Model-level problems: Problems with Variance

D. Alex Hughes

November 25, 2014

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables

2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!

3. Assumptions about X's
   - Measurement Error

Remember our assumptions

- We have the right model
- Errors are iid $N(0, \sigma^2)$
- X's are fixed, or orthogonal to errors

We will consider these, though not entirely in order.

The Right Model

Presume there is indeed a linear relationship between Y and our X's.

Three ways to mess this up:

1. Non-linearity
2. Including Irrelevant Variables
3. Omitting Relevant Variables

The Right Model

Presume there is indeed a linear relationship between Y and our X's. Three ways to mess this up:

1. Non-linearity
2. Including Irrelevant Variables
3. Omitting Relevant Variables

The Right Model

Presume there is indeed a linear relationship between Y and our X's.
Three ways to mess this up:

1. Non-linearity
2. Including Irrelevant Variables
3. Omitting Relevant Variables

1. Non-Linearity

$$Y = X\beta$$
$$\rightarrow E[\epsilon|X] = 0$$

Cost of being wrong:

- Meaningless estimates
- Conditional expectation (on Xs) doesn't reliably tell us about Y
- Error in *functional form*

Solutions:

- Transform variables appropriately
- Estimate a nonlinear model, using nonlinear least squares, or maximum likelihood methods.

1. Linearity: Identifying

This looks actually a lot like what we did with residuals last week.

- Partial residual plot of $E_i|X$
- Unfortunately, can not provide *correct* functional forms; only identify *incorrect* ones...

Example

```
Go to code chunk 1
```

Including Irrelevant Variables

What is the cost of "throwing in the kitchen sink"? An irrelevant variable in this case means one that does not have an impact on Y, and is not correlated with the errors.

$$Cov(X_O, \epsilon) = 0$$

- Our estimates of β and of Σ, the variance-covariance matrix, are unbiased.
- Our estimates, however, are less efficient. The standard errors on our estimates become larger, and we are less likely to reject the null when we should.
- This is bad, but not as bad estimating with bias or committing type I error (maybe?)

Including Irrelevant Variables

What is the cost of "throwing in the kitchen sink"? An irrelevant variable in this case means one that does not have an impact on Y, and is not correlated with the errors.

$$Cov(X_O, \epsilon) = 0$$

- Our estimates of β and of Σ, the variance-covariance matrix, are unbiased.
- Our estimates, however, are less efficient. The standard errors on our estimates become larger, and we are less likely to reject the null when we should.
- This is bad, but not as bad estimating with bias or committing type I error (maybe?)

Including Irrelevant Variables

What is the cost of "throwing in the kitchen sink"? An irrelevant variable in this case means one that does not have an impact on Y, and is not correlated with the errors.

$$Cov(X_O, \epsilon) = 0$$

- Our estimates of β and of Σ, the variance-covariance matrix, are unbiased.
- Our estimates, however, are less efficient. The standard errors on our estimates become larger, and we are less likely to reject the null when we should.
- This is bad, but not as bad estimating with bias or committing type I error (maybe?)

Go to code chunk 2

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables
2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!
3. Assumptions about X's
   - Measurement Error

Omitted Variable Bias

- What if you exclude relevant variables?
- The other predictors will try to "make up the difference" - and will do so unless they are TOTALLY unrelated to the excluded variable.
- Example:
  - $Success_{PhD} = IQ + GPA - GRE + \epsilon$
  - What if you leave out IQ?
  - GPA is likely correlated with IQ, so it will try to step in and make up the difference:
  - BUT - this means we've now overstated the true effect of GPA!

Omitted Variable Bias

- What if you exclude relevant variables?
- The other predictors will try to "make up the difference" - and will do so unless they are TOTALLY unrelated to the excluded variable.
- Example:
  - $Success_{PhD} = IQ + GPA - GRE + \epsilon$
  - What if you leave out IQ?
  - GPA is likely correlated with IQ, so it will try to step in and make up the difference:
  - BUT - this means we've now overstated the true effect of GPA!

Go to the board, Alex!

