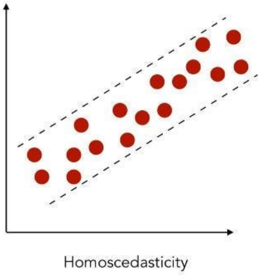
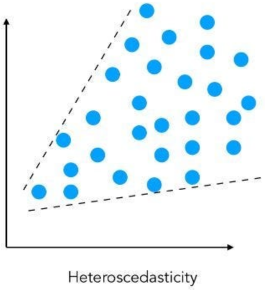
Fundamentals of Econometrics

Lecture 7: Heteroskedasticity





Section 1

Heteroskedasticity

Model Assumptions: Classical Linear Models

In order to have unbiased and consistent estimates, the classical linear model assumptions must hold:

1. **Linearity**: The true model is linear in parameters.
2. **Random Sampling**: The data are a random sample from the population.
3. **No Perfect Collinearity**: The regressors are not perfectly collinear.
4. **Zero Conditional Mean**: $E(u|x) = 0$.
5. **Homoskedasticity**: $Var(u|x) = \sigma^2$.
6. **Normality**: $u|x \sim N(0, \sigma^2)$.

Thought

How do we know if any assumption is violated? And, what to do if they are?

Heteroskedasticity

- **Heteroskedasticity** is the violation of the homoskedasticity assumption.
- It occurs when the variance of the error term varies for different values of $\mathbf{x}$.

Consequences

- OLS is still unbiased and consistent under heteroskedasticity.
- Interpretations of R^2 and $\bar{R}^2$ are not changed: $R^2 = 1 - \sigma_u^2/\sigma_y^2$ where σ_y^2 is the **unconditional** error variance. Heteroskedasticity affects the **conditional** error variance.
- Main issue is inference:
  - Variance formulas for OLS estimator are no longer valid.
  - Usual F-tests are no longer valid.
  - OLS is no longer BLUE. There might be more efficient linear estimators.

Robust Standard Errors

Consider the univariate model:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

If Assumptions 1-4 hold but 5 does not, then

$$Var(u_i|x_i) = \sigma_i^2$$

The OLS estimator is given by:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The variance of the OLS estimator is now given by:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{SST_x^2}; \quad SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1}$$

Under homoskedasticity, $\sigma_i^2 = \sigma^2$ for all i. In this case, $Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x^2}$.

Robust Standard Errors

- Since the standard error of $\hat{\beta}_1$ is based on directly estimating $var(\hat{\beta}_1)$, we will need a way to estimate Equation (1) when σ_i^2 under heteroskedasticity.
- White (1980) proposed a consistent estimator for the variance of the OLS estimator under any form of heteroskedasticity:

$$\widehat{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$$

  where $\hat{u}_i$ is the OLS residual.
- Consistent means that as $n \to \infty$, $\widehat{Var}(\hat{\beta}_1) \to Var(\hat{\beta}_1)$.

Robust Standard Errors

In a multiple regression model, the White estimator for the variance of the OLS estimator is given by:

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

where $\hat{r_{ij}}$ is the ith residual from regressing x_j on all other independent variables, and SSR_j is the sum of squared residuals from this regression. Recall the concept of **partialling out** from earlier.

- The square root of Equation (2) is referred to as the *heteroskedasticity-robust standard error* for $\hat{\beta}_j$.

  Usual covariance matrix:

  $$\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

  Robust covariance matrix:

  $$\widehat{Var}(\hat{\beta}) = (X'X)^{-1} X' \Omega' X (X'X)^{-1}$$

  where $\Omega = diag(\hat{u}_1^2, \ldots, \hat{u}_n^2)$

Table 1:

	Dependent variable:	
	wage	
	OLS	Robust SE
	(1)	(2)
educ	0.556***	0.556***
	(0.050)	(0.061)
exper	0.255***	0.255***
	(0.035)	(0.033)
expersq	−0.004***	−0.004***
	(0.001)	(0.001)
female	−2.110***	−2.110***
	(0.263)	(0.250)
Constant	−2.320***	−2.320***
	(0.739)	(0.818)
Observations	526	526
R^2	0.350	0.350
Adjusted R^2	0.345	0.345
Residual Std. Error (df = 521)	2.990	2.990
F Statistic (df = 4; 521)	70.200***	70.200***

```
m1 <- lm(wage ~ educ + exper + expersq + female, data = wage1)
# Heteroskedasticity-robust standard errors
# require("sandwich"); require ("lmtest")
# coeftest(m1, vcov = vcovHC(m1, type = "HC1"))

cov1 <- vcovHC(m1, type = "HC0") # Robust covariance matrix
robust.se <- sqrt(diag(cov1)) # Robust standard errors

stargazer(m1, m1, se = list(NULL,robust.se), font.size = "scri
          header = FALSE, column.labels = c("OLS", "Robust SE"
```

```r
# require("ggfortify")
autoplot(m1, which = c(1:3,5), ncol = 2, label.size = 3)
```

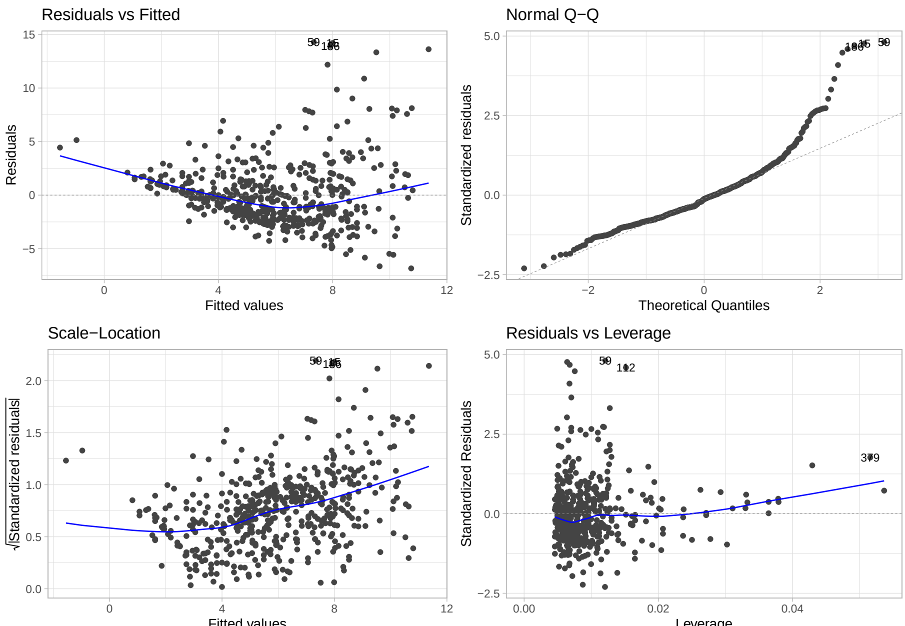
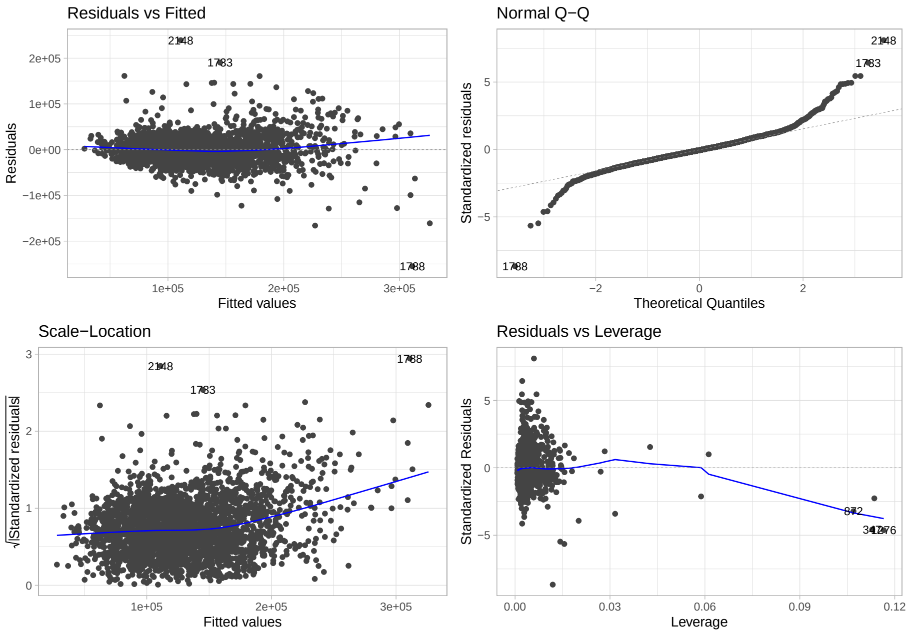


