

# Regression III

## Lecture 10: Non-Normality and Heteroskedasticity

Dave Armstrong

University of Wisconsin – Milwaukee  
Department of Political Science

e: armstrod@uwm.edu  
w: www.quantoid.net/ICPSR.php

1 / 41

### Goals of this Lecture

- Discuss methods for detecting non-normality, non-constant error variance, and nonlinearity
  - Each of these reflect problems with the specification of the model
- Discuss various ways that transformations can be used to remedy these problems
- Also explore polynomial regression as a potential solution to complicated nonlinear patterns
- Explore maximum likelihood methods that embed the linear model in a more general nonlinear model incorporating transformations as parameters

2 / 41

### Non-Normal Errors

#### Assessing Non-normality

### Non-constant Error Variance

Assessing Non-constant Error Variance  
Testing for Non-constant Error Variance  
Fixing Non-constant Error Variance: Without a Model  
Fixing Non-constant Error Variance: With a Model

3 / 41

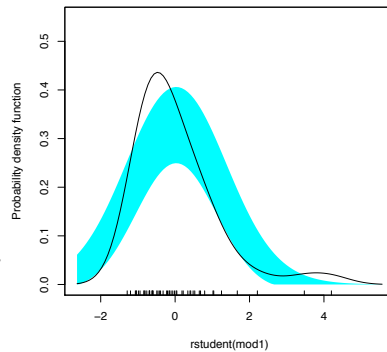
### Non-normally distributed errors

- The least-squares fit is based on the conditional mean
  - The mean is not a good measure of center for either a highly skewed distribution or a multi-modal distribution
- Non-Normality does not produce bias in the coefficient estimates, but it does have two important consequences:
  - It poses problems for efficiency - i.e., the OLS standard errors are no longer the smallest. Weighted least squares (WLS) is more efficient
  - Standard errors can be biased - i.e., confidence intervals and significance test may lead to wrong conclusions. Robust standard errors can compensate for this problem
- Transformations can often remedy the heavy-tailed problem
- Re-specification of the model - i.e., include a missing discrete predictor - can sometimes fix a multi-modal problem

4 / 41

## Distribution of the Residuals Example: Inequality data

- Quantile comparison plots and density estimates of the residuals from a model are useful for assessing normality
- The density estimate of the studentized residuals clearly shows a positive skew, and the possibility of a grouping of cases to the right

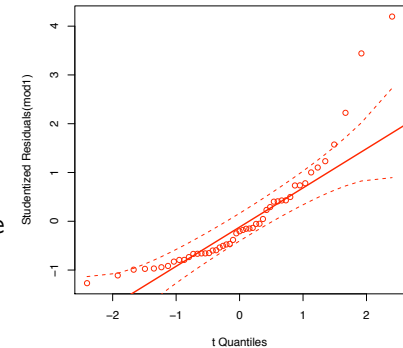


```
> Weakliem <- read.table("Weakliem.txt")
> mod1 <- lm(secpay ~ gini + democrat, data=Weakliem)
> sm.density(rstudent(mod1), model="normal")
```

5 / 41

## Assessing Unusual Cases

- A quantile comparison plot can give us a sense of which observations depart from normality.
- We can see that the points with the biggest departure are the Czech Republic and Slovakia.

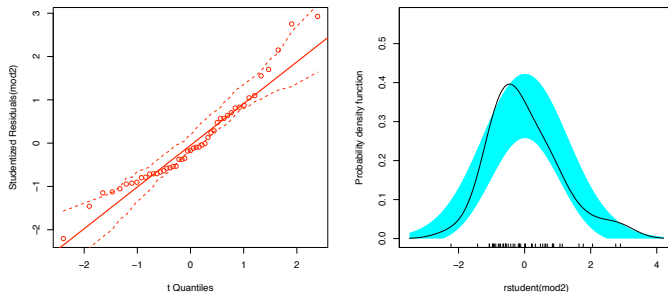


```
> library(car)
> qq.plot(mod1)

integer(0)
```