Omitted Variables, More Formally

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Estimate β_1 without including β_2. Call this estimate $\tilde{\beta}_1$

$$\begin{aligned}
\tilde{\beta}_1 &= (X_1' X_1)^{-1} X_1' Y \\
&= (X_1' X_1)^{-1} X_1' [X_1 \beta_1 + X_2 \beta_2 + \epsilon] \\
&= (X_1' X_1)^{-1} X_1' X_1 \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2 + (X_1' X_1)^{-1} X_1' \epsilon \\
E(\tilde{\beta}_1) &= \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2 + 0 \\
&= \beta_1 + \frac{Cov(x_1, x_2)}{Var(x_1)} \beta_2
\end{aligned}$$

Go to code chunk 4

Omitted Variable Bias

- If X2 is omitted, slope estimates for X1 will not be biased unless X1 and X2 are correlated.
- If they are correlated, X1 will try to make up for X2, instead of leaving it for the residuals.
- Key factor is the variance in X_1 versus the covariance of X_1 and X_2
- However, the omission of a relevant variable will always bias estimates of coefficients' standard errors upward, even if the excluded variable is orthogonal to the included variables.

Omitted Variable Bias

- If X2 is omitted, slope estimates for X1 will not be biased unless X1 and X2 are correlated.
- If they are correlated, X1 will try to make up for X2, instead of leaving it for the residuals.
- Key factor is the variance in X_1 versus the covariance of X_1 and X_2
- However, the omission of a relevant variable will always bias estimates of coefficients' standard errors upward, even if the excluded variable is orthogonal to the included variables.

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables

2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!

3. Assumptions about X's
   - Measurement Error

Assumptions about Errors

$$\epsilon_i \sim N(0, \sigma^2)$$

Assumptions about Errors

$$\epsilon_i \sim N(0, \sigma^2)$$

How do things go wrong?

1. The mean isn't zero
2. The errors are not normally distributed
3. The variance of errors is different across observations
4. The covariance of errors across observations is not zero

Assumptions about Errors

$$\epsilon_i \sim N(0, \sigma^2)$$

How do things go wrong?

1. The mean isn't zero
2. The errors are not normally distributed
3. The variance of errors is different across observations
4. The covariance of errors across observations is not zero

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables

2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!

3. Assumptions about X's
   - Measurement Error

What if the mean of the error distribution isn't zero?

This is the easy case. Why?

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$
$$\epsilon \sim N(\theta, \sigma^2)$$

What if the mean of the error distribution isn't zero?

This is the easy case. Why?

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \epsilon &\sim N(\theta, \sigma^2) \\ E[\beta_0^*] &= \beta_0 + \theta \end{aligned}$$

What if the errors are not normally distributed?

- Normality: we either assume it or we rely on the central limit theorem and big datasets.
- Even without normality, regression is the best linear unbiased estimator (BLUE) of the model parameters β. This comes from the Gaus-Markov Theorem.
- However, the normality assumption provides t statistics for hypothesis testing - so inference is suspect without normality.
- If you don't feel comfortable being normal...You could specify an alterative distribution and come up with your own standard errors for hypothesis testing
- Or you could use a *nonparametric* method to run the regresion, depending on your beliefs about the errors.
- Check for normality: formal test, or quantile plots of residuals. What's a quantile plot?

What if the errors are not normally distributed?

- Normality: we either assume it or we rely on the central limit theorem and big datasets.
- Even without normality, regression is the best linear unbiased estimator (BLUE) of the model parameters β. This comes from the Gaus-Markov Theorem.
- However, the normality assumption provides t statistics for hypothesis testing - so inference is suspect without normality.
- If you don't feel comfortable being normal...You could specify an alterative distribution and come up with your own standard errors for hypothesis testing
- Or you could use a *nonparametric* method to run the regresion, depending on your beliefs about the errors.
- Check for normality: formal test, or quantile plots of residuals. What's a quantile plot?

What if the errors are not normally distributed?

