Chapter 8
Heteroskedasticity

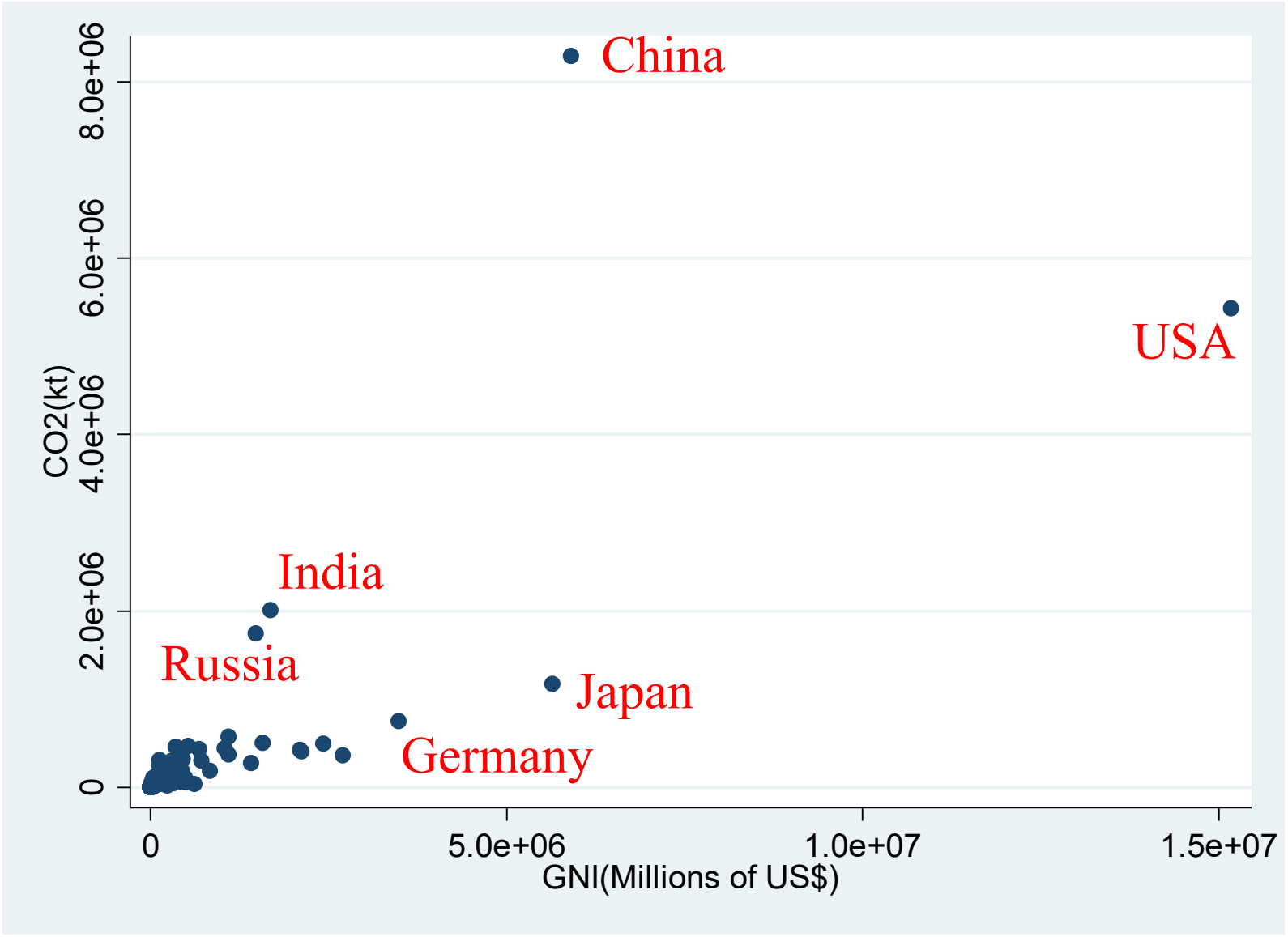
Learning Objectives

- Demonstrate the problem of heteroskedasticity and its implications
- Conduct and interpret tests for heteroscedasticity
- Correct for heteroscedasticity using White's heteroskedasticity-robust estimator
- Correct for heteroscedasticity by getting the model right

What is Heteroscedasticity?