6 / 41

## Studentized Residuals after Removing the Czech Republic and Slovakia



```
> mod2 <- lm(secpay ~ gini + democrat, data=Weakliem,
+ subset=-c(25,49))
> qq.plot(mod2, simulate=T, labels=FALSE)
> sm.density(rstudent(mod2), model="normal")
```

7 / 41

## Non-Normal Errors Assessing Non-normality

### Non-constant Error Variance

Assessing Non-constant Error Variance  
Testing for Non-constant Error Variance  
Fixing Non-constant Error Variance: Without a Model  
Fixing Non-constant Error Variance: With a Model

8 / 41

## Non-constant Error Variance

- Also called *Heteroskedasticity*
- An important assumption of the least-squares regression model is that the variance of the errors around the regression surface is everywhere the same:  $V(E) = V(Y|x_1, \dots, x_k) = \sigma^2$ .
- Non-constant error variance does not cause biased estimates, but it does pose problems for efficiency and the usual formulas for standard errors are inaccurate
  - OLS estimates are inefficient because they give equal weight to all observations regardless of the fact that those with large residuals contain less information about the regression
- Two types of nonconstant error variance are relatively common:
  - Error variance increases as the expectation of  $Y$  increases;
  - There is a systematic relationship between the errors and one of the  $X$ 's

9 / 41

## Non-Normal Errors Assessing Non-normality

- Non-constant Error Variance
- Assessing Non-constant Error Variance
- Testing for Non-constant Error Variance
- Fixing Non-constant Error Variance: Without a Model
- Fixing Non-constant Error Variance: With a Model

10 / 41

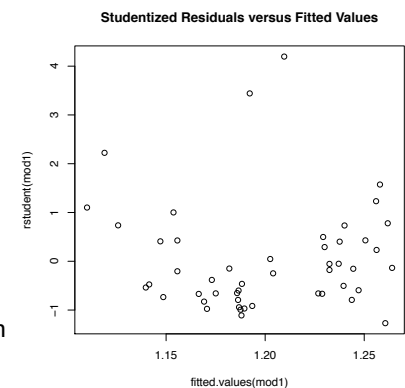
## Assessing Non-constant Error Variance

- Direct examination of the data is usually not helpful in assessing non-constant error variance, especially if there are many predictors. Instead, we look to the residuals to uncover the distribution of the errors.
  - It is also not helpful to plot  $Y$  against the residuals  $E$ , because there is a built-in correlation between  $Y$  and  $E$ :
$$Y = \hat{Y} + E$$
- The least squares fit ensures that the correlation between  $\hat{Y}$  and  $E$  is 0, so a plot of these (residual plot) can help us uncover nonconstant error variance.
  - The pattern of changing spread is often more easily seen using studentized residuals  $E_i^*$  against  $\hat{Y}$
- If the values of  $Y$  are all positive, we can use a Spread-level plot
  - plot  $\log(|E_i^*|)$  (called the log spread) against  $\log \hat{Y}$  (called the log level)
  - The slope  $b$  of the regression line fit to this plot suggests the variance-stabilizing transformation  $Y^{(p)}$ , with  $p = 1 - b$

11 / 41

## Assessing Heteroskedasticity: Example - Inequality Data (1)

- Two things are obvious:
  1. The residuals have a recognizable pattern, suggesting the model is missing something systematic
  2. There are two outlying cases (Czech Republic and Slovakia)
- We next take out the outliers and fit the model including an interaction between democracy and gini.



```
> plot(fitted.values(mod1), rstudent(mod1),  
+      main="Studentized Residuals versus Fitted Values")
```

