Heteroskedasticity in Regression

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

2020

Outline

1 Introduction

2 Fix #1: Robust Standard Errors

3 Weighted Least Squares
- Combine Subsets of a Sample
- Random coefficient model
- Aggregate Data

4 Testing for heteroskedasticity
- Categorical Heteroskedasticity
- Checking for Continuous Heteroskedasticity
- Toward a General Test for Heteroskedasticity

5 Appendix: Robust Variance Estimator Derivation

6 Practice Problems

Remember the Basic OLS

- Theory behind the Linear Model

$$y_i = \beta_0 + \beta_1 x1_i + e_i$$

- Error term, we assumed, for all i,
  - $E(e_i) = 0$ *for all* i (errors are "symmetric" above and below)
  - $Var(e_i) = E[(e_i - E(e_i))^2] = \sigma^2$ (Homoskedasticity: same variance).
- Heteroskedasticity: the assumption of homogeneous variance is violated.

Homoskedasticity means

$$Var(e) = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}$$

Heteroskedasticity depicted in one of these ways

$$Var(e) = \begin{bmatrix} \sigma_e^2 w_1 & & & 0 \\ & \sigma_e^2 w_2 & & \\ & & \sigma_e^2 w_3 & \\ 0 & & & \sigma_e^2 w_N \end{bmatrix} \text{ or } \begin{bmatrix} \frac{\sigma_e^2}{w_1} & & & 0 \\ & \frac{\sigma_e^2}{w_2} & & \\ & & \frac{\sigma_e^2}{w_3} & \\ 0 & & & \frac{\sigma_e^2}{w_N} \end{bmatrix}$$

I get confused a lot when comparing textbooks because of this problem!

Consequences of Heteroskedasticity 1: $\hat{\beta}^{OLS}$ still unbiased, consistent

- OLS Estimates of β_0 and β_1 are still unbiased and consistent.
  - Unbiased: $E[\hat{\beta}^{OLS}] = \beta$
  - Consistent: As $N \to \infty$, $\hat{\beta}^{OLS}$ tends to β in probability limit.
- If the predictive line was "right" before, It's still right now.
- However, these are incorrect
  - standard error of $\hat{\beta}$
  - *RMSE*
  - confidence / prediction intervals

Proof: $\hat{\beta}^{OLS}$ Still Unbiased

- Begin with "mean centered" data. OLS with one variable:

$$\hat{\beta}_1 = \frac{\sum x_i \cdot y_i}{\sum x_i^2} = \frac{\sum x_i(b \cdot x_i + e_i)}{\sum x_i^2} = \frac{\beta_1 \sum x_i^2 + \sum x_i \cdot e_i}{\sum x_i^2} = \beta_1 + \frac{\sum x_i \cdot e_i}{\sum x_i^2}$$

- Apply the Expected value operator to both sides:

$$E[\hat{\beta}_1] = E(\beta_1) + E(\frac{\sum x_i \cdot e_i}{\sum x_i^2})$$

$$E[\hat{\beta}_1] = \beta_1 + E(\frac{\sum x_i \cdot e_i}{\sum x_i^2}) = \beta_1 + (\frac{\sum E[x_i \cdot e_i]}{\sum x_i^2})$$

- Assume x_i is uncorrelated with e_i, $E[x_i e_i] = 0$, the work is done

$$E(\hat{\beta}_1) = \beta_1$$

Consequence 2. OLS formula for $\widehat{Var(\hat{\beta})}$ is wrong

1. Usual formula to estimate $Var(\hat{\beta})$, $\widehat{Var(\hat{\beta})}$ is wrong. And it's square root, the $std.err.(\hat{\beta})$ is wrong.
2. Thus t-tests are WRONG (too big).

Proof: OLS $\widehat{Var(\hat{\beta})}$ Wrong

- Variance of e_i: $Var(e_i)$.
- The variance of the OLS slope estimator, $Var(\hat{b}_1)$, in "mean-centered (or deviations) form":

$$Var(\hat{\beta}_1) = Var\left[\frac{\sum x_i \cdot e_i}{\sum x_i^2}\right] = \frac{Var[\sum x_i e_i]}{(\sum x_i^2)^2} = \frac{\sum Var(x_i e_i)}{(\sum x_i^2)^2} = \frac{\sum x_i^2 \cdot Var(e_i)}{(\sum x_i^2)^2} \tag{1}$$

- We assume all $Var(e_i)$ are equal, and we put in the MSE as an estimate of it.

$$\widehat{Var(\hat{\beta}_1)} = \frac{MSE}{\sum x_i^2} \tag{2}$$

Proof: OLS $Var(\hat{\beta})$Wrong (page 2)

- Instead, suppose the "true" variance

$$Var(e_i) = s^2 + s_i^2 \tag{3}$$

(a common minimum variance s^2 plus an additional individualized variance s_i^2).