- Hetero = different
- Scedastic – from a Greek word meaning dispersion

CO_2 Emissions vs National Income



The Problem

- The population model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- The errors come from a distribution with constant standard deviation

- What if the standard deviation is not constant? => Heteroskedasticity
  - Maybe it's higher for higher levels of X_{1i}, X_{2i}, or some combination of the two.

The Good News

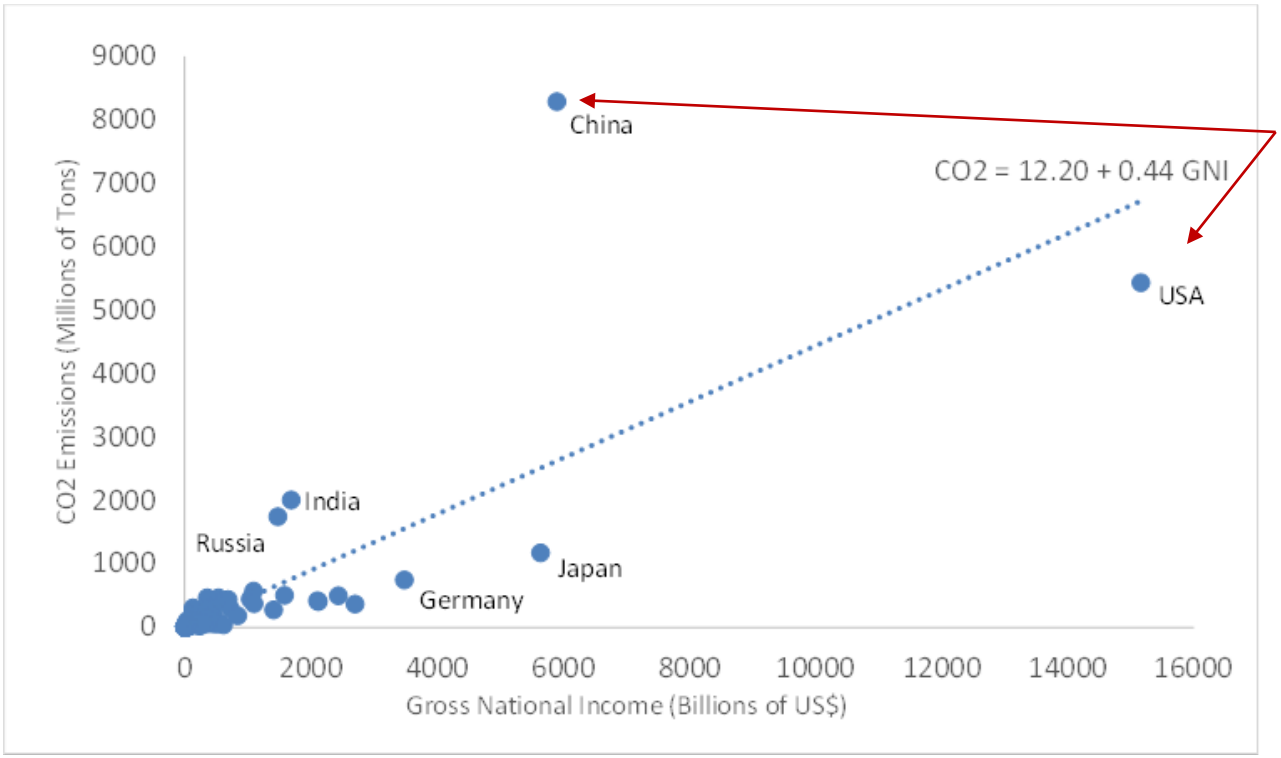
- As long as you still have a representative sample
    - …the error on average is zero and your estimates will still be unbiased

The Bad News

- …But the errors will not have a constant variance
  - That could make a big difference when it comes to testing hypotheses and setting up confidence intervals around your estimates.