- Normality: we either assume it or we rely on the central limit theorem and big datasets.
- Even without normality, regression is the best linear unbiased estimator (BLUE) of the model parameters β. This comes from the Gaus-Markov Theorem.
- However, the normality assumption provides t statistics for hypothesis testing - so inference is suspect without normality.
- If you don't feel comfortable being normal...You could specify an alterative distribution and come up with your own standard errors for hypothesis testing
- Or you could use a *nonparametric* method to run the regresion, depending on your beliefs about the errors.
- Check for normality: formal test, or quantile plots of residuals. What's a quantile plot?

What if the errors are not normally distributed?

- Normality: we either assume it or we rely on the central limit theorem and big datasets.
- Even without normality, regression is the best linear unbiased estimator (BLUE) of the model parameters β. This comes from the Gaus-Markov Theorem.
- However, the normality assumption provides t statistics for hypothesis testing - so inference is suspect without normality.
- If you don't feel comfortable being normal...You could specify an alterative distribution and come up with your own standard errors for hypothesis testing
- Or you could use a *nonparametric* method to run the regresion, depending on your beliefs about the errors.
- Check for normality: formal test, or quantile plots of residuals. What's a quantile plot?

What if the errors are not normally distributed?

- Normality: we either assume it or we rely on the central limit theorem and big datasets.
- Even without normality, regression is the best linear unbiased estimator (BLUE) of the model parameters β. This comes from the Gaus-Markov Theorem.
- However, the normality assumption provides t statistics for hypothesis testing - so inference is suspect without normality.
- If you don't feel comfortable being normal...You could specify an alterative distribution and come up with your own standard errors for hypothesis testing
- Or you could use a *nonparametric* method to run the regresion, depending on your beliefs about the errors.
- Check for normality: formal test, or quantile plots of residuals. What's a quantile plot?

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables

2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!

3. Assumptions about X's
   - Measurement Error

What if the variance of the errors is not constant?

Instead of:

$$\epsilon \sim N(0, \sigma^2)$$

or variance-covariance matrix:

$$\sigma^2 \mathbf{I}$$

We have:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

What if the variance of the errors is not constant?

Instead of:

$$\epsilon \sim N(0, \sigma^2)$$

or variance-covariance matrix:

$$\sigma^2 \mathbf{I}$$

We have:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity

Or even worse, we have:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \sigma_2^2 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity

Or even worse, we have:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \sigma_2^2 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity

Let's start simple:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity

If we know the σ_i's up to some proportion, we can write:

$$\mathbf{\Sigma} = \sigma^2 * \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

How might this happen?

Heteroskedasticity

Suppose we have:

- Data on average school achievement tests for all of Texas
- The proportion voting in census blocks
- Or something like that...

Heteroskedasticity

What's the sampling distribution of the sample proportion?
What's the sampling distribution of the sample mean?

$$\hat{p} \sim N(p, \frac{p * (1-p)}{n})$$

$$\bar{x} \sim t(\mu, \frac{se}{n-1})$$

- We might expect the error to be partly related to population size.
- Implies that we weighting larger samples more heavily than smaller samples
- More information about large sample than small samples

Heteroskedasticity

What's the sampling distribution of the sample proportion?
What's the sampling distribution of the sample mean?

$$\hat{p} \sim N(p, \frac{p * (1-p)}{n})$$

$$\bar{x} \sim t(\mu, \frac{se}{n-1})$$

- We might expect the error to be partly related to population size.
- Implies that we weighting larger samples more heavily than smaller samples
- More information about large sample than small samples

Heteroskedasticity, Cost

- $\hat{\beta}$ is unbiased. (Why?) Hooray!
- However, $\hat{\beta}$ is less efficient than an alternative
- p-values from t-tests and F-tests will be *wrong*

Heteroskedasticity, Cost

- $\hat{\beta}$ is unbiased. (Why?) Hooray!
- However, $\hat{\beta}$ is less efficient than an alternative
- p-values from t-tests and F-tests will be *wrong*

Heteroskedasticity, Cost

- $\hat{\beta}$ is unbiased. (Why?) Hooray!
- However, $\hat{\beta}$ is less efficient than an alternative
- p-values from t-tests and F-tests will be *wrong*

Testing for heteroskedasticity

- Best way: plot residuals!
- White test (p. 297)
- Breusch Pagan (p. 296)
- Breusch Godfrey

We'll return to these tests in a bit

Heteroskedasticity, Case 1

Case 1: the error variance varies as a function of something we do know...