- Plug this into (1):

$$\frac{\sum x_i^2(s^2 + s_i^2)}{(\sum x_i^2)^2} = \frac{s^2}{\sum x_i^2} + \frac{\sum x_i \cdot s_i^2}{(\sum x_i^2)^2} \tag{4}$$

- The first term is "roughly" what OLS would calculate for the variance of $\hat{\beta}_1$.
- The second term is the additional "true variance" in $\hat{\beta}_1$ that the OLS formula $\widehat{V(\hat{\beta}_1)}$ does not include.

Consequence 3: $\hat{\beta}^{OLS}$ is Inefficient

1. $\hat{\beta}_i^{OLS}$ is *inefficient*: It has higher variance than the "weighted" estimator.
2. Note that to prove an estimator is "inefficient", it is necessary to provide an alternative estimator that has lower variance.
3. WLS: Weighted Least Squares estimator, $\hat{\beta}_1^{WLS}$.
4. The Sum of Squares to be minimized now includes a weight for each case

$$SS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{N} W_i (y - \hat{y}_i)^2 \tag{5}$$

5. The weights chosen to "undo" the heteroskedasticity.

$$W_i^2 = 1/Var(e_i) \tag{6}$$

Covariance matrix of error terms

- This thing is a weighting matrix

$$\begin{bmatrix} \frac{1}{Var(e_1)} & & & 0 \\ & \frac{1}{Var(e_2)} & & \\ & & \ddots & \\ 0 & & & \frac{1}{Var(e_N)} \end{bmatrix}$$

Is usually simplified in various ways.

- Factor out a common parameter, so each individual's error variance is proportional

$$\frac{1}{\sigma_e^2} \begin{bmatrix} \frac{1}{w_1} & & & 0 \\ & \frac{1}{w_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{w_N} \end{bmatrix}$$

Example. Suppose variance proportional to x_i^2

The "truth" is

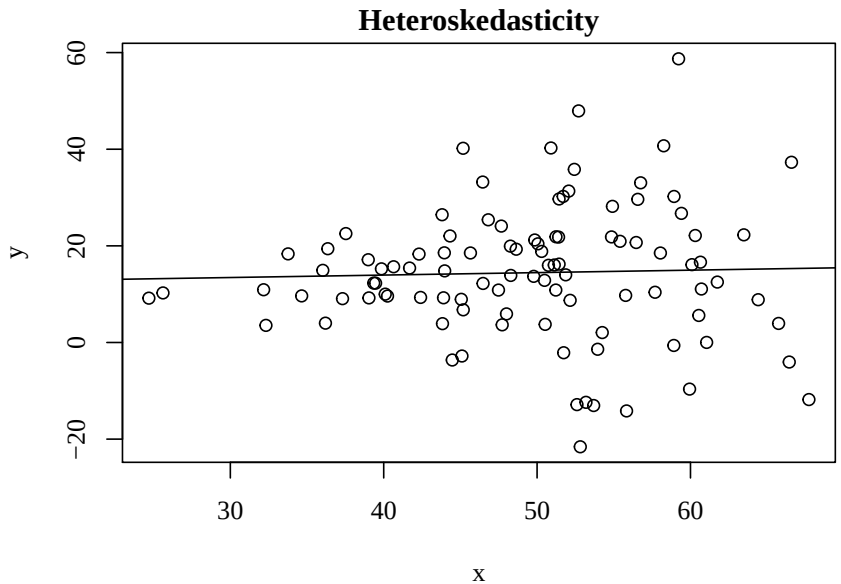
$$y_i = 3 + 0.25x_i + e_i \qquad (7)$$

Homoskedastic:

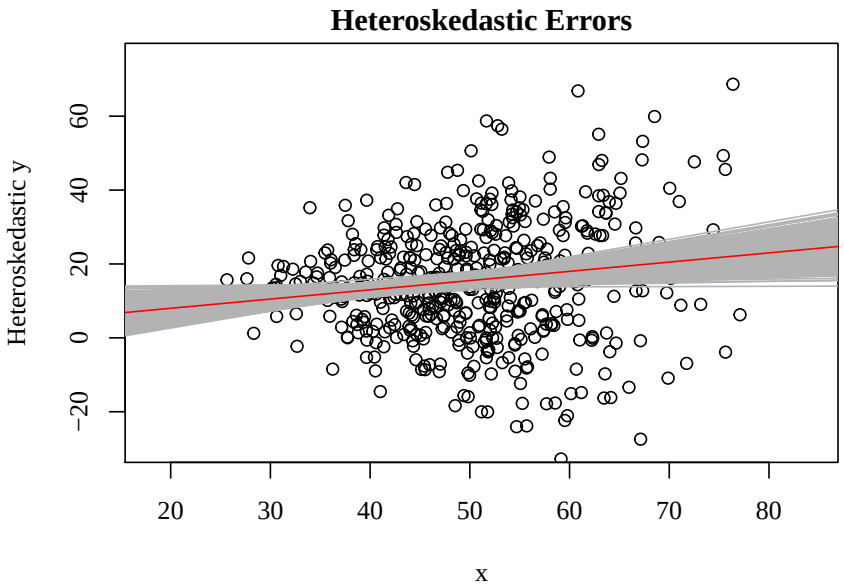
$$Std.Dev.(e_i) = \sigma_e = 10 \qquad (8)$$

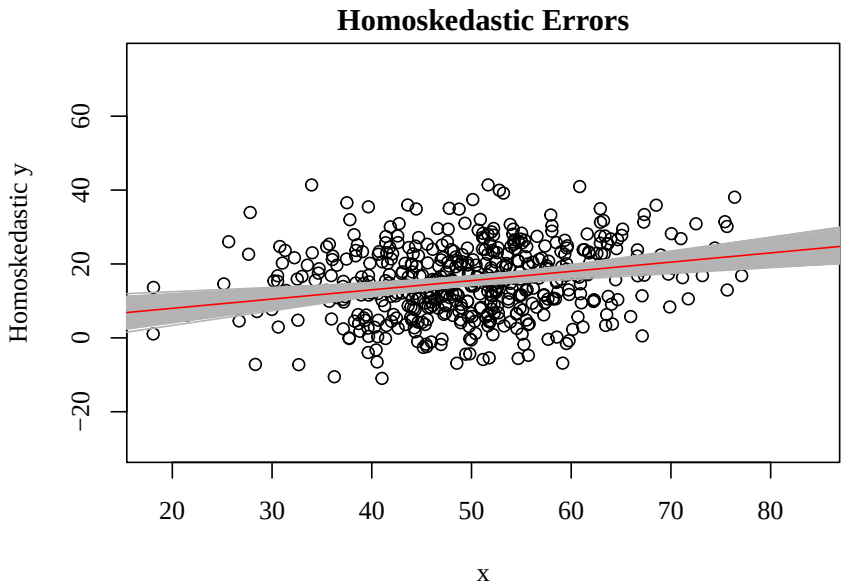
Heteroskedastic:

$$Std.Dev.(e_i) = 0.05 * (x_i - min(x_i)) * \sigma_e \qquad (9)$$

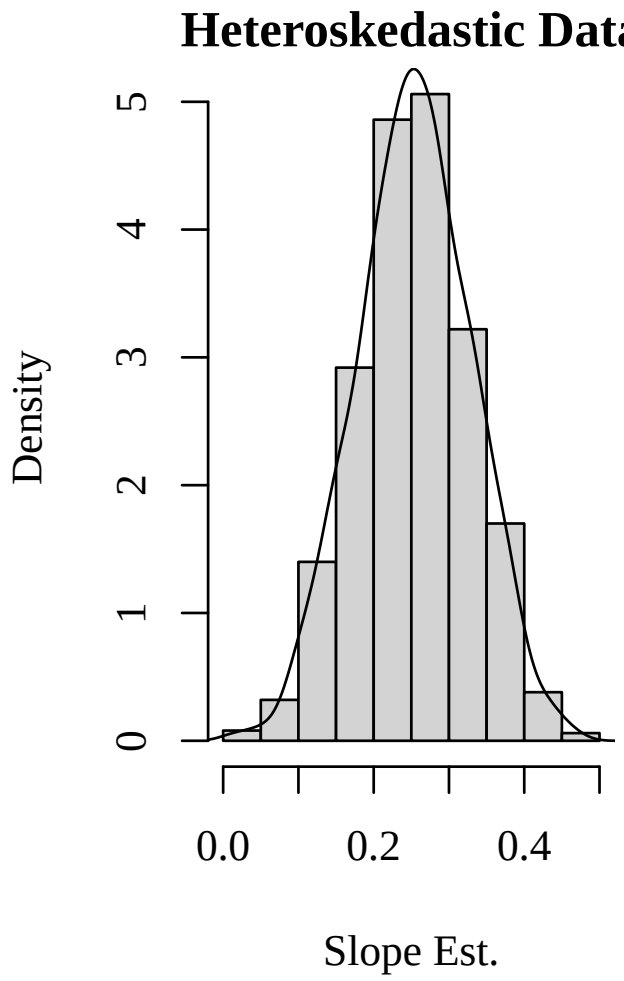


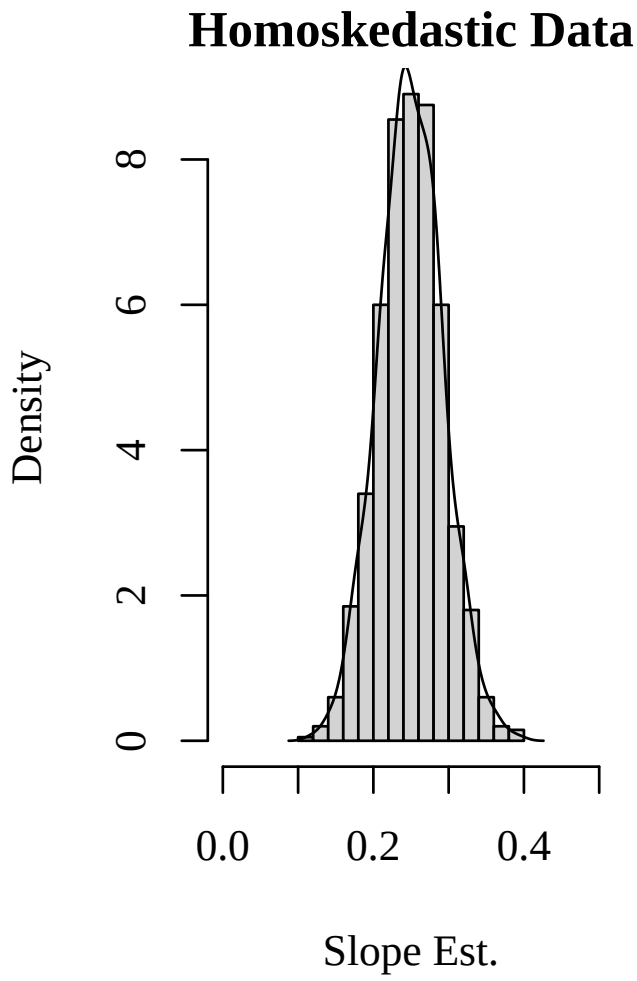
Compare Lines of 1000 Fits (Homo vs Heteroskedastic)





Histograms of Slope Estimates (w/Kernel Density Lines)





So, the Big Message Is

- Heteroskedasticity inflates the amount of uncertainty in the estimates.
- Distorts t-ratios.

Outline

1 Introduction

2 Fix #1: Robust Standard Errors

3 Weighted Least Squares
- Combine Subsets of a Sample
- Random coefficient model
- Aggregate Data

4 Testing for heteroskedasticity
- Categorical Heteroskedasticity
- Checking for Continuous Heteroskedasticity
- Toward a General Test for Heteroskedasticity

5 Appendix: Robust Variance Estimator Derivation

6 Practice Problems

Robust estimate of the variance of $\hat{b}$

- Replace OLS formula for $\widehat{Var(\hat{b})}$ with a more "robust" version
- Robust "heteroskedasticity consistent" variance estimator: weaker assumptions.
- No known small sample properties
  - But are consistent / asymptotically valid
- Note: This does not "fix" $\hat{b}^{OLS}$. It just gives us more accurate t-ratios by correcting $std.err(\hat{b})$.