Table 2:

	Dependent Variable: saleprice			
	OLS	HC0	HC1	HC2
	(1)	(2)	(3)	(4)
gdistance	-1.610^{***}	-1.610^{***}	-1.610^{***}	-1.610^{***}
	(0.177)	(0.171)	(0.172)	(0.172)
wdistance	0.665^{***}	0.665^{***}	0.665^{***}	0.665^{***}
	(0.159)	(0.148)	(0.148)	(0.149)
cdistance	2.440^{***}	2.440^{**}	2.440^{**}	2.440^{**}
	(0.929)	(1.040)	(1.040)	(1.040)
bathrooms	2,460.000	2,460.000	2,460.000	2,460.000
	(1,759.000)	(2,000.000)	(2,004.000)	(2,021.000)
bedrooms	$-5{,}855.000^{***}$	$-5{,}855.000^{***}$	$-5{,}855.000^{***}$	$-5{,}855.000^{**}$
	(1,164.000)	(1,335.000)	(1,337.000)	(1,340.000)
sqftbuilding	72.800^{***}	72.800^{***}	72.800^{***}	72.800^{***}
	(1.940)	(3.070)	(3.080)	(3.100)
sqftlot	0.506^{***}	0.506^{***}	0.506^{***}	0.506^{***}
	(0.048)	(0.127)	(0.127)	(0.133)
age	-506.000^{***}	-506.000^{***}	-506.000^{***}	-506.000^{***}
	(37.300)	(44.600)	(44.700)	(44.800)
Constant	$46{,}319.000^{***}$	$46{,}319.000^{***}$	$46{,}319.000^{***}$	$46{,}319.000^{**}$
	(4,308.000)	(5,327.000)	(5,336.000)	(5,354.000)
Observations	2,661			
R^2	0.678			


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized Residuals
Leverage
2148
1783
1788
872

Why not always use robust standard errors?

- Robust errors are easily computed in `R`. So why not use them all the time?
- For small samples, the robust standard errors from the White estimator (`HC0`), for example can be produce inaccurate test statistics.
- Other robust standard errors measures might be better for small samples and might prove more conservative.

How can we test for Heteroskedasticity?

Testing for Heteroskedasticity

- **Breusch-Pagan Test**: The null hypothesis is homoskedasticity.

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \ldots + \delta_k x_{ik} + error$$

We will regress the squared residuals on the independent variables and test whether this auxiliary regression has explanatory power.

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

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

Alternatively, we can use the LM test:

$$LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2(k)$$

- In both cases, a large $R^2_{\hat{u}^2}$ provides evidence against (rejection of) the null.

Breusch-Pagan Test

```r
h1 <- lm(price ~ lotsize + sqrft + bdrms, data = hprice1)
h1_aux <- lm(resid(h1)^2 ~ lotsize + sqrft + bdrms, data = hprice1)
r2_u <- summary(h1_aux)$r.squared
N <- nobs(h1)
k <- length(coef(h1_aux)) - 1
Fstat <- (r2_u/k) / ((1 - r2_u)/(N - k - 1))
F_crit <- qf(0.95, k, N - k - 1)
pval <- 1 - pf(Fstat, k, N - k - 1)
LM <- N * r2_u
LM_crit <- qchisq(0.95, k)
LM_pval <- 1 - pchisq(LM, k)
```