- Heteroskedasticity means that some observations give you more information about β than others.

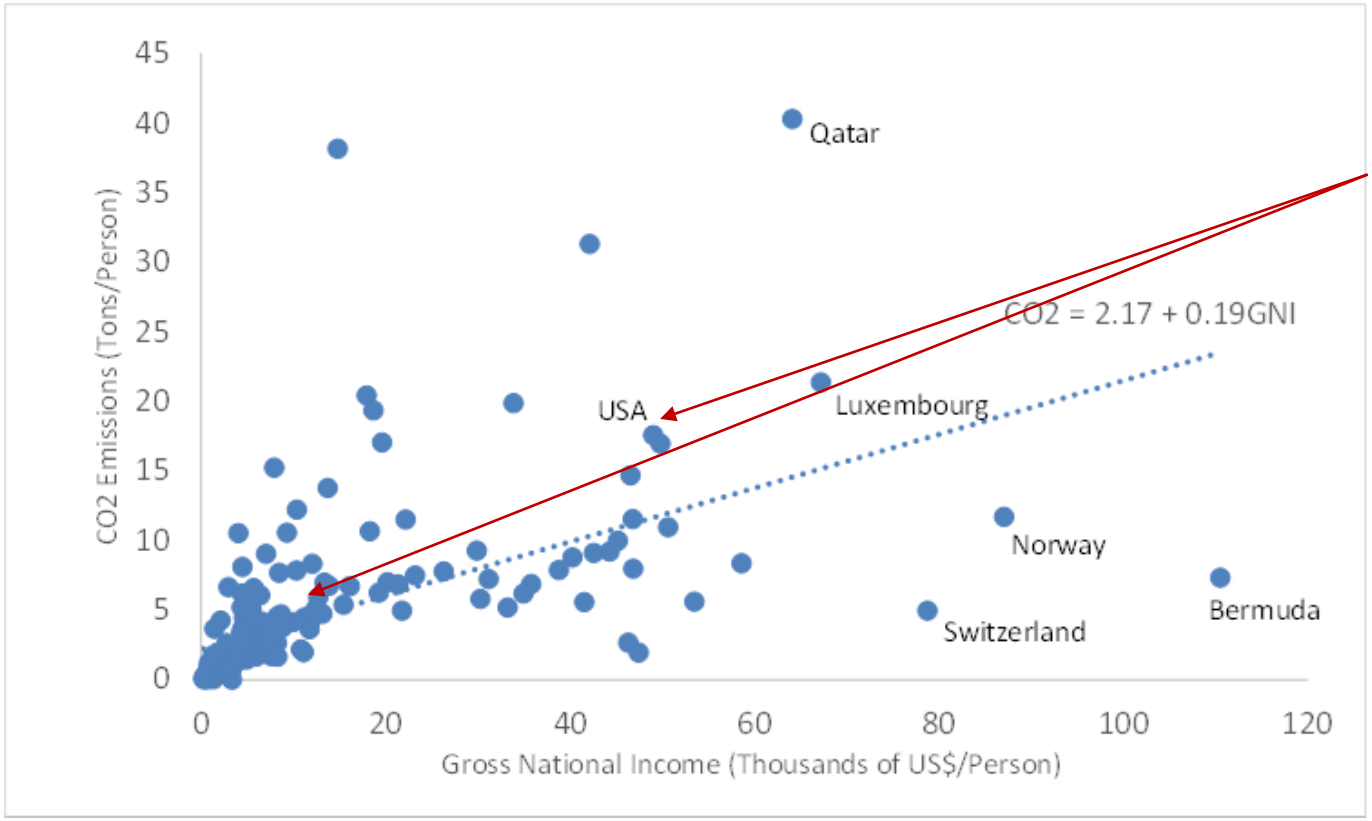
Example 1: Variance is correlated with RH Variable



What happens if we drop one of these?