12 / 41

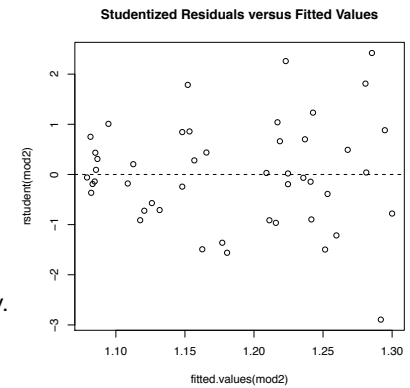
## Assessing Heteroskedasticity: Example - Inequality Data (3)

```
> infl.outliers <- which(rownames(Weakliem) %in%
+   c("Slovakia", "CzechRepublic"))
> mod2 <- lm(secpay~gini*democrat, data=Weakliem,
+   subset=-infl.outliers)
> plot(fitted.values(mod2), rstudent(mod2),
+   main="Studentized Residuals versus Fitted Values")
> abline(h=0, lty=2)
```

13 / 41

## Assessing Heteroskedasticity: Example - Inequality Data (4)

- The non-constant error variance is not as obviously problematic as it was before, but as we will see below, recent research offers assistance on testing for and remediating heteroskedasticity.



14 / 41

## Assessing Heteroskedasticity: Example - DHS Data (1)

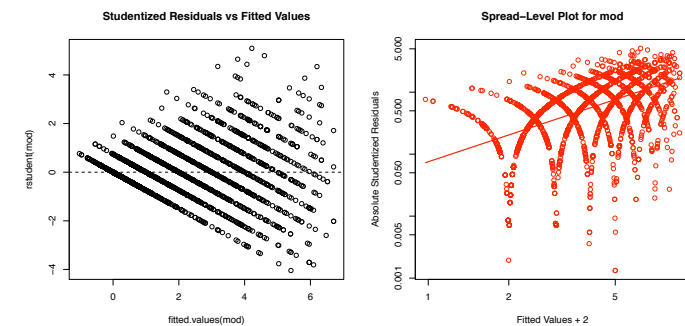
```
> library(apsrtable)
> dat <- read.dta("dhs_sl.dta")
> mod <- lm(ceb ~ age + christian + educ + notv, data=dat)
> apsrtable(mod, model.names="", Sweave=T, digits=3)
```

(Intercept)	-2.603*
	(0.097)
age	0.185*
	(0.003)
christian	-0.239*
	(0.054)
educjunior secondary	-0.321*
	(0.061)
educsenior secondary	-0.896*
	(0.075)
eductech school	-1.518*
	(0.120)
educhigher	-1.458*
	(0.209)
notv	0.226*
	(0.056)
N	2766
R <sup>2</sup>	0.610
adj. R <sup>2</sup>	0.609
Resid. sd	1.343

Standard errors in parentheses  
\* indicates significance at  $p < 0.05$

15 / 41

## Assessing Heteroskedasticity: Example - DHS Data (2)



- In the residual plot (left), we see the familiar “fanning” out of the data - i.e., the variance of the residuals is increasing as the fitted values get larger
- The slope of the spread-level plot (right) helps us find an appropriate transformation

16 / 41

## Assessing Heteroskedasticity: DHS Data (3)

```
> plot(fitted.values(mod), rstudent(mod),  
+ main="Studentized Residuals vs Fitted Values")  
> abline(h=0, lty=2)
```

```
> spread.level.plot(mod)
```

Suggested power transformation: -0.4358929

- The `spread.level.plot` command in `car` returns a suggested transformation for  $Y$  of  $\lambda = -0.436$  (1-slope of the spread level plot)

17 / 41

## Non-Normal Errors Assessing Non-normality

### Non-constant Error Variance

Assessing Non-constant Error Variance

Testing for Non-constant Error Variance

Fixing Non-constant Error Variance: Without a Model

Fixing Non-constant Error Variance: With a Model

18 / 41

## Testing for Non-Constant Error variance (1)

- Assume that a discrete  $X$  (or combination of  $X$ 's) partitions the data into  $m$  groups.
- Let  $Y_{ij}$  denote that  $i^{\text{th}}$  of  $n_j$  outcome-variable scores in group  $j$
- Within-group sample variances are then calculated as follows:

$$S_j^2 = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n_j - 1}$$

- We could then compare these within-group sample variances to see if they differ
- If the distribution of the errors is non-normal, however, tests that examine  $S_j^2$  directly are not valid because the mean is not a good summary of the data