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

and

$$\mathbf{V}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

for example,

$$\mathbf{W} = \sigma^2 * \begin{bmatrix} \frac{1}{n_1} & 0 & \cdot & 0 \\ 0 & \frac{1}{n_2} & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{n_m} \end{bmatrix}$$

Easy solution

Divide each observation of the linear model by the *known* certainty σ_i

$$\frac{Y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{X_{1,i}}{\sigma_i} + \frac{\epsilon_i}{\sigma_i}$$
$$Y_i^* = X^* \beta + u_i^*$$

Where:

- $u_i^* = \frac{\epsilon_i}{\sigma_i}$
- $V(u_i^*) = V(\frac{\epsilon_i}{\sigma_i}) = 1$

Heteroskedasticity

Go to code block 5

Solution for Case 1

```
library(survey)
data(api)
m0 <- lm(pcttest ~ growth + as.factor(stype)
         , data = apiclus2)
m1 <- lm(pcttest ~ growth + as.factor(stype)
         , weights = pw, data = apiclus2)
```

Problems with WLS

Unless we have constructed sampling weights (maybe our sample was constructed by the inverse probability of answering like yesterday's job candidate?), *we don't know* σ_i.

Two ways forward:

1. We could *guess* for σ_i (hard)
2. Keep OLS for $\hat{\beta}$ (it *is* unbiased, good!), but develop a fix for $V(\hat{\beta})$ to fix the bias (bad...)

Problems with WLS

Unless we have constructed sampling weights (maybe our sample was constructed by the inverse probability of answering like yesterday's job candidate?), *we don't know* σ_i.

Two ways forward:

1. We could *guess* for σ_i (hard)
2. Keep OLS for $\hat{\beta}$ (it *is* unbiased, good!), but develop a fix for $V(\hat{\beta})$ to fix the bias (bad...)

Heteroskedasticity, Case 2

Case 2: the error variance varies, but that's all we know....

$$\mathbf{\Sigma} = \begin{bmatrix} ? & 0 & \cdot & 0 \\ 0 & ? & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & ? \end{bmatrix}$$

What to do?

Heteroskedasticity, Case 2

Recall $V(\hat{\beta})$.

$$V(\hat{\beta}) = (X'X)^{-1}X'V(y)X(X'X)^{-1}$$

If $V(y) = \sigma_\epsilon^2$, then the classic case is

$$V(\hat{\beta}) = \sigma_\epsilon^2(X'X)^{-1}$$

But, what if different variances, but independent ($Cov(X_i, X_j) = 0$)?

- Define $\Sigma \equiv V(y) = diag(VCOV) = \{\sigma_1^2 \cdots \sigma_n^2\}$
- Then,

$$V(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

Zut alors! We don't know Σ...

Heteroskedasticity, Case 2

But, we can estimate it! Use the residuals as weights

$$\hat{\mathbf{V}}(\hat{\beta})_{\mathbf{HC}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} E_1^2 & 0 & \cdot & 0 \\ 0 & E_2^2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & E_n^2 \end{bmatrix}$$

This is the **White Heteroskedastic Consistent** estimator.

White Standard Errors

What's good?

- Does not require assumptions about form of heteroskedasticity
- Unbiased, consistent estimator of VCOV

What's bad?

- Inefficient
- If we knew the form of heteroskedasticity we would use WLS

Heteroskedasticity, Case 2

This implies several steps:

1. Run an unweighted regression.
2. Calculate residuals
3. Construct variance-covariance matrix.
4. or...

Go to code block 8

Heteroskedasticity, Case 2

This implies several steps:

1. Run an unweighted regression.
2. Calculate residuals
3. Construct variance-covariance matrix.
4. or...

Go to code block 8

Heteroskedasticity, Case 2

This implies several steps:

1. Run an unweighted regression.
2. Calculate residuals
3. Construct variance-covariance matrix.
4. or...

`Go to code block 8`

What type of error to use?