- The F-statistic is 5.339 with a p-value of 0.002.
- The F-critical value is 2.713.
- The LM statistic is 14.092 with a critical value ($\chi^2(3, 5\%)$) of 7.815.
- The LM test p-value is 0.003.
- **What do we conclude?**

Breusch-Pagan Test

How do the results above change if we used logged variables instead?

```
h2 <- lm(lprice ~ llotsize + lsqrft + bdrms, data = hprice1)
h2_aux <- lm(resid(h2)^2 ~ llotsize + lsqrft + bdrms,
             data = hprice1)
nr2_u2 <- summary(h2_aux)$r.squared*nobs(h2)
LM_pval2 <- 1 - pchisq(nr2_u2, length(coef(h2_aux)) - 1)
cat("p-value(LM) for the logged variables is",
    round(LM_pval2, 3))
```

```
## p-value(LM) for the logged variables is 0.238
```

White test for Heteroskedasticity

- Modified the Breusch-Pagan test to include quadratic and interaction terms.

Trade-offs??

- Generating all the extra terms adds lots of variables to the model thereby using up a lot of the degrees of freedom.
- Even a small number of variables can result in a large number of extra terms.
  - For example k = 6 leads to 27 parameters to be estimated.

Issues with Heteroskedasticity Tests

What do we do if we Reject the null of homoskedasticity?

White test for Heteroskedasticity

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

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_9 = 0$$
$$LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2(9)$$

Conducting the tets

- **Breusch-Pagan test**: `bptest()` in the `lmtest` package.

```
bptest(h1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  h1
## BP = 14, df = 3, p-value = 0.003
```

Conducting the tets

- **White test**: using the `bptest()` function.

```
bptest(h1, ~ lotsize + sqrft + bdrms + I(lotsize^2) +
         I(sqrft^2) +  I(bdrms^2) +  I(lotsize*sqrft) +
         I(lotsize*bdrms) + I(sqrft*bdrms), data = hprice1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  h1
## BP = 34, df = 9, p-value = 1e-04
```

Alternative Form of White Test

- We can indirectly test the dependence of the squared residuals on the explanatory variables, their squares, and their cross-products (interactions), using the predicted values of y.
- This works because the predicted values of y and its square implicitly contain all these squared and cross-product terms.

$$\hat{u}_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + error$$
$$H_0 : \delta_1 = \delta_2 = 0, \text{(Homoskedastic)}$$
$$H_1 : \text{At least one is not zero, (Heteroskedastic)}$$

The LM test is given by:

$$LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2(2)$$
$$R^2_{\hat{u}^2} = 0.0392, LM = 0.0392 \times 88 \approx 3.45$$
$$LM_{p-value} = 1 - \text{pchisq}(3.45,\ 2) = 0.178$$

```r
# Using the log house price equation
bptest(h2, ~fitted(h2) + I(fitted(h2)^2), data = hprice1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  h2
## BP = 3, df = 2, p-value = 0.2
```

Weighted Least Squares

- If the form of heteroskedasticity is known, we can use weighted least squares (WLS) to estimate the model.

Assume that

$$var(u_i|x_1) = \sigma^2 h(\mathbf{x})$$

where $h(\mathbf{x})$ is a known function of the independent variables that determines the heteroskedasticity.

- Because variances must be positive, $h(\mathbf{x}) > 0$ for all possible values of the independent variables.

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

$$\Rightarrow \frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \frac{\beta_1 x_{i1}}{\sqrt{h_i}} + \ldots + \frac{\beta_k x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

The transformed model is:

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

Example: Savings and Income

$$sav_i = \beta_0 + \beta_1 inc_i + u_i, \quad var(u_i|inc_i) = \sigma^2 inc_i$$

The transformed model is (note, no intercept):

$$\frac{sav_i}{\sqrt{inc_i}} = \beta_0 \frac{1}{\sqrt{inc_i}} + \beta_1 \frac{inc_i}{\sqrt{inc_i}} + \frac{u_i}{\sqrt{inc_i}}$$

The transformed model is now homoskedastic:

$$E(u_i^{*2}|x_i) = E\left[\left(\frac{u_i^2}{\sqrt{(inc_i)}}\right)^2 |x_i\right] = \frac{E(u_i^2|x_i)}{inc_i} = \frac{\sigma^2 \cdot inc_i}{inc_i} = \sigma^2$$

If the GM assumptions hold, OLS applied to the transformed model will be BLUE.