19 / 41

## Testing for Non-Constant Error variance (2): Score Test

- A score test for the null hypothesis that all of the error variances  $\sigma^2$  are the same provides a better alternative

1. We start by calculating the standardized squared residuals

$$U_i = \frac{E_i^2}{\hat{\sigma}^2} = \frac{E_i^2}{\frac{\sum E_i^2}{n}}$$

2. Regress the  $U_i$  on all of the explanatory variable  $X$ 's, finding the fitted values:

$$U_i = \eta_0 + \eta_1 X_{i1} + \dots + \eta_p X_{ip} + \omega_i$$

3. The score test, which is distributed as  $\chi^2$  with  $p$  degrees of freedom is:

$$S_0^2 = \frac{\sum (\hat{U}_i - \bar{U})^2}{2}$$

20 / 41

## R-script testing for non-constant error variance

- The `ncv.test` function in the `car` library provides a simple way to carry out the score test
- The result below shows that the nonconstant error variance is statistically significant

```
> ncv.test(mod, ~age + christian + educ + notv, data=dat)
```

Non-constant Variance Score Test

Variance formula: ~ age + christian + educ + notv

Chisquare = 1338.552 Df = 7 p = 0

21 / 41

## Non-Normal Errors Assessing Non-normality

### Non-constant Error Variance

Assessing Non-constant Error Variance

Testing for Non-constant Error Variance

Fixing Non-constant Error Variance: Without a Model

Fixing Non-constant Error Variance: With a Model

22 / 41

## Robust Standard Errors (1)

- Robust standard errors can be calculated to compensate for an *unknown* pattern of non-constant error variance
- Robust standard errors require fewer assumptions about the model than WLS (which is better if there is increasing error variance in the level of  $Y$ )
  - Robust standard errors do not change the OLS coefficient estimates or solve the inefficiency problem, but do give more accurate  $p$ -values.
- There are several methods for calculating heteroskedasticity consistent standard errors (e.g., known variously as White, Eicker or Huber standard errors) but most are variants on the method originally proposed by White (1980).

23 / 41

## Robust Standard Errors (2): White's Standard Errors

- The covariance matrix of the OLS estimator is:

$$\begin{aligned} V(\mathbf{b}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Where  $V(\mathbf{y}) = \sigma_\varepsilon^2 \mathbf{I}_n$  if the assumptions of normality and homoskedasticity are satisfied. The variance simplifies to:

$$V(\mathbf{b}) = \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- In the presence of nonconstant error variance, however,  $V(\mathbf{y})$  contains nonzero covariance and unequal variances
  - In these cases, White suggests a consistent estimator of the variance that constrains  $\boldsymbol{\Sigma}$  to a diagonal matrix containing only squared residuals

24 / 41

## Robust Standard Errors (3): White's Standard Errors

- The *heteroskedasticity consistent covariance matrix* (HCCM) estimator is then:

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Phi}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where  $\hat{\mathbf{\Phi}} = e_i^2 \mathbf{I}_n$  and the  $e_i$  are the OLS residuals

- This is what is known as HC0 - White's (1980) original recipe.

25 / 41

## Other HCCM's

MacKinnon and White (1985) considered three alternatives: HC1, HC2 and HC3, each of which offers a different method for finding  $\hat{\mathbf{\Phi}}$ .

- HC1:  $\frac{N}{N-K} \times \text{HC0}$ .
- HC2:  $\hat{\mathbf{\Phi}} = \text{diag} \left[ \frac{e_i^2}{1-h_{ii}} \right]$  where  $h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$
- HC3:  $\hat{\mathbf{\Phi}} = \text{diag} \left[ \frac{e_i^2}{(1-h_{ii})^2} \right]$

Long and Ervin (2000) find that in small samples (e.g., < 500) the HC3 errors are the "best" in terms of size and power.

