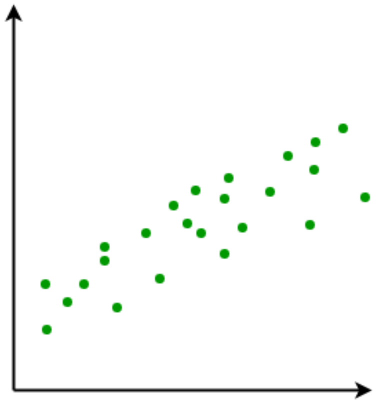
Class 24 - MRM: Heteroskedasticity (Part I)

Pedram Jahangiry

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Homoskedasticity (Equal variances)



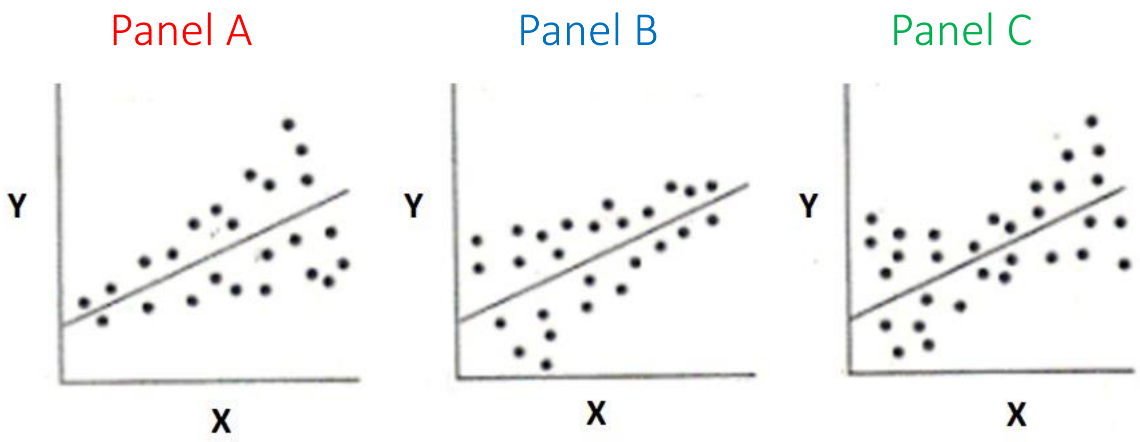
Homoscedasticity

Assumption MLR.5 **Homoskedasticity**

> The error u has the same variance given any value of the explanatory variables. In other words, $\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2$.

The value of the explanatory variable must contain no information about the variability of the unobserved factors

Heteroskedasticity (Unequal variances)



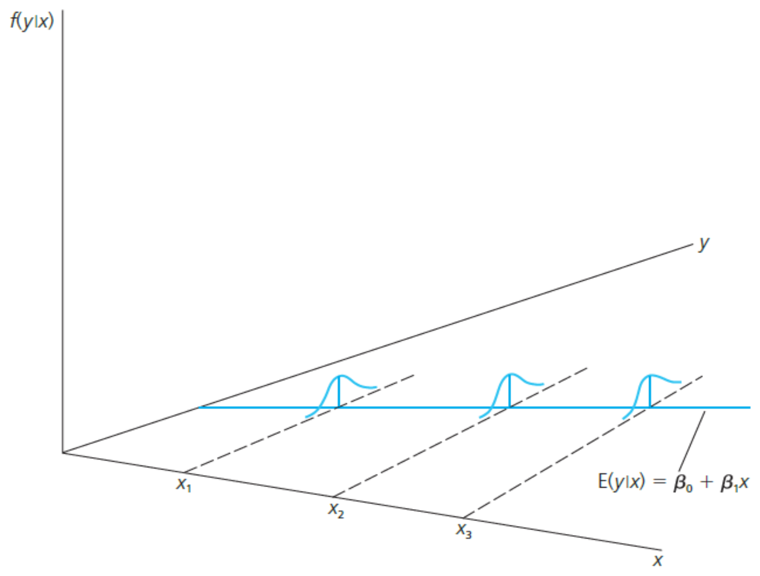
Examples:

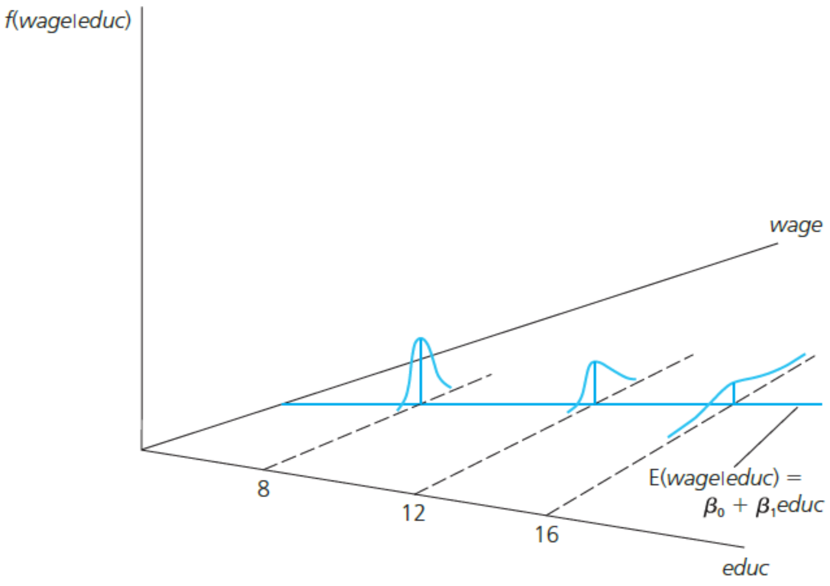
Panel A: Income vs consumption

Panel B: Learning by doing (practice vs mistakes) or better data collection methods (GDP pre/post war)

Panel C: Outliers at both ends!

Graphical illustration of homoskedasticity VS heteroskedasticity





The variance of the unobserved determinants does not depend on the value of the explanatory variable

The variance of the unobserved determinants does depend on the value of the explanatory variable

THE GAUSS-MARKOV ASSUMPTIONS

The following is a summary of the five Gauss-Markov assumptions that we used in this chapter. Remember, the first four were used to establish unbiasedness of OLS, whereas the fifth was added to derive the usual variance formulas and to conclude that OLS is best linear unbiased.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the independent variables. In other words,

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0.$$

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variables. In other words,

$$\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

THEOREM 3.2

SAMPLING VARIANCES OF THE OLS SLOPE ESTIMATORS

Under Assumptions MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)}, \quad [3.51]$$

for $j = 1, 2, \ldots, k$, where $\mathrm{SST}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

The sampling variability of the estimated regression coefficients depends on 4 things:

1. Variability of the unobserved factors (σ^2)
2. Variation in the explanatory variable $var(X_j)$ or SST_j
3. Number of observations n
4. Linear relationships among the independent variables (R^2)

Consequences of heteroskedasticity for OLS

- OLS still unbiased and consistent under heteroskedastictiy!
- Also, interpretation of R-squared is not changed
- Heteroskedasticity invalidates **variance** formulas for OLS estimators
- The usual **F tests** and **t tests** are not valid under heteroskedasticity
- Under heteroskedasticity, OLS is no longer the best linear unbiased estimator (**BLUE**); there may be more efficient linear estimators

Testing for Heteroskedasticity

There are many tests for heteroskedasticity; two popular:

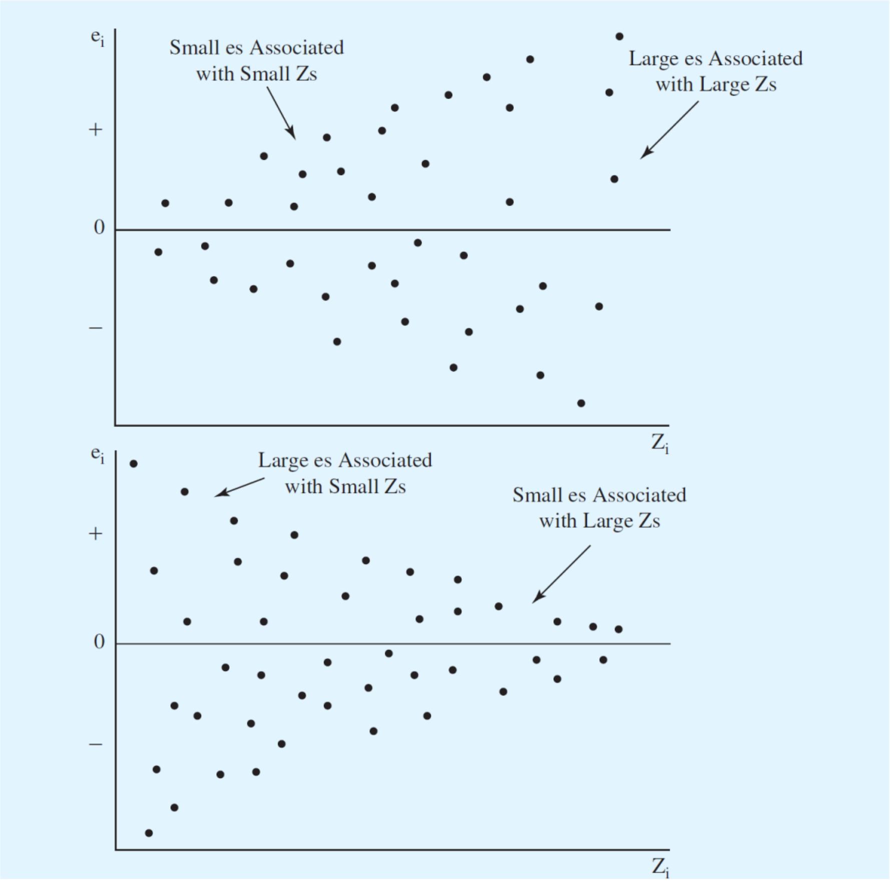
- Breusch-Pagan test
- White test

Before testing for heteroskedasticity, start with asking:

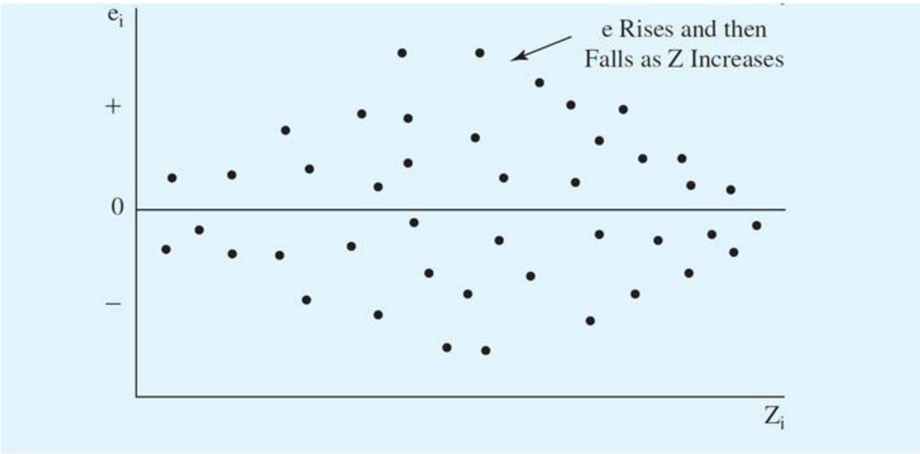
1. Are there any obvious specification errors?
2. Are there any early warning signs of heteroskedasticity?
3. Does a graph of the residuals show any evidence of heteroskedasticity?

Testing for Heteroskedasticity (cont'd)

Eyeballing Residuals for Possible Heteroskedasticity



If you plot the residuals of an equation with respect to a potential explanatory variable Z, a pattern in the residuals is an indication of possible heteroskedasticity.



The Breusch-Pagan Test for Heteroskedasticity:

Steps:

1. Estimate the model $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u}$ by OLS, as usual. Obtain the squared OLS residuals $\hat{u}$
2. Run the regression in $\boxed{\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + error}$ Keep the R-squared from this regression $R^2_{\hat{u}^2}$
3. Form either the *F* statistic or the *LM* statistic and compute the *p*-value. If the *p*-value is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity.