Robust Std.Err. in a Nutshell

- Recall: the variance-covariance matrix of the errors assumed by OLS.

$$Var(e) = E(e \cdot e'|X) = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 \\ \dots & & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 \end{bmatrix} \tag{10}$$

Heteroskedastic Covariance Matrix

- If there's heteroskedasticity, we have to allow the possibility like this:

$$Var(e) = E[e \cdot e' | X] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \sigma_{N-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_N^2 \end{bmatrix}$$

Robust Std.Err. : Use Variance Estimates

- Fill in estimates for the case-specific error variances

$$\widehat{Var(e)} = \begin{bmatrix} \widehat{\sigma_1^2} & 0 & 0 & 0 & 0 \\ 0 & \widehat{\sigma_2^2} & 0 & 0 & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \widehat{\sigma_{N-1}^2} & 0 \\ 0 & 0 & 0 & 0 & \widehat{\sigma_N^2} \end{bmatrix}$$

- Embed those estimates into the larger formula that is used to calculate the robust standard errors.
- Famous paper
  White, Halbert. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4), 817-838.
  Robust estimator originally proposed by Huber (1967), but was forgotten

Derivation: Calculating The Robust Estimator of $Var(\hat{b})$

- The true variance of the OLS estimator is

$$Var(\hat{b}_1) = \frac{\sum x_i^2 Var(e_i)}{(\sum x_i^2)^2} \tag{11}$$

Assuming Homoskedasticity, estimate σ_e^2 with MSE.

$$\widehat{Var(\hat{b}_1)} = \frac{MSE}{\sum x_i^2} \text{ and the square root of that is } std.err.(\hat{b}_1) \tag{12}$$

- The robust versions replace $Var(e_i)$ with other estimates. White's suggestion was

$$Robust\ \widehat{Var(\hat{b}_1)} = \frac{\sum x_i^2 \cdot \hat{e}_i^2}{(\sum x_i^2)^2} \tag{13}$$

$\hat{e}_i{}^2$: the "squared residual", used in place of the unknown error variance.

Outline

1 Introduction

2 Fix #1: Robust Standard Errors

3 Weighted Least Squares
- Combine Subsets of a Sample
- Random coefficient model
- Aggregate Data

4 Testing for heteroskedasticity
- Categorical Heteroskedasticity
- Checking for Continuous Heteroskedasticity
- Toward a General Test for Heteroskedasticity

5 Appendix: Robust Variance Estimator Derivation

6 Practice Problems

WLS is Efficient

- Change OLS:
  - $$SS(\hat{\beta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$
- to WLS:

$$minimize\, SS(\hat{\beta}) = \sum_{i=1}^{N} W_i (y_i - \hat{y}_i)^2$$

- In practice, weights are guesses about the standard deviation of the error term

$$W_i = \frac{1}{\sigma_i} \tag{14}$$

Feasible Weighted Least Squares.

- Analysis proceeds in 2 steps.
  - Regression is estimated to gather information about the error variance.
  - That information is used to fill in the Weight matrix with WLS
- May revise the weights, re-fit the WLS, repeatedly until convergence.

Data From Categorical Groups

- Suppose you separately investigate data for men and women

$$men: \ y_i = \beta_0 + \beta_1 x_i + e_i \tag{15}$$

$$women: \ y_i = c_0 + c_1 x_i + u_i \tag{16}$$

- Then you wonder, "can I combine the data for men and women to estimate one model"

$$humans: \ y_i = \beta_0 + \beta_1 x_i + \beta_2 sex_i + \beta_3 sex_i x_i + e_i \tag{17}$$

- This "manages" the differences of intercept and slope for men and women by adding coefficients β_2 and β_3.
- But this ASSUMED that $Var(e_i) = Var(u_i)$.
- We should have tested for homoskedasticity (the ability to pool the 2 samples).

Methods Synonyms

The basic idea is to say that the linear model has "extra" random error terms.

Synonyms

- Random effects models
- Mixed Models
- Hierarchical Linear Models (HLM)
- Multi-level Models (MLM)

This "Laird and Ware" notation has now become a standard. Let the "fixed" coefficients be β's, but suppose in addition there are random coefficients $b \sim N(0, \sigma_b^2)$.

$$y = X\beta + Zb + e \tag{18}$$

- I'll probably write something on the board.

Simple Random Coefficient Model

- Start with the regression model that has a different slope for each case:

$$y_i = \beta_0 + \beta_i x_i + e_i \tag{19}$$

- Slope is a "random coefficient" with 2 parts

$$\beta_i = \beta_1 + u_i$$

  - β_1 is the "same" for all cases
  - u_i is noise in the slope that is individually assigned. It has expected value 0 and a variance σ_u^2.