There are *different* penalizations of outliers.

- Probably beyond class today
- For an internet-searchable discussion see **this link** to scholars at Indiana University

1. What we have shown today is known as the "HC0" estimate, or the White Heteroskedastic Errors.
2. Formula (12.6) on p. 275 gives the HC3 "jackknife" errors

$$\begin{aligned}\tilde{V}^*(\hat{\beta}) &= (X'X)^{-1}X'\hat{\Sigma}^*X(X'X)^{-1} \\ &= (X'X)^{-1}X'diag\left[\frac{e_i^2}{(1-h_{ii})^2}\right]X(X'X)^{-1}\end{aligned}$$

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables

2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!

3. Assumptions about X's
   - Measurement Error

Heteroscedasticity

- $\sigma_i = f(z_i)$?
- Covariance between ϵ_i and ϵ_j

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \sigma_2^2 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

- Time Series
- Notice that there are potentially more parameters than there are observations....

We're going to leave this for now...

How do you know if it is a problem?

- Plot residuals
- Test for it.
- Intuition tells you it is a problem.

1. Plotting Residuals

- Plot fitted values against Studentized residuals
- If we're fitting well, should be no relationship – normal "cloud"
- If we're missing more at one or several points in the $\hat{y}$ range, need HC errors

2. White's Test for Heteroscedasticity

$$H_0 : \sigma_i^2 = \sigma^2 \forall i$$
$$H_A : \neg H_0$$

Test statistic:

- $nR^2 \sim \chi^2_{P-1}$ where:
  - n is the number of observations
  - R^2 is the r-squared from a regression of the e_i^2 on all $X's$ and all squares and cross products of the x's, and a constant
  - P is the number of covariates in the test regression

White's Test for Heteroscedasticity

Test statistic:

$$nR^2 \sim \chi^2_{P-1}$$

- Extremely general and requires no assumptions about heteroscedasticity....
- But...
  - Potentially lower power
  - May detect other problems apart from heteroscedasticity
  - Does not offer guidance on how to proceed.

Practically:

`Go to code block 7`

Breusch-Pagan/Godfrey LM Test

Test based on Lagrange multiplier. Model:

$$\sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha' \mathbf{z_i})$$

Null:

$$\alpha = 0$$

Test Statistic:

$$LM = \frac{1}{2}(\mathbf{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{g})$$

```
library(lmtest)
bptest(...)
```

Breusch-Pagan/Godfrey LM Test

Essentially, this is the vector of estimated regression premultiplied by the dependent variable in that regression. Specifically:

- $g = \frac{e^2}{\mathbf{e}'\mathbf{e}/\mathbf{n}} - 1$
- z is a matrix of predictors of heteroscedasticity, with $P - 1$ covariates and an intercept.

$$LM \sim \chi^2_{P-1}$$

Null: All the error variances are equal;
Alternative: Error variances are linear product of RHS regressor

`bgtest(...)`

Heteroscedasticity

- This is a huge topic that requires extended study and we don't have time to do more with it.
- There are a diverse set of tests for detecting and fixing the many potential forms of heteroscedasticity.
- Consider courses in Econ or IRPS, or just read Greene's Econometrics.

How bad is it?

Impact depends on

- Sample size
- Degree of variation in σ_i^2
- Configuration of Xs
- Relationship between X and $V(\sigma_i^2)$

How bad is it?

We can think of the penalty we pay as a regularized cost – *relative* to the OLS estimator.

- Suppose $Y_i = \alpha + \beta X_i + \epsilon_i$ where errors are independent, and normal with $\mu = 0$ but standard deviations proportional to X.
- So, $\sigma_i = \sigma_\epsilon X_i$ – the wedge!
- then, the OLS estimator (B) of β is less efficient that the WLS estimator ($\hat{\beta}$) (which is BLUE in this case).

How bad is it?