- They suggest using HC3 all of the time, as they do not do poorly in the presence of homoskedasticity and outperform all other options in the presence of heteroskedasticity.

26 / 41

## Small Sample Properties of Screening Tests

- Long and Ervin (2000) performed a monte carlo study on screening tests for heteroskedasticity.
- They find that in small samples (i.e., < 250), the standard tests (as discussed above) have very little power.
- With small samples, if there is *any* reason to suspect heteroskedasticity may be a problem, use HC3 robust SEs

27 / 41

## Function for lm output that includes White Standard Errors

- The `robust.se` function below relies on the `hccm` function in `car` and the `summary` function in the base **R** package
- ```
> robust.se <- function(model, type="hc3") {
+   require(car)
+   s <- summary(model)
+   wse <- sqrt(diag(hccm(model, type=type)))
+   t <- model$coefficients/wse
+   p <- 2*pt(abs(t), model$df.residual, lower.tail=F)
+   results <- round(cbind(model$coefficients, wse, t, p), 3)
+   dimnames(results) <- dimnames(s$coefficients)
+   results
+ }
```

28 / 41

## Summary of DHS Model

|                      | Estimate   | Std. Error  | t value    | Pr(> t )      |
|----------------------|------------|-------------|------------|---------------|
| (Intercept)          | -2.6027223 | 0.097424711 | -26.715217 | 4.964023e-140 |
| age                  | 0.1853348  | 0.002903105 | 63.840194  | 0.000000e+00  |
| christian            | -0.2388662 | 0.053968393 | -4.426038  | 9.970676e-06  |
| educjunior secondary | -0.3205895 | 0.061121063 | -5.245156  | 1.679835e-07  |
| educsenior secondary | -0.8964320 | 0.074886759 | -11.970500 | 3.144259e-32  |
| educotech school     | -1.5181394 | 0.120193419 | -12.630803 | 1.354791e-35  |
| educhigher           | -1.4575078 | 0.208853997 | -6.978597  | 3.719421e-12  |
| notv                 | 0.2262339  | 0.056013017 | 4.038952   | 5.515526e-05  |

29 / 41

## Output for robust .se

```
> robust.se(mod)
```

|                      | Estimate | Std. Error | t value | Pr(> t ) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -2.603   | 0.100      | -26.142 | 0        |
| age                  | 0.185    | 0.004      | 47.678  | 0        |
| christian            | -0.239   | 0.053      | -4.521  | 0        |
| educjunior secondary | -0.321   | 0.062      | -5.164  | 0        |
| educsenior secondary | -0.896   | 0.073      | -12.224 | 0        |
| educotech school     | -1.518   | 0.129      | -11.794 | 0        |
| educhigher           | -1.458   | 0.177      | -8.228  | 0        |
| notv                 | 0.226    | 0.053      | 4.288   | 0        |

30 / 41

## Robust Standard Errors (4)

- Since the HCCM is found without a formal model of the heteroskedasticity, relying instead on only the regressors and residuals from the OLS for its computation, it can be easily adapted to many applications
- For example. robust standard errors can be used to improve statistical inference from clustered data, pooled time-series data with autocorrelated errors (e.g., Newey-West SE's) and panel data
- **Cautions**
  - Robust standard errors should not be seen as a substitute for careful model specification. In particular, if the pattern of heteroskedasticity is known, it can often be more effectively corrected - and the model more efficiently estimated - using WLS