$\boxed{H_0 : Var(u|x_1, x_2, \ldots, x_k) = Var(u|\mathbf{x}) = \sigma^2} \longrightarrow H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$

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

$$F = \frac{R^2_{\hat{u}^2}/k}{1 - R^2_{\hat{u}^2}/(n-k-1)} \qquad LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2_k$$

A large **F statistic** or a large **Lagrange multiplier** statistic, (LM) lead to rejection of the null hypothesis.

The White Test for Heteroskedasticity

Steps:

1. Estimate the model $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u}$ by OLS, as usual. Obtain the squared OLS residuals $\hat{u}$
2. Run the regression in $\boxed{\begin{array}{c}\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error.\end{array}}$ Keep the R-squared from this regression $R^2_{\hat{u}^2}$
3. Form either the *F* statistic or the *LM* statistic and compute the *p*-value. If the *p*-value is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity.

$$\boxed{H_0 : Var(u|x_1, x_2, \ldots, x_k) = Var(u|\mathbf{x}) = \sigma^2} \longrightarrow H_0 : \delta_1 = \delta_2 = \cdots = \delta_9 = 0$$

Regress squared residuals on all explanatory variables, their squares, and interactions (here: example for k=3)

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

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

A large **F statistic** or a large **Lagrange multiplier** statistic, (LM) lead to rejection of the null hypothesis.

The White test for heteroskedasticity (cont'd)

Advantage:

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

Disadvantage of this form of the White test:

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

Alternative form of the White test

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

This regression indirectly tests the dependence of the squared residuals on the explanatory variables, their squares, and interactions, because the predicted value of y and its square implicitly contain all of these terms. (what is the number of restrictions in this alternative form?)

Example: Heteroskedasticity in housing price equations
BP test using F statistic

$$\widehat{price} = -21.77 + .00207\, lotsize + .123\, sqrft + 13.85\, bdrms$$
$$\quad (29.48) \quad (.00064) \qquad (.013) \qquad (9.01)$$
$$n = 88,\ R^2 = .672.$$

```
> # BP test (using F statistics)
> u_hat <- resid(MRM)
> summary(lm( u_hat^2     ~ lotsize+sqrft+bdrms, data=hprice1))

Call:
lm(formula = u_hat^2 ~ lotsize + sqrft + bdrms, data = hprice1)

Residuals:
   Min     1Q Median     3Q    Max 
 -9044  -2212  -1256    -97  42582 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) -5.523e+03  3.259e+03  -1.694  0.09390 . 
lotsize      2.015e-01  7.101e-02   2.838  0.00569 **
sqrft        1.691e+00  1.464e+00   1.155  0.25128   
bdrms        1.042e+03  9.964e+02   1.046  0.29877   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6617 on 84 degrees of freedom
Multiple R-squared:  0.1601,	Adjusted R-squared:  0.1301 
F-statistic: 5.339 on 3 and 84 DF,  p-value: 0.002048
```

Reject Homoskedasticity

$$\widehat{\log(price)} = -1.30 + .168 \log(lotsize) + .700 \log(sqrft) + 0.37\, bdrms$$
$$\quad (.65) \quad (.038) \qquad (.093) \qquad (.028)$$
$$n = 88,\ R^2 = .643.$$

```
> # BP test (using F statistics)
> u_hat_log <- resid(MRM_log)
> summary(lm( u_hat_log^2 ~ log(lotsize)+log(sqrft)+bdrms, data=hprice1))

Call:
lm(formula = u_hat_log^2 ~ log(lotsize) + log(sqrft) + bdrms, 
    data = hprice1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.05601 -0.03011 -0.01687  0.00523  0.40978 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.509994   0.257857   1.978   0.0512 .
log(lotsize) -0.007016   0.015156  -0.463   0.6446  
log(sqrft)   -0.062737   0.036767  -1.706   0.0916 .
bdrms         0.016841   0.010900   1.545   0.1261  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07309 on 84 degrees of freedom
Multiple R-squared:  0.04799,	Adjusted R-squared:  0.01399 
F-statistic: 1.411 on 3 and 84 DF,  p-value: 0.2451
```