Then,

- We can easily derive the sampling variance for B and $\hat{\beta}$ (uh, oh...) and can compare the ratio of the sampling variances:

$$\frac{V(\hat{\beta})}{V(B)}$$

- For concreteness – suppose $X \sim U[x_0, ax_0]$
- Then, a is the ratio of the largest to the smallest value of X (and therefore the largest to smallest $sigma_i$).
- Then, by the sample variance of B and $\hat{\beta}$:
  - The relative precision of the OLS estimator stabilized quickly, for as small as n=20.
  - In this case, if a $= 2$, B_{OLS} is 98% as efficient at $\hat{\beta}_{WLS}$
  - Even if $a = 10$ the efficiency is $> 93\%$

So, we're really only worried about this when the variance of the errors is greater than a factor of 3 (or ten times). To be conservative, you could use a factor of 2 (or about 4 times).

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables
2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!
3. Assumptions about X's
   - Measurement Error

Assumptions about X's

1. We assume that X's are fixed, or uncorrelated with errors.
2. When they are correlated, regression leads to biased estimates, regardless of sample size.
3. One common solution is the use of an instrument, via an "instrumental variables regression", or a "two-stage least square regression".

What the assumption means

We assume that X's are fixed, or uncorrelated with errors.

- Knowing that X is big doesn't change our expectation of the errors - we aren't more likely to see a positive error with a big x than a little x.

What the assumption means

We assume that X's are fixed, or uncorrelated with errors.

- Knowing that X is big doesn't change our expectation of the errors - we aren't more likely to see a positive error with a big x than a little x.

When they are correlated, we get biased estimates

1. Consider a positive correlation between X and ϵ. That means that when there is a big x, there is a big positive ϵ. Given the solid true line, this will lead to the observed data, and an estimated dashed line.
2. The direction of the bias will depend on the correlation between X and ϵ. In this case, the correlation is positive, so the estimate is inflated.
3. This bias does *not* go away as *n* gets big.

Go to code block 9

Why might X and ϵ be correlated?

1. Measurement error: X is measured with error
2. Omitted Variable Bias: A relevant variable is excluded, and correlated with some other variable in the model.
3. Simultaneity: two variables codetermine each other. Price and demand are classic examples.
4. Temporal codetermination: cross-period dependence of factors.

Note that all of these are sometimes called "endogeneity". One solution is an Instrumental Variables approach using two-stage least squares.

Why might X and ϵ be correlated?

1. Measurement error: X is measured with error
2. Omitted Variable Bias: A relevant variable is excluded, and correlated with some other variable in the model.
3. Simultaneity: two variables codetermine each other. Price and demand are classic examples.
4. Temporal codetermination: cross-period dependence of factors.

Note that all of these are sometimes called "endogeneity". One solution is an Instrumental Variables approach using two-stage least squares.

1. The Wrong Model
   - Non Linearity: Functional Form
   - Including Irrelevant Variables
   - Omitting Relevant Variables

2. Problems in the Errors
   1. Mean error isn't zero
   2. Non-normal errors
   3. Nonconstant error Variance
   4. Covariance for days!

3. Assumptions about X's
   - Measurement Error

Measurement Error

Not a problem when the error is in Y:

We are modeling Y, but only observe $Y + \gamma$, where γ are iid with mean 0. What does this do?

$$Y + \gamma = \beta_0 + \beta_1 x + \epsilon$$

$$Y = \beta_0 + \beta_1 x + \epsilon - \gamma$$

Impact on $E(\hat{\beta}_1)$? Impact on $Var(\hat{\beta}_1)$?

Measurement Error

Not a problem when the error is in Y:
We are modeling Y, but only observe $Y + \gamma$, where γ are iid with mean 0.
What does this do?

$$Y + \gamma = \beta_0 + \beta_1 x + \epsilon$$

$$Y = \beta_0 + \beta_1 x + \epsilon - \gamma$$

Impact on $E(\hat{\beta}_1)$? Impact on $Var(\hat{\beta}_1)$?

```
Go to code block 10
```

Measurement Error

Measurement Error in Y

- Random measurement error in Y goes into the residuals if iid mean zero. No bias problem, just less efficient.