Figure 8.1. 2010 CO_2 Emissions for 183 Countries. Data Source: World Bank World Development Indicators. http://data.worldbank.org/data-catalog/world-development-indicators

Using Average GNI per Person



Now what happens if we drop one of these?

- Still have lots of heteroscedasticity, but the extremes are less extreme than in the first figure

Example 2: Working with Average Data

Model for person j: $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j$

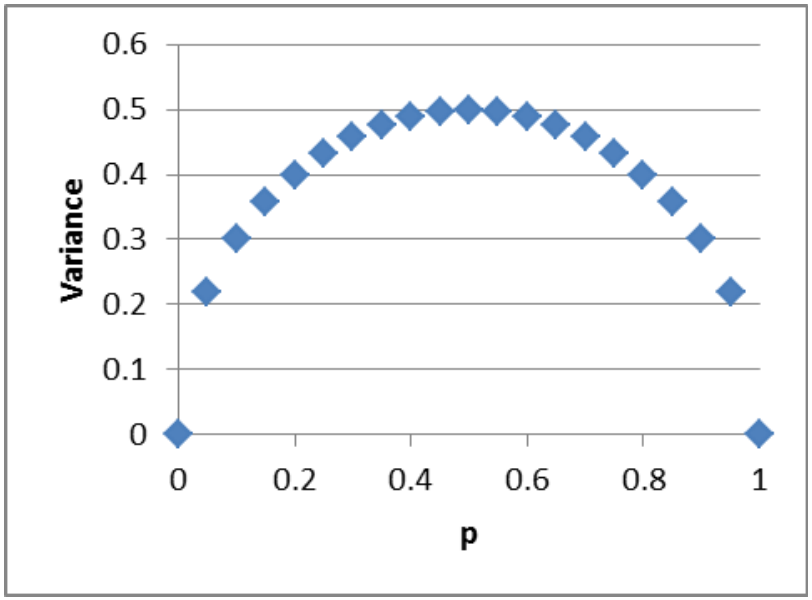
You estimate (country i): $\overline{Y}_i = \beta_0 + \beta_1 \overline{X}_i + \overline{\varepsilon}_i, \qquad \overline{\varepsilon}_i = \frac{1}{n}\sum_{j=1}^{n_i} \varepsilon_{ij}$

$$Var(\varepsilon_{ij}) = \sigma^2 , \quad \text{so} \qquad Var(\overline{\varepsilon}_i) = \frac{\sum_{j=1}^{n_i} Var(\varepsilon_{ij})}{n_i^2} = \frac{n_i \sigma^2}{n_i^2} = \frac{\sigma^2}{n_i}$$

- The variance is lower for bigger countries (which makes sense, right?)

