$	$88*0.185=16.27$
P-value for LM test	0.003	0.0001	0.0003
Conclusion	heteroscedasticity	heteroscedasticity	heteroscedasticity

All tests show heteroscedasticity in price. The regression for price needs correction for heteroscedasticity.

Heteroskedasticity tests summary for *log of price* regressions

	Breusch-Pagan test	White test	Alternative White test
Observations n	88	88	88
R-squared $R_{\hat{u}^2}^2$	0.040	0.082	0.012
k	3	9	2
F-stat	$(0.04/3)/((1-0.04)/(88-3-1))=1.17$	$(0.082/9)/((1-0.082)/(88-9-1))=0.77$	$(0.012/2)/((1-0.012)/(88-2-1))=0.53$
P-value for F-test	0.32	0.64	0.59
LM-stat	$88*0.04=3.54$	$88*0.082=7.19$	$88*0.012=1.08$
P-value for LM test	0.32	0.61	0.58
Conclusion	homoscedasticity	homoscedasticity	homoscedasticity

All tests show homoscedasticity in log price. The regression for log price does not need correction for heteroscedasticity.

8.4 Weighted Least Squares Estimation

Before the development of heteroskedasticity-robust statistics, the response to a finding of heteroskedasticity was to specify its form and use a weighted least squares method, which we develop in this section

If we have correctly specified the form of the variance (as a function of explanatory variables), then weighted least squares (WLS) is more efficient than OLS, and WLS leads to new t and F statistics that have t and F distributions.

We will also discuss the implications of using the wrong form of the variance in the WLS procedure.

a) The Heteroskedasticity Is Known Up to a Multiplicative Constant

$$\text{Var}(u_i \mid \mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i), h(\mathbf{x}_i) = h_i > 0$$

The functional form of the heteroskedasticity is known: $h(\mathbf{x}_i) = h_i > 0$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

$$\left[\frac{y_i}{\sqrt{h_i}} \right] = \beta_0 \left[\frac{1}{\sqrt{h_i}} \right] + \beta_1 \left[\frac{x_{i1}}{\sqrt{h_i}} \right] + \cdots + \beta_k \left[\frac{x_{ik}}{\sqrt{h_i}} \right] + \left[\frac{u_i}{\sqrt{h_i}} \right]$$

Transformed model

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + u_i^*$$

Example: Savings and income

$$\text{sav}_i = \beta_0 + \beta_1 \text{inc}_i + u_i, \text{Var}(u_i \mid \text{inc}_i) = \sigma^2 \text{inc}_i$$

Note that this regression model has no intercept

$$\left[\frac{\text{sav}_i}{\sqrt{\text{inc}_i}} \right] = \beta_0 \left[\frac{1}{\sqrt{\text{inc}_i}} \right] + \beta_1 \left[\frac{\text{inc}_i}{\sqrt{\text{inc}_i}} \right] + u_i^*$$

The transformed model is homoskedastic

$$E(u_i^{*2} \mid \mathbf{x}_i) = E\left[\left(\frac{u_i}{\sqrt{h_i}}\right)^2 \mid \mathbf{x}_i\right] = \frac{E(u_i^2 \mid \mathbf{x})}{h_i} = \frac{\sigma^2 h_i}{h_i} = \sigma^2$$

If the other Gauss-Markov assumptions hold as well, OLS applied to the transformed model is the best linear unbiased estimator.

OLS in the transformed model is weighted least squares (WLS)

$$\min \sum_{i=1}^n \left(\left[\frac{y_i}{\sqrt{h_i}} \right] - b_0 \left[\frac{1}{\sqrt{h_i}} \right] - b_1 \left[\frac{x_{i1}}{\sqrt{h_i}} \right] - \cdots - b_k \left[\frac{x_{ik}}{\sqrt{h_i}} \right] \right)^2$$

$$\min \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2 / h_i$$

Observations with a large variance get a smaller weight in the optimization problem

Why is WLS more efficient than OLS in the original model?

- Observations with a large variance are less informative than observations with small variance and therefore should get less weight.