Simple Random Coefficient Model

- The regression model becomes

$$y_i = \beta_0 + (\beta_1 + u_i)x_i + e_i$$

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

- Note: My "new" error term is $u_i x_i + e_i$. NOT homoskedastic
- What's the variance of that? Apply the usual rule:

$$Var[u_i x_i + e_i] = x_i^2 Var(u_i) + Var(e_i) + 2x_i Cov(u_i, e_i)$$

- Get rid of the last part by asserting that the 2 random effects are uncorrelated, so we have

$$= x_i^2 \sigma_u^2 + \sigma_e^2$$

With Aggregated Data, the Variance is Almost Never Homogeneous.

- Each row in the data set represents a collection of observations ("group averages" like "mean education" or "mean salary")
- The averaging process causes heteroskedasticity.
- The mean $\bar{y} = \frac{\sum y_i}{N}$ and standard deviation σ_y^2 imply the variance of the mean is

$$Var(\bar{y}) = \frac{Var(y_i)}{N} = \frac{\sigma_y^2}{N}$$

- Regression Weights proportional to $\sqrt{N_{group}}$ should be used.

Outline

1 Introduction

2 Fix #1: Robust Standard Errors

3 Weighted Least Squares
- Combine Subsets of a Sample
- Random coefficient model
- Aggregate Data

4 Testing for heteroskedasticity
- Categorical Heteroskedasticity
- Checking for Continuous Heteroskedasticity
- Toward a General Test for Heteroskedasticity

5 Appendix: Robust Variance Estimator Derivation

6 Practice Problems

Adapt tests from Analysis of Variance

Idea: Estimate the error variances for the subgroups, try to find out if they are different.

- Bartlett's test: Assuming normality of observations, derives a statistic that is distributed as a χ^2.

```
library(lmtest)
plot(count ~ spray, data = InsectSprays)
bartlett.test(count ~ spray, data = InsectSprays)
```

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London Series A* 160, 268–282.

- Fligner-Killeen Test : Robust against non-normality (less likely to confuse non-normality for heteroskedasticity)

```
library(lmtest)
fligner.test(count ~ spray, data = InsectSprays)
```

Adapt tests from Analysis of Variance ...

William J. Conover, Mark E. Johnson and Myrle M. Johnson (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23, 351–361.

- Levene's test

```
library(car)
leveneTest (y~x*z, data=dat) ##x and z must be factors
```

Goldfield Quandt test

S.M. Goldfeld & R.E. Quandt (1965), Some Tests for Homoskedasticity. *Journal of the American Statistical Association* 60, 539–547

- Consider a continuous numeric predictor. Exclude observations "in the middle" and then compare observed variances of the left and right.
- Draw a picture on Board here!
- HOWTO: compare the Error Sum of Squares for 2 chunks of data.

$$F = \frac{ESS_2}{ESS_1} = \frac{\text{the "lower set" ESS}}{\text{the "upper set" ESS}}$$

and the degrees of freedom for both numerator and denominator are $(N - d - 4)/2$, where d is the number of excluded observations.

- The more observations you exclude, the smaller will be your degrees of freedom, meaning your F value must be larger.

```
library(lmtest)
gqtest(y ~ x, fraction=0.2, order.by=c(z))
gqtest(y ~ x, point=0.4, order.by=c(z))
```

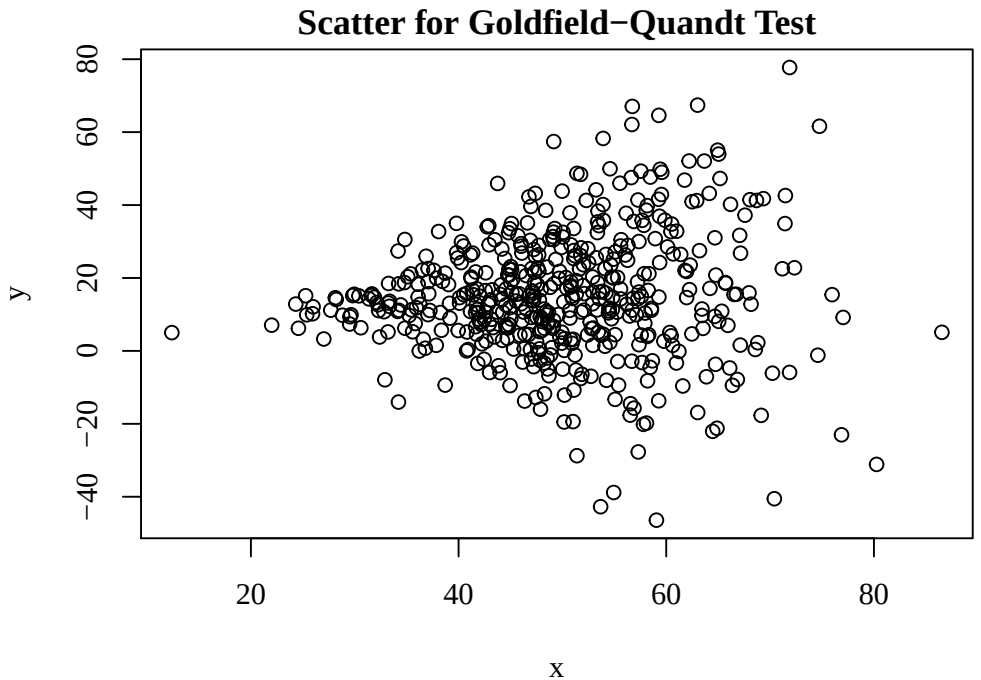
Example of Goldfield-Quandt Test: Continuous X