Example 3: Modeling Probabilities

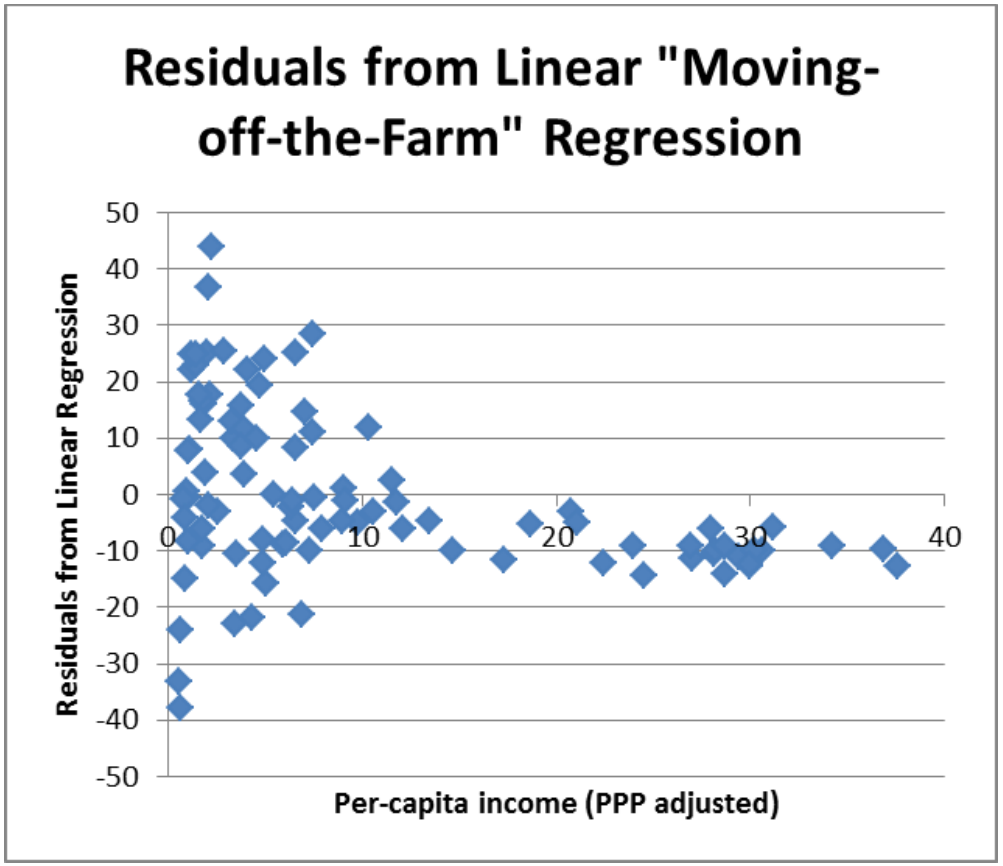
- *Suppose you drop a laptop to see whether it breaks*
- B_i=1 if laptop breaks,
  0 otherwise
- p=Prob(breaks)
- (1-p) = Prob (doesn't break)
- Variance = $p(1-p)$
- HEIGHT=height you drop the laptop from



$$B_i = \beta_0 + \beta_1 HEIGHT_i + \varepsilon_i$$

Low height: never breaks; Large height: always breaks;
in between, the variance is greater than zero

Example 4: Moving Off the Farm (Ch. 7)



$$AGL_i = \beta_0 + \frac{\beta_1}{PCY_i} + \varepsilon_i$$

- Variance decreases to almost nothing in rich countries (where nobody wants to do hired farm work)

Two Solutions

1. Test and fix after-the-fact (ex-post)

2. Change the model to eliminate the heteroscedasticity

Testing: Let Every Observation Have Its Own Variance

- Heteroskedasticity means different variances for different observations
- The squared residual is a good proxy for the variance σ_i^2
  - It's different for each observation
  - It even looks like a variance!
- So why not regress the squared residuals on all the RH variables and see if there's a correlation?

Breusch-Pagan Test

- Estimate the regression:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

- Save the residuals and square them
- Regress the squared residuals on X_1 and X_2
  - For the two RH variable case:

$$e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$$

- Under homoscedasticity, NR_a^2 is distributed as a chi-squared with df equal to the total number of right-hand variables in the regression (*df=2*).

White's Test (More Flexible than Breusch-Pagan. Why?)

- Estimate the regression:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

- Save the residuals and square them
- Regress the squared residuals on X_1, X_2, their squares, and their interactions
  - For the two RH variable case:

Careful with dummies!

$$e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{1i}^2 + \alpha_3 X_{2i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{1i} X_{2i} + u_i$$

- Under homoscedasticity, NR_a^2 is distributed as a chi-squared with df equal to the total number of right-hand variables in the White regression, a (here, $a=5$).