WLS is a special case of generalized least squares (GLS)


```

# When the heteroscedasticity form is known  $\text{var}(u|x)=(\sigma^2)*(sqrft)$ ,
# use WLS with  $\text{weight}=1/sqrft$ .

# WLS: estimate model with  $\text{weight}=1/sqrft$ 
model_WLS1 <- lm(formula = price ~ lotsize + sqrft + bdrms,
                 data = hprice1, weights = 1/sqrft)

# Multiply all variables and the constant by  $1/sqrt(sqrft)$ 
hprice1 %<>% mutate(pricestar = price/sqrt(sqrft),
                  lotsizestar = lotsize/sqrt(sqrft),
                  sqrftstar = sqrft/sqrt(sqrft),
                  bdrmsstar = bdrms/sqrt(sqrft),
                  constantstar = 1/sqrt(sqrft))

# WLS: estimate model with transformed variables by OLS
model_WLS2 <- lm(pricestar ~ 0 + constantstar + lotsizestar + sqrftstar + bdrmsstar,
                 hprice1)

```

	Model 1	Model 2
(Intercept)	4.199	
	(29.698)	
lotsize	0.002**	
	(0.001)	
sqrft	0.118***	
	(0.014)	
bdrms	10.607	
	(8.659)	
constantstar		4.199
		(29.698)
lotsizestar		0.002**
		(0.001)
sqrftstar		0.118***
		(0.014)
bdrmsstar		10.607
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

Important special case of heteroskedasticity

If the observations are reported as averages at the city/county/state/-country/firm level, they should be weighted by the size of the unit

Average contribution to pension plan in firm i Average earnings and age in firm i Percentage firm contributes to plan Heteroskedastic error term

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earns}_i + \beta_2 \overline{age}_i + \beta_3 \overline{mrate}_i + \overline{u}_i$$

$\Rightarrow Var(\overline{u}_i) = Var\left(\frac{1}{m_i} \sum_{e=1}^{m_i} u_{i,e}\right) = \sigma^2/m_i$ Error variance if errors are homoskedastic at the individual-level

- If errors are homoskedastic at the individual-level, WLS with weights equal to firm size m_i should be used.
- If the assumption of homoskedasticity at the individual-level is not exactly right, one can calculate robust standard errors after WLS (i.e. for the transformed model).

b) The heteroskedasticity function must be estimated: Feasible GLS

- Unknown heteroskedasticity function (feasible GLS)
- In many cases is difficult to find the function $h(x)$ so we can use data to estimate it $\widehat{h(x)}$

Assumed general form of heteroskedasticity

- exponential function is used to ensure positivity

$$\text{Var}(u \mid \mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k) = \sigma^2 h(\mathbf{x})$$

Multiplicative error (assumption: independent of the explanatory variables)

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k) \cdot v$$

$$\Rightarrow \log(u^2) = \alpha_0 + \delta_1 x_1 + \cdots + \delta_k x_k + e$$

Use inverse values of the estimated heteroskedasticity function as weights in WLS

$$\log(\hat{u}^2) = \hat{\alpha}_0 + \hat{\delta}_1 x_1 + \cdots + \hat{\delta}_k x_k + \text{error}$$

$$\Rightarrow \hat{h}_i = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_1 + \cdots + \hat{\delta}_k x_k)$$

Steps to compute a Feasible GLS

1. Run the regression of y on x_1, x_2, \dots, x_k and obtain the residuals, \hat{u} .
2. Create $\log(\hat{u}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression in equation $\log(\hat{u}^2)$ on x_1, x_2, \dots, x_k and obtain the fitted values, \hat{g} .
4. Exponentiate the fitted values $\hat{h} = \exp(\hat{g})$.
5. Estimate the equation $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$ by WLS, using weights $1/\hat{h}$.
 - If instead we first transform all variables and run OLS, each variable gets multiplied by $1/\sqrt{\hat{h}_i}$, including the intercept.