- Use heteroskedastic model from previous illustration.

```
mymod <- lm(y~x)
gqtest(mymod, fraction=0.2,
       order.by= ~ x)
```

```
  Goldfeld-Quandt test

data:  mymod
GQ = 4.497, df1 = 198, df2 =
    198, p-value < 2.2e-16
alternative hypothesis:
    variance increases from
    segment 1 to 2
```



This excludes 20% of the cases from the middle, and compares the variances of the outer sections.

Test for Predictable Squared Residuals

- Versions of this test were proposed in Breusch & Pagan (1979) and White (1980).
- Basic Idea: If errors are homogeneous, the variance of the residuals should not be predictable with the use of input variables.
- T.S. Breusch & A.R. Pagan (1979), A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* 47, 1287–1294

Breusch-Pagan test

- Model the squared residuals with the other predictors ($Z1_i$, etc) like this:

$$\frac{\widehat{e_i}^2}{\widehat{\sigma^2}} = \gamma_o + \gamma_1 Z1_1 + \gamma_2 Z2_i$$

Here, $\widehat{\sigma^2} = MSE$.

- If the error is homoskedastic/Normal, the coefficients γ_0, γ_1, and γ_2 will all equal zero. The input variables Z can be the same as the original regression, but may are include squared values of those variables.
- BP contend that $\frac{1}{2}RSS$ (the regression sum of squares) should be distributed as χ^2 with degrees of freedom equal to the number of Z variables.

```
library(lmtest)
mod <- lm (y ~ x1 + x2 +x3, data=dat)
bptest( mod , studentize=F) ##for the classic bp test
```

A Robust Version of the Test

- The original form of the BP test assumed Normally distributed errors. Non-normal, but homoskedastic, error, might cause the test to indicate there is heteroskedasticity.
- A "studentized" version of the test was proposed by Koenker (1981), that's what lmtest's bptest uses by default.

```
library(lmtest)
mod <- lm (y ~ x1 + x2 +x3, data=dat)
bptest( mod ) ## Koenker's robust version
```

White's Version of the Test

- White's general test for heteroskedasticity is another view of the same exercise. Run the regression

$$\widehat{e_i}^2 = \gamma_o + \gamma_1 Z1_1 + \gamma_2 Z2_i + \dots$$

The Z's should include the predictors, their squares, and cross products.

- Under the assumption of homoskedasticity, $N \cdot R^2$ is asymptotically distributed as χ^2_p, where N is the sample size, R^2 is the coefficient of determination from the fitted model, and p is the number of Z variables used in the regression.
- Algebraically equivalent to robust version of bp test (Waldman, 1983).

Outline

1 Introduction

2 Fix #1: Robust Standard Errors

3 Weighted Least Squares
- Combine Subsets of a Sample
- Random coefficient model
- Aggregate Data

4 Testing for heteroskedasticity
- Categorical Heteroskedasticity
- Checking for Continuous Heteroskedasticity
- Toward a General Test for Heteroskedasticity

5 Appendix: Robust Variance Estimator Derivation

6 Practice Problems

Where Robust $Var(\hat{\beta})$ Comes From

- The OLS estimator in matrix form

$$\hat{b} = (X'X)^{-1}X'Y \tag{20}$$

- If e_i is homoskedastic, the "true variance" of the estimates of the b's is

$$Var(\hat{b}) = \sigma^2 \cdot (X'X)^{-1} \tag{21}$$

Replace σ^2, with the Mean Squared Error (MSE).