Fail to reject Homoskedasticity

Example: Heteroskedasticity in housing price equations
White test using F statistic

$$\widehat{price} = -21.77 + .00207\, lotsize + .123\, sqrft + 13.85\, bdrms$$
$$\qquad (29.48)\ (.00064) \qquad (.013) \qquad (9.01)$$
$$n = 88,\ R^2 = .672.$$

```
> # white test (using F statistics)
> u_hat <- resid(MRM)
> y_hat <- predict(MRM)
> summary(lm( u_hat^2   ~ y_hat  + I(y_hat^2) ))

Call:
lm(formula = u_hat^2 ~ y_hat + I(y_hat^2))

Residuals:
   Min     1Q Median     3Q    Max 
-16805  -2185  -1521    324  40853 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 19071.5921  8876.2273   2.149  0.03451 * 
y_hat        -119.6554    53.3172  -2.244  0.02742 * 
I(y_hat^2)      0.2089     0.0746   2.801  0.00631 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6480 on 85 degrees of freedom
Multiple R-squared:  0.1849,	Adjusted R-squared:  0.1657 
F-statistic: 9.639 on 2 and 85 DF,  p-value: 0.0001687
```

Reject Homoskedasticity

$$\widehat{\log(price)} = -1.30 + .168 \log(lotsize) + .700 \log(sqrft) + 0.37\, bdrms$$
$$\qquad (.65)\ (.038) \qquad (.093) \qquad (.028)$$
$$n = 88,\ R^2 = .643.$$

```
> # white test (using F statistics)
> u_hat_log <- resid(MRM_log)
> logy_hat  <- predict(MRM_log)
> summary(lm( u_hat_log^2   ~ logy_hat  + I(logy_hat^2) ))

Call:
lm(formula = u_hat_log^2 ~ logy_hat + I(logy_hat^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.06004 -0.03177 -0.01364  0.00528  0.41808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.0468     3.3450   1.509    0.135
logy_hat       -1.7092     1.1633  -1.469    0.145
I(logy_hat^2)   0.1451     0.1010   1.437    0.154

Residual standard error: 0.07299 on 85 degrees of freedom
Multiple R-squared:  0.03917,	Adjusted R-squared:  0.01657 
F-statistic: 1.733 on 2 and 85 DF,  p-value: 0.183
```

Fail to reject Homoskedasticity

Example: Heteroskedasticity in housing price equations
BP and White tests using LM ratio

$$\widehat{price} = -21.77 + .00207\, lotsize + .123\, sqrft + 13.85\, bdrms$$
$$\quad (29.48)\quad (.00064) \qquad (.013) \qquad (9.01)$$
$$n = 88,\ R^2 = .672.$$

```
> # BP test (using LM statistics)
> bptest(MRM)

        studentized Breusch-Pagan test

data:  MRM
BP = 14.092, df = 3, p-value = 0.002782
```

```
> # White test (using LM statistics)
> y_hat <- predict(MRM)
> bptest(MRM,          ~ y_hat      + I(y_hat^2) )

        studentized Breusch-Pagan test

data:  MRM
BP = 16.268, df = 2, p-value = 0.0002933
```

Reject Homoskedasticity

$$\widehat{\log(price)} = -1.30 + .168 \log(lotsize) + .700 \log(sqrft) + 0.37\, bdrms$$
$$\quad (.65)\quad (.038) \qquad (.093) \qquad (.028)$$
$$n = 88,\ R^2 = .643.$$

```
> # BP test (using LM statistics)
> bptest(MRM_log)

        studentized Breusch-Pagan test

data:  MRM_log
BP = 4.2232, df = 3, p-value = 0.2383
```

```
> # White test (using LM statistics)
> logy_hat <- predict(MRM_log)
> bptest(MRM_log,      ~ logy_hat   + I(logy_hat^2) )

        studentized Breusch-Pagan test

data:  MRM_log
BP = 3.4473, df = 2, p-value = 0.1784
```

Fail to reject Homoskedasticity