Example: Demand for cigarettes

Estimation by OLS

Cigarettes smoked per day Logged income and cigarette price

$$\widehat{cigs} = - \begin{matrix} 3.64 \\ (24.08) \end{matrix} + \begin{matrix} .880 \\ (.728) \end{matrix} \log(income) - \begin{matrix} .751 \\ (5.773) \end{matrix} \log(cigpric)$$
$$- \begin{matrix} .501 \\ (.167) \end{matrix} educ - \begin{matrix} .771 \\ (.160) \end{matrix} age - \begin{matrix} .0090 \\ (.0017) \end{matrix} age^2 - \begin{matrix} 2.83 \\ (1.11) \end{matrix} restaurn$$

Smoking restrictions in restaurants

$$n = 807, R^2 = .0526, p\text{-value}_{Breusch-Pagan} = .000$$


Reject homoskedasticity

Example: Demand for cigarettes

Estimation by FGLS

$$\begin{aligned}\widehat{cigs} = & - \frac{5.64}{(17.80)} + \frac{1.30}{(.44)} \log(income) - \frac{2.94}{(4.46)} \log(cigpric) \\ & - .463 \frac{educ}{(.120)} + .482 \frac{age}{(.097)} - .0056 \frac{age^2}{(.0009)} - 3.46 \frac{restaurn}{(.80)}\end{aligned}$$

n = 807, R^2 = .1134

Now statistically significant


Discussion:

- The income elasticity is now statistically significant; other coefficients are also more precisely estimated (without changing qualitative results).

```

# Feasible GLS (FGLS) -----

# When the heteroscedasticity form is not known,
#  $\text{var}(u|x) = \sigma^2(\delta_0 + \delta_1 \text{lotsize} + \delta_2 \text{sqrft} + \delta_3 \text{bdrms})$ 
# estimate hhat and use WLS with weight=1/hhat.

# Heteroscedasticity form, estimate hhat
# model_0 <- lm(price ~ lotsize + sqrft + bdrms, hprice1)
summary(model_0)
hprice1 %<>% mutate(u = resid(model_0),
                    g = log(u^2))
model_g <- lm(g ~ lotsize + sqrft + bdrms, hprice1)
hprice1 %<>% mutate(ghat = fitted(model_g),
                    hhat = exp(ghat))

# FGLS: estimate model using WLS with weight=1/hhat
model_FGLS1 <- lm(formula = price ~ lotsize + sqrft + bdrms,
                  data = hprice1,
                  weights = 1/hhat)
#summary(model_FGLS1)

```



```
##
## Call:
## lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.026  -38.530   -6.555   32.323  209.376
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
## lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
## sqrft       1.228e-01  1.324e-02   9.275 1.66e-14 ***
## bdrms       1.385e+01  9.010e+00   1.537  0.12795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.83 on 84 degrees of freedom
## Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6607
## F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```

```
# Multiply all variables and the constant by 1/sqrt(hhat)
hprice1 %<>% mutate(pricestar1 = price/sqrt(hhat),
                    lotsizestar1 = lotsize/sqrt(hhat),
                    sqrftstar1 = sqrft/sqrt(hhat),
                    bdrmsstar1 = bdrms/sqrt(hhat),
                    constantstar1 = 1/sqrt(hhat))

# FGLS: estimate model with transformed variables by OLS
model_FGLS2 <- lm(pricestar1 ~ 0 + constantstar1 + lotsizestar1 + sqrftstar1 +
                  bdrmsstar1, hprice1)

#summary(model_FGLS2)
```



```
fgls<-list(model_FGLS1,model_FGLS2)
modelsummary(fgls,stars = TRUE,fmt = 3,gof_omit = "R2 | R2 Within |AIC|BIC|Log.Lik.|R2 Pseudo")
```

	Model 1	Model 2
(Intercept)	45.912	
	(30.824)	
lotsize	0.004**	
	(0.001)	
sqrft	0.092***	
	(0.015)	
bdrms	6.175	
	(8.894)	
constantstar1		45.912
		(30.824)
lotsizestar1		0.004**
		(0.001)
sqrftstar1		0.092***
		(0.015)
bdrmsstar1		6.175
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

c) What if the assumed heteroskedasticity function is wrong?

- If the heteroskedasticity function is misspecified, WLS is still consistent under MLR.1 – MLR.4, but robust standard errors should be computed.
- If OLS and WLS produce very different estimates, this typically indicates that some other assumptions (e.g. MLR.4) are wrong.
- If there is strong heteroskedasticity, it is still often better to use a wrong form of heteroskedasticity in order to increase efficiency.