Measurement Error

Measurement Error in X's. A problem?

$$X = X* + \gamma$$

$$Y = \beta_0 + \beta_1 X* + \epsilon$$

where $\gamma \sim N(0, \psi^2)$.

$$X* = X - \gamma$$

$$Y = \beta_0 + \beta_1 X - \beta_1 \gamma + \epsilon$$

$$Y = \beta_0 + \beta_1 X - \upsilon$$

Measurement Error

Continuing...

$$Y = \beta_0 + \beta_1 X - \upsilon$$

- Problem: X is correlated with υ, because $X = X * + \gamma$, and γ is part of the error.
- Recall assumptions of regression: X's are fixed or not correlated with the errors.
- So what happens to our estimate of β?

When estimating blindly via OLS, we get

$$\hat{\beta}_1 \rightarrow \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\gamma^2}\beta_1$$

So on average, measurement error in Y disappears into the residuals, and measurement error in X *when just one X* biases our estimated coefficients (too bad for us).

When there is more than one independent variable, things get messy.

When estimating blindly via OLS, we get

$$\hat{\beta}_1 \rightarrow \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\gamma^2}\beta_1$$

So on average, measurement error in Y disappears into the residuals, and measurement error in X *when just one X* biases our estimated coefficients (too bad for us).
When there is more than one independent variable, things get messy.

Go to code block 11

One Solution: Instrumental Variables via 2SLS

The Logic

- We don't observe the true X, just $X + \gamma$
- There is some other variable Z that is correlated with X, but not with γ
    - Z is called the "instrument" - it is a proxy for X, and isn't correlated with the measurement error.
- We use the "unpolluted" covariance between X and Z to get a better estimate of β
- Examples?

Instrumental Variables - Solution

"Two Stage Least Squares"

- Regress the mismeasured X on Z, and get a predicted value of X, $\hat{X}$
- Regress Y on the "unpolluted" $\hat{X}$
- Since there is no correlation between Z and the measurement error, the $\hat{X}$ get us unbiased estimates
- This can be used to solve the problem of "endogeneity"
- Pretty amazing, there's just one catch... Where do you find a good instrument? Where should you look?

Institutions and Growth

- Do Institutions affect growth rates? IMPORTANT question.
- But over time, growth probably affects institutions too.
- So how do we get the independent impact of institutional change?
- Find something that affects institutions, but not growth.
- Settler mortality from hundreds of years ago?

Instrumental Variables

- Want a variable Z that is highly correlated with X, but not with γ
- We won't actually ever know if Z is correlated with the errors or not...
- Exogenous changes can help - randomized audits, and so on
- Also try creativity: # of checks, # of dead settlers, and so on.
- Note that behind every methods lies lots of assumptions...

Examples

- Health = a + b*Smoking
  Problem?
  Instrument for smoking?
  Taxes

Examples

- Health = a + b*Smoking
  Problem?
  Instrument for smoking?
  Taxes

Examples

- Health = a + b*Smoking
  Problem?
  Instrument for smoking?
  Taxes

Examples

- Tennancy = a + b*Land Conflict Problem?

  Instrument for Land Conflict?

  Number of Priests in 1700's

Examples

- Tennancy = a + b*Land Conflict
  Problem?
  Instrument for Land Conflict?
  Number of Priests in 1700's

Examples

- Tennancy = a + b*Land Conflict
  Problem?
  Instrument for Land Conflict?
  Number of Priests in 1700's

Examples

- Civic Values = a + b*Military Service
  Problem?
  Instrument for Military Service?
  Draft Eligibility
  Caveats?

Examples

- Civic Values = a + b*Military Service
  Problem?
  Instrument for Military Service?
  Draft Eligibility
  Caveats?

Examples

- Civic Values = a + b*Military Service
  Problem?
  Instrument for Military Service?
  Draft Eligibility
  Caveats?

Examples

- Civic Values = a + b*Military Service
  Problem?
  Instrument for Military Service?
  Draft Eligibility
  Caveats?

More generally

There are lots of ways regressions can go bad, but there is a big literature on detecting these and fixing them. We don't have time for them all, but most are easy to implement and programmed into stata. You can take an advanced class for more detail.