Set up your Breusch-Pagan or White test

$H_0 : \sigma_i^2 = \sigma^2$ for all i vs $H_1 : \text{not } H_0$

Test statistic: NR_a^2

Reject the null hypothesis if: $NR_a^2 > \chi_a^2$

Rejecting the null implies that have heteroskedasticity

White Test of Moving Off the Farm

$$e_i^2 = \alpha_0 + \alpha_1 PCY_i + \alpha_2 PCY_i^2 + u_i$$

Variables in White's Auxiliary Regression	Estimated Coefficient	Standard Error
PCY	-38.27	12.41
PCY-squared	0.91	0.37
Constant	418.11	61.47
R-squared	0.14	
N	95.00	
N*R-squared	13.09	
Critical Chi-Square at α=.05	5.99	

Basis for most tests of heteroskedasticity

- White's test shows a quadratic relationship between PCY and the variance
- NR_a^2 exceeds the critical Chi-Square value, so we reject the null of homoskedasticity

Fixing the Problem at Its Source

- If your data are averages:

$$\overline{Y}_i = \beta_0 + \beta_1 \overline{X}_i + \overline{\varepsilon}_i$$

$$Var(\overline{\varepsilon}_i) = \frac{\sigma^2}{n_i}$$

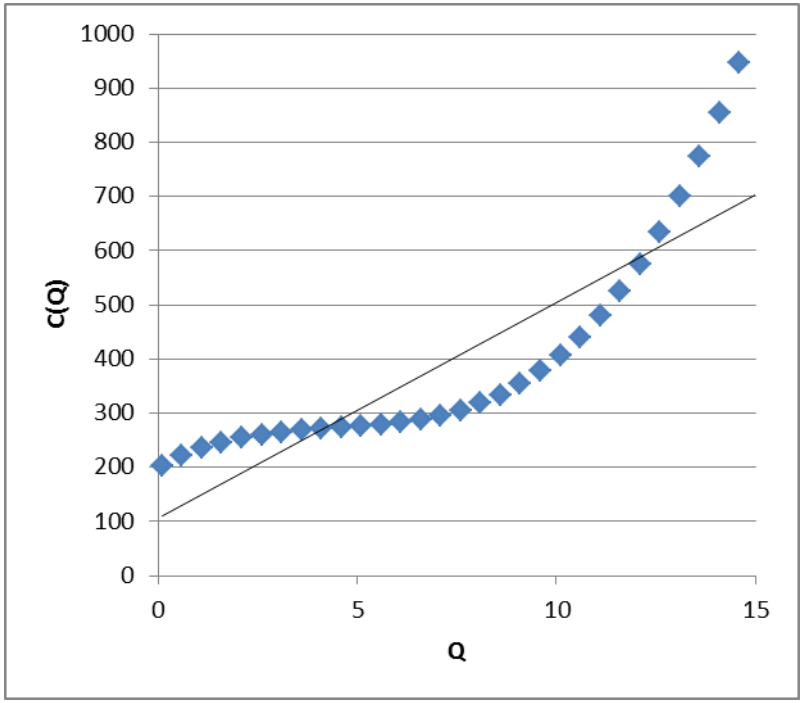
- We know the cause of the heteroskedasticity, so transform the regression equation by multiplying through by $\sqrt{n_i}$

$$\sqrt{n_i}\overline{Y}_i = \sqrt{n_i}\beta_0 + \beta_1\sqrt{n_i}\overline{X}_i + \sqrt{n_i}\overline{\varepsilon}_i$$

$$Var(\sqrt{n_i}\overline{\varepsilon}_i) = \frac{n_i^2\sigma^2}{n_i^2} = \sigma^2$$

Problem Fixed!