$$\widehat{Var(\hat{b})} = MSE \cdot (X'X)^{-1} \tag{22}$$

Where OLS Exploits Homoskedastic Assumption

- In the OLS derivation of (22), one arrives at this intermediate step:

$$OLS: \ Var(\hat{b}) = (X'X)^{-1}(X'Var(e)X)(X'X)^{-1} \tag{23}$$

- The OLS derivation exploits homoskedasticity, which appears as

$$Var(e) = E(e \cdot e'|X) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ \dots & & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix} \tag{24}$$

$$= \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \dots & & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \sigma^2 \cdot I \tag{25}$$

$$OLS\ Var(\hat{b}) = (X'X)^{-1}(X' \cdot \sigma^2 \cdot X)(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

But Heteroskedasticity Implies

- If there's heteroskedasticity, we have to allow the possibility like this:

$$Var(e) = E[e \cdot e' | X] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ 0 & 0 & 0 & \sigma_{N-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_N^2 \end{bmatrix}$$

- Those "true variances" are unknown. How can we estimate $Var(\hat{b})$?

White's Idea

- The variance of e_1, for example, is never observed, but the best estimate we have for it is the mean square for that one case:

$$\widehat{e_1}^2 = (y_1 - X_1\hat{b})(y_1 - X_1\hat{b})$$

- Hence, Replace $Var(e)$ with a matrix of estimates like this:

$$\widehat{Var(e)} = \begin{bmatrix} \widehat{e_1}^2 & & & \\ & \widehat{e_2}^2 & & \\ & & \widehat{e_{N-1}}^2 & \\ & & & \widehat{e_N}^2 \end{bmatrix}$$

Heteroskedasticity Consistent Covariance Matrix

- The "heteroskedasticity consistent covariance matrix of $\hat{b}$" uses $\widehat{Var(e)}$ in the formula to calculate estimated variance.

$$hccm\ Var(\hat{b}) = (X'X)^{-1}(X'\widehat{Var(e)}X)(X'X)^{-1}$$

- White proved that the estimator is consistent, i.e, for large samples, the value converges to the true $Var(\hat{b})$.
- Sometimes called an "information sandwich" estimator. The matrix $(X'X)^{-1}$ is the "information matrix". This equation gives us a "sandwich" of $X'Var(e)X$ between two pieces of information matrix.

■echo=F■= dir.create("plots", showWarnings=F)

Outline

1 Introduction

2 Fix #1: Robust Standard Errors

3 Weighted Least Squares
- Combine Subsets of a Sample
- Random coefficient model
- Aggregate Data

4 Testing for heteroskedasticity
- Categorical Heteroskedasticity
- Checking for Continuous Heteroskedasticity
- Toward a General Test for Heteroskedasticity

5 Appendix: Robust Variance Estimator Derivation

6 Practice Problems

Problems

1. Somebody says "your regression results are obviously plagued by heteroskedasticity. And they are correct!" Explain what might be wrong, and what the consequences might be.
2. Run a regression on any data set you like. Suppose you call it "mod". Run a few garden variety heteroskedasticity checks.

```
library(lmtest)
bptest(mod)
```

If you have a continuous predictor "MYX" and you want to check for heteroskedasticity with the Goldfield-Quandt test, it is best to specify a fraction to exclude from the "middle" of the data. If you order the data frame by "MYX" before fitting the model and running the gqtest, it works a bit more smoothly. If you do not do that, you have to tell gqtest to order the data for you.

```
library(lmtest)
gqtest(mod, fraction=0.2, order.by=~MYX)
```

Problems ...

On the other hand, the help page for gqtest also suggests using it to test a dichotomous predictor, but in that case don't exclude a fraction in the middle, just specify the division point that splits the range of MYX in two. You better sort the dataset by MYX before trying this, it will be tricky.

```
library(lmtest)
dat <- dat[dat$MYX,    ] ##Sorts rows by MYX
gqtest(mod, point=67) ## splits data at row 67.
```

3. We have quite a few different ways to check for "categorical heteroskedasticity". I think I've never compared them side by side, but maybe you can. Run a regression that has at least one dichotomous predictor, and then run the various tests. I have in mind Bartlett's test, Fligner-Killeen test, and Levene's test. I noticed at the last minute we can also use the Goldfield Quandt test, in the method demonstrated in ?gqtest (or in previous question).

Problems ...

Run those tests, check to see if they all lead to the same conclusion or not. Try to understand what they are testing so you could explain them to one of your students.