31 / 41

## Non-Normal Errors

### Assessing Non-normality

## Non-constant Error Variance

### Assessing Non-constant Error Variance

### Testing for Non-constant Error Variance

### Fixing Non-constant Error Variance: Without a Model

### Fixing Non-constant Error Variance: With a Model

32 / 41

## Weighted least squares (1)

- If the error variances are proportional to a particular  $X$  (i.e., the error variances are known up to a constant of proportionality  $\sigma_\varepsilon^2$ , so that  $V(\varepsilon_1) = \sigma_\varepsilon^2 w_i$ ), weighted least squares provides a good alternative to OLS
- WLS minimizes the weighted sum of squares  $\sum w_i E_i^2$  giving greater weight to observations with smaller variance
- The WLS maximum likelihood estimators are defined as:

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum \frac{E_i^2}{w_i}}{n}$$

- The estimated asymptotic covariance for the estimators is:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}_\varepsilon^2 (X'WX)^{-1}$$

- Here  $W$  is a square diagonal matrix with individual weights  $w_i$  on the diagonal and zeros elsewhere

33 / 41

## Weighted Least Squares Example: DHS Data

- The “fanning” pattern in the residual plot for the DHS model indicates that the error variance is proportional to age of respondent
- We could then proceed to estimate a WLS using the weight  $\frac{1}{age_i}$
- This amounts to estimating the following equation:

$$\frac{CEB}{\sqrt{Age}} = \frac{\beta_0}{\sqrt{Age}} + \beta_1 \sqrt{Age} + \beta_2 \frac{Christian}{\sqrt{Age}} + \beta_3 \frac{ED:JS}{\sqrt{Age}} + \beta_4 \frac{ED:SS}{\sqrt{Age}} + \beta_5 \frac{ED:T}{\sqrt{Age}} + \beta_6 \frac{ED:SS}{\sqrt{Age}} + \beta_7 \frac{NOTV}{\sqrt{Age}} + \frac{u}{\sqrt{Age}}$$

34 / 41

## WLS Example: DHS Data (2)

In **R**, we simply add a weight argument to the `lm` function:

```
> mod.wls <- lm(ceb ~ age + christian + educ + notv, data=dat,
+             weight=1/age)
> with(summary(mod.wls), coefficients)
```

|                      | Estimate   | Std. Error  | t value    | Pr(> t )      |
|----------------------|------------|-------------|------------|---------------|
| (Intercept)          | -2.7073426 | 0.082672817 | -32.747676 | 5.519108e-199 |
| age                  | 0.1887246  | 0.002779522 | 67.898216  | 0.000000e+00  |
| christian            | -0.1581601 | 0.045586879 | -3.469423  | 5.295772e-04  |
| educjunior secondary | -0.3134806 | 0.050975974 | -6.149575  | 8.893743e-10  |
| educsenior secondary | -0.8293801 | 0.064403195 | -12.877934 | 6.782367e-37  |
| eductech school      | -1.4611152 | 0.112660305 | -12.969210 | 2.215528e-37  |
| educhigher           | -1.4134369 | 0.191916237 | -7.364863  | 2.329690e-13  |
| notv                 | 0.1683805  | 0.047340311 | 3.556811   | 3.817048e-04  |

35 / 41

## Interpreting WLS Results

- It is important to remember that the parameters in the WLS model are estimators of  $\beta$ , just like the OLS parameters are (these are just more efficient in the presence of heteroskedasticity).
- Thus interpretation takes the same form as it does with the OLS parameters.
- The  $R^2$  is less interesting here because we are explaining variance in  $\frac{CEB}{\sqrt{Age}}$ , rather than CEB.

36 / 41

## Generalized Least Squares (1)

- Sometimes, we do not know the relationship between  $x_i$  and  $var(u_i|x_i)$ .
- In this case, we can use a Feasible GLS model.
- FGLS estimates the weight from the data. That weight is then used in a WLS fashion.

37 / 41

## GLS: Steps

1. Regress  $y$  on  $x_i$  and obtain residuals  $\hat{u}_i$ .
2. Create  $\log(\hat{u}_i^2)$  by squaring and then taking the natural log of the OLS residuals from step 1.
3. Run a regression of  $\log(\hat{u}_i^2)$  on  $x_i$  and obtain the fitted values  $\hat{g}_i$ .
4. Generate  $\hat{h}_i = \exp(\hat{g}_i)$ .
5. Estimate the WLS of  $y$  on  $x_i$  with weights of  $\frac{1}{\hat{h}_i}$ .