What is WLS doing?

$$\min \sum_{i=1}^{n} \left(\frac{y_i}{\sqrt{h_i}} - \beta_0 \frac{1}{\sqrt{h_i}} - \ldots - \beta_k \frac{x_{ik}}{\sqrt{h_i}} \right)^2$$

Obs with larger h_i will have smaller weights in the optimization problem.

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik} \right)^2 / h_i$$

- WLS is 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)

Table 3:

	Dependent Variable: nettfa			
	OLS	WLS	OLS	WLS
	(1)	(2)	(3)	(4)
inc	0.821***	0.787***	0.771***	0.740***
	(0.104)	(0.063)	(0.099)	(0.064)
I((age - 25)^2)			0.025***	0.018***
			(0.004)	(0.002)
male			2.480	1.840
			(2.060)	(1.560)
e401k			6.890***	5.190***
			(2.280)	(1.700)
Constant	−10.600***	−9.580***	−21.000***	−16.700***
	(2.530)	(1.650)	(3.490)	(1.960)
Observations		2,017		2,017
R^2		0.071		0.112
Adjusted R^2		0.070		0.110
Residual Std. Error		7.220 (df = 2015)		7.070 (df = 201
F Statistic		154.000*** (df = 1; 2015)		63.100*** (df = 4;
Note:				*p<0.1; **p<0.05; ***p

```r
single.OLS1 <- lm(nettfa ~ inc, data=k401ksubs, subset = fsize==1)
single.WLS1 <- lm(nettfa ~ inc, data=k401ksubs, subset = fsize==1,
                  weights = 1/inc)
single.OLS2 <- lm(nettfa ~ inc + I((age-25)**2) + male + e401k,
                  data=k401ksubs, subset = fsize==1)
single.WLS2 <- lm(nettfa ~ inc + I((age-25)**2) + male + e401k,
                  data=k401ksubs, subset = fsize==1,
                  weights = 1/inc)
## Robust OLS standard errors
rob.OLS1 <- coeftest(single.OLS1, vcov = vcovHC(single.OLS1, type = "HC0"))
rob.OLS2 <- coeftest(single.OLS2, vcov = vcovHC(single.OLS2, type = "HC0"))

stargazer(rob.OLS1, single.WLS1, rob.OLS2, single.WLS2,
          header = FALSE, font.size = "tiny",
          dep.var.caption = "Dependent Variable: nettfa",
          column.labels = c(rep(c("OLS", "WLS"), 2)),
          dep.var.labels.include = FALSE, model.names = FALSE)
```

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.

For example:

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

where $\overline{contrib}_i$, $\overline{earns}_i$, $\overline{age}_i$, and $\overline{mrate}_i$ are the average contribution, earnings, age, and firm contribution to the plan, respectively. The error term is assumed to be heteroskedastic.

$$\Rightarrow var(u_i) = var\left(\frac{1}{m}\sum_{i=1}^{m_i} u_{i,e}\right) = \frac{\sigma^2}{m_i}$$

where m_i is the number of observations (workers) in the i^{th} group. The error variance is assumed to be homoskedastic at the individual level.

Unknown Form of Heteroskedasticity

- If the form of heteroskedasticity is unknown, we can use **Feasible Generalized Least Squares (FGLS)**.

Option 1:

Assume a general form of heteroskedasticity:

$$var(u_i|x_i) = \sigma^2 \underbrace{exp(\delta_0 + \delta_1 x_{i1} + \ldots + \delta_k x_{ik})}_{\text{Ensures positive values}} = \sigma^2 h(x)$$

We need to estimate the δ's, to get $\hat{h}(x)$.