Fix It By Getting the Model Right



- Fit a line instead of a cubic and it LOOKS like the variance is related to Q (heteroskedasticity)
- Fit a cubic and you should get rid of the problem

What If the Model's Right and There's Still Heteroskedasticity?

- White's solution: Replace s^2 with $\hat{e}_i^2$ when calculating the variance

Before:

$$s_{b_1}^2 = \frac{\sum_{i=1}^{N} x_i^2 s^2}{\left(\sum_{i=1}^{N} x_i^2\right)^2} = \frac{s^2}{\sum_{i=1}^{N} x_i^2}$$

White's Method

(For simple regression;
for multiple, use v method):

$$Vb_1 = \frac{\sum_{i=1}^{N} x_i^2 e_i^2}{\left(\sum_{i=1}^{N} x_i^2\right)^2}$$

Heteroskedasticity-Consistent Estimator, HCE

Stata: reg y x1 x2, robust

Moving Off the Farm With and Without White's Correction

Variable	Regression A: Uncorrected			Regression B: Corrected		
	Estimated Coefficient	Standard Error	t-statistic	Estimated Coefficient	Standard Error	t-statistic
1/PCY	65.71	4.04	16.25	65.71	5.86	11.21
Constant	11.81	2.05	5.77	11.81	1.96	6.02
R-squared	0.74			0.74		
N	95			95		

- Here, the heteroskedasticity is not enough to change hypothesis test results, but it has a *big* effect on confidence intervals
  - The corrected standard error is almost 50% larger!
- In other cases, test results can change, too

White-corrected Confidence Intervals

Uncorrected Regression:

$$65.71 \pm 4.04 * 1.98 = 65.71 \pm 8.01$$

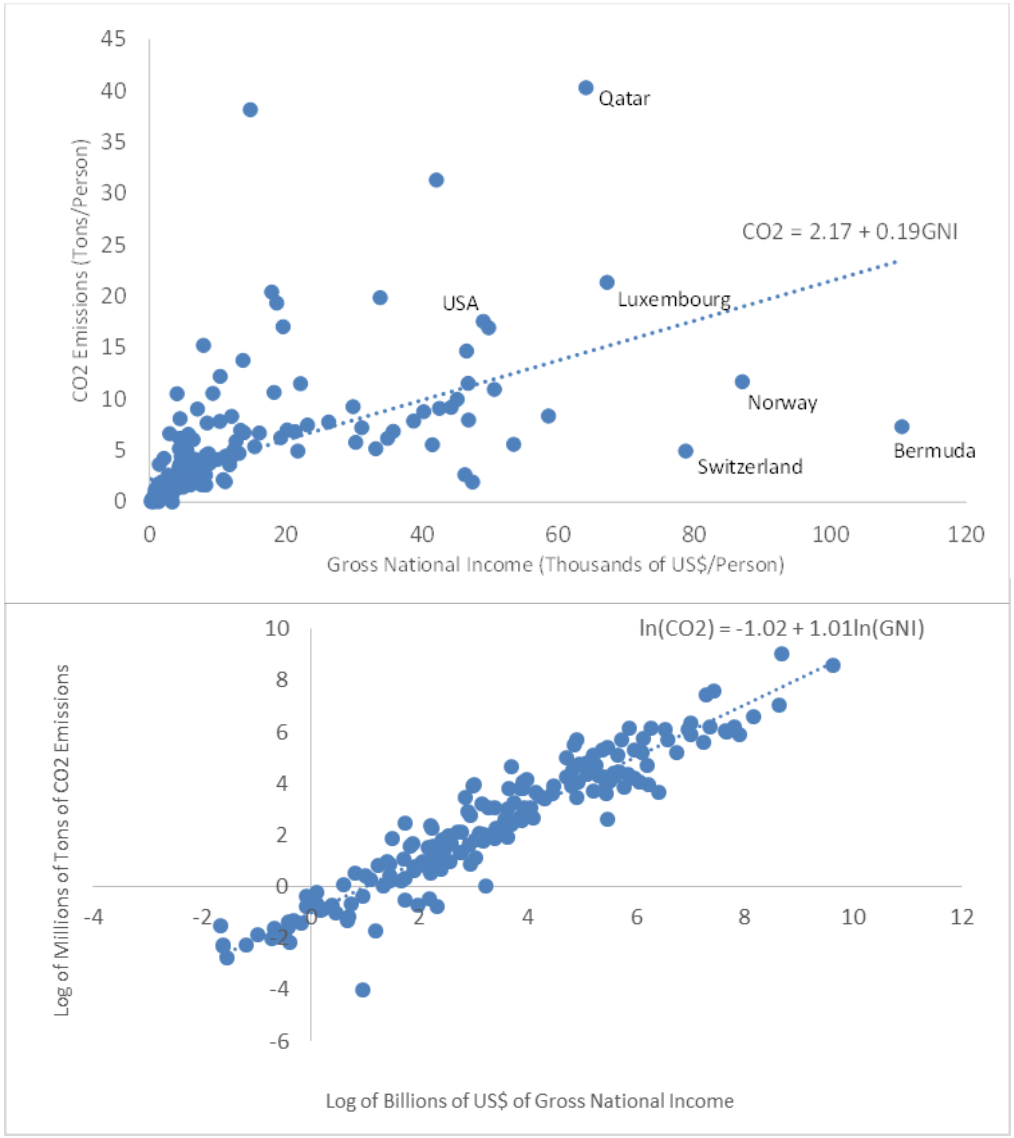
Robust Regression:

$$65.71 \pm 5.86 * 1.98 = 65.71 \pm 11.61$$

In this example, we get a confidence interval that is quite a bit smaller than it should be when we ignore heteroskedasticity

Log Regression Gets Rid of Heteroskedasticity in CO2 Study

Taking the log compresses the huge differences in income evident in figure 8.2. Country GNIs range from $0.19 billion to $15,170 billion. The log ranges from –1.66 to 9.63.



Heteroskedasticity-Robust Standard Errors

5% critical value for $\chi^2_{(2)} = \mathbf{5.99}$

Uncorrected	Robust
$CO2_i = 12282.71 + 0.44GNI_i + e_i$ (36713.99) (0.03)	$CO2_i = 12282.71 + 0.44GNI_i + e_i$ (16752.29) (0.12)
Sample Size = 182	Sample Size = 182
R-squared = 0.61	R-squared = 0.61
White test: $NR^2 = 50.84$	

Uncorrected	Robust
$\frac{CO2_i}{POP_i} = 2155.05 + 0.20\frac{GNI_i}{POP_i} + e_i$ (458.28) (0.02)	$\frac{CO2_i}{POP_i} = 2155.05 + 0.20\frac{GNI_i}{POP_i} + e_i$ (407.25) (0.04)
Sample Size = 182	Sample Size = 182
R-squared = 0.33	R-squared = 0.33
White test: $NR^2 = 16.52$	

Uncorrected	Robust
$\ln(CO2_i) = -1.09 + 1.01\ln(GNI_i) + e_i$ (0.27) (0.026)	$\ln(CO2_i) = -1.09 + 1.01\ln(GNI_i) + e_i$ (0.25) (0.023)
Sample Size = 182	Sample Size = 182
R-squared = 0.90	R-squared = 0.90
White test: $NR^2 = 0.73$	

Success!

What We Learned

- Heteroskedasticity means that the error variance is different for some values of *X* than for others; it can indicate that the model is misspecified.

- Heteroskedasticity causes OLS to lose its "best" property and it causes the standard error formula to be wrong (i.e., estimated standard errors are biased).

- The standard errors can be corrected with White's heteroskedasticity-robust estimator.

- Getting the model right by, for example, taking logs can sometimes eliminate the heteroskedasticity problem.