38 / 41

## FGLS Example: DHS Data

```
> mod1.ols <- lm(ceb ~ age + christian + educ + notv, data=dat)
> aux.mod1 <- lm(log(resid(mod1.ols)^2) ~ age + christian + educ + notv, data=dat)
> h <- exp(predict(aux.mod1))
> mod.fgls <- lm(ceb ~ age + christian + educ + notv, data=dat, weights=1/h)
> with(summary(mod.fgls), coefficients)
```

|                      | Estimate    | Std. Error  | t value    | Pr(> t )      |
|----------------------|-------------|-------------|------------|---------------|
| (Intercept)          | -2.64405992 | 0.068554862 | -38.568525 | 1.175615e-260 |
| age                  | 0.18262406  | 0.002948531 | 61.937298  | 0.000000e+00  |
| christian            | -0.04483191 | 0.029444272 | -1.522602  | 1.279730e-01  |
| educjunior secondary | -0.28653608 | 0.037977879 | -7.544815  | 6.113167e-14  |
| educsenior secondary | -0.69179453 | 0.044419466 | -15.574130 | 1.757188e-52  |
| eductech school      | -1.21986641 | 0.085580676 | -14.253994 | 1.554346e-44  |
| educhigher           | -1.22527349 | 0.125311667 | -9.777809  | 3.205361e-22  |
| notv                 | 0.10271727  | 0.030208606 | 3.400265   | 6.827801e-04  |

39 / 41

Table: Comparing Models

|                            | OLS                | WLS                | FGLS               |
|----------------------------|--------------------|--------------------|--------------------|
| (Intercept)                | -2.603*<br>(0.097) | -2.707*<br>(0.083) | -2.644*<br>(0.069) |
| age                        | 0.185*<br>(0.003)  | 0.189*<br>(0.003)  | 0.183*<br>(0.003)  |
| christian                  | -0.239*<br>(0.054) | -0.158*<br>(0.046) | -0.045<br>(0.029)  |
| educjunior secondary       | -0.321*<br>(0.061) | -0.313*<br>(0.051) | -0.287*<br>(0.038) |
| educsenior secondary       | -0.896*<br>(0.075) | -0.829*<br>(0.064) | -0.692*<br>(0.044) |
| eductech school            | -1.518*<br>(0.120) | -1.461*<br>(0.113) | -1.220*<br>(0.086) |
| educhigher                 | -1.458*<br>(0.209) | -1.413*<br>(0.192) | -1.225*<br>(0.125) |
| notv                       | 0.226*<br>(0.056)  | 0.168*<br>(0.047)  | 0.103*<br>(0.030)  |
| <i>N</i>                   | 2766               | 2766               | 2766               |
| <i>R</i> <sup>2</sup>      | 0.610              | 0.637              | 0.593              |
| adj. <i>R</i> <sup>2</sup> | 0.609              | 0.636              | 0.592              |
| Resid. sd                  | 1.343              | 0.236              | 1.845              |

Standard errors in parentheses  
\* indicates significance at  $p < 0.05$

40 / 41

## Conclusions

- Heteroskedasticity, while not bias-inducing, can cause problems with efficiency of the model.
- Tests of heteroskedasticity only have sufficient power when  $n$  is large (e.g.,  $> 250$ ).
- If the errors are found to be heteroskedastic, there are a number of potential fixes that all require various assumptions:
  1. If heteroskedasticity is thought to exist and no suitable functional form of the variance can be found, then HC3 robust standard errors are the best bet (especially in small samples).
  2. If the heteroskedasticity is proportional to a single variable, weighted least squares can be used.
  3. FGLS is required to model the functional form of the heteroskedasticity as a function of the  $x$  variables.