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_{i1} + \ldots + \delta_k x_{ik}) \cdot \nu$$

Assuming ν is independent of x, we can write:

$$log(u^2) = \alpha_0 + \delta_1 x_{i1} + \ldots + \delta_k x_{ik} + e$$

Replacing the unobserved u^2 with residuals, we run the regression:

$$\log(\hat{u}^2) = \alpha_0 + \delta_1 x_{i1} + \ldots + \delta_k x_{ik} + e$$

Collect fitted values, $\hat{g}_i$ and exponentiate to get $\hat{h}_i$.

$$\hat{h}_i = \exp(\hat{g}_(x)_i)$$

Summary of Option 1

1. Run regression of y on $x_1, x_2, \ldots, x_k$ and obtain residuals, $\hat{u}_i$.
2. Create $log(\hat{u}^2)$
3. Regress $log(\hat{u}^2)$ on $x_1, x_2, \ldots, x_k$ and obtain the fitted values, $\hat{g}_i$.
4. Exponentiate $\hat{g}_i$ to get $\hat{h}(x)$.
5. Run WLS with weights $1/\hat{h}(x)$.

Unknown Form of Heteroskedasticity

Option 2:

- As we saw in the case of the White model modifications of the Breusch-Pagan test, we can estimate h_i using the predicted and squared predicted values of y, $\hat{y}_i$ and $\hat{y}_i^2$ instead.

Summary of Step 2

1. Run regression of y on $x_1, x_2, \ldots, x_k$ and obtain residuals, $\hat{u}_i$.
2. Create $log(\hat{u}^2)$
3. Regress $log(\hat{u}^2)$ on $\hat{y}_i$ and $\hat{y}_i^2$ and obtain the fitted values, $\hat{g}_i$.
4. Exponentiate $\hat{g}_i$ to get $\hat{h}(x)$.
5. Run WLS with weights $1/\hat{h}(x)$.

```r
ols.cig <- lm(cigs ~ lincome + lcigpric + educ + age + agesq +
                restaurn, data=smoke)
summary(ols.cig) |> coef()
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.63984   24.07866  -0.151 8.80e-01
## lincome      0.88027    0.72778   1.210 2.27e-01
## lcigpric    -0.75086    5.77334  -0.130 8.97e-01
## educ        -0.50150    0.16708  -3.002 2.77e-03
## age          0.77069    0.16012   4.813 1.78e-06
## agesq       -0.00902    0.00174  -5.176 2.86e-07
## restaurn    -2.82508    1.11179  -2.541 1.12e-02
```

```r
# Test for heteroskedasticity
bptest(ols.cig)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ols.cig
## BP = 32, df = 6, p-value = 1e-05
```

```r
g_cig <- lm(log(resid(ols.cig)^2) ~ lincome + lcigpric + educ +
              age + agesq + restaurn, data = smoke) |> fitted()
h.hat_cig <- exp(g_cig)

wls.cig <- lm(cigs ~ lincome + lcigpric + educ + age + agesq + restaurn,
              data = smoke, weights = 1/h.hat_cig)
summary(wls.cig) |> coef()
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.63546   1.78e+01   0.317 7.52e-01
## lincome      1.29524   4.37e-01   2.964 3.13e-03
## lcigpric    -2.94031   4.46e+00  -0.659 5.10e-01
## educ        -0.46345   1.20e-01  -3.857 1.24e-04
## age          0.48195   9.68e-02   4.978 7.86e-07
## agesq       -0.00563   9.39e-04  -5.990 3.17e-09
## restaurn    -3.46106   7.96e-01  -4.351 1.53e-05
```

Table 4:

	Dependent Variable: cigs	
	OLS	WLS
	(1)	(2)
lincome	0.880	1.290***
	(0.728)	(0.437)
lcigpric	−0.751	−2.940
	(5.770)	(4.460)
educ	−0.501***	−0.463***
	(0.167)	(0.120)
age	0.771***	0.482***
	(0.160)	(0.097)
agesq	−0.009***	−0.006***
	(0.002)	(0.001)
restaurn	−2.830**	−3.460***
	(1.110)	(0.796)

What if our 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.
- WLS is consistent under MLR.4 but not necessarily under MLR.4'
- 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.