Introduction to Econometrics

Chapter 9 - Assessing Studies Based on Multiple Regression

Nishant Yonzan

Fall 2019

Goal of this class

- Internal versus External Validity
- Testing for Heteroskedasticity
- BP Test

Internal and External Validity

- Internal validity is satisfied if the statistical inference about the causal effects are valid for the population being studied.
- External validity would need the inferences and conclusions to be generalized from the population and setting studied to other populations and setting.
- Example: Think about the california test example.
- Is the slope, β_{str}, unbiased and consistent?
- Is this a valid estimate for NY?, WV?, NM?,. . .

Threats to Internal Validity

1. Omitted Variable Bias
2. Misspecification of Functional Form of Regression Model
3. Measurement Error
4. Missing Data and Sample Selection
5. Simultaneous Causality

In each case, OLS assumption #1 is violated: $E(u_i \mid X_{1i}, X_{2i}, ..., X_{ki}) \neq 0$.

1. Omitted Variable Bias

- If you have the data for omitted variables:
  - Be specific about your coefficient of interest.
  - Use a priori reason for adding variables.
  - Use statistical tests for questionable variables (t and F).
  - Provide all the potential specifications in tabular form.
- What if you don't have the data:
  - Panel data (next chapter)
  - IV method (chapter 12)
  - Randomized control experiments

2. Misspecification of Functional Form of Regression Model

- For continuous *dependent variable*: Use methods discussed in chapter 8 to modify functional forms.
- For discrete or binary *dependent variable*: Chapter 11 (we will get there. . .).

3. Measurement Error in X

- General regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- What you would like to measure is X_i, but what you do actually measure is $\tilde{X}_i$. Then the *error* is $X_i - \tilde{X}_i$.
- Regression model with mesurement error ($\tilde{X}_i$ instead of X_i):

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i]$$

- Rewrite $\nu_i = \beta_1 (X_i - \tilde{X}_i) + u_i$:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \nu_i$$

- If the error term is correlated with the $\tilde{X}_i$, then $\hat{\beta}_1$ will be biased and inconsistent.

Classical measurement error

- Suppose the error is purely random such that $\tilde{X}_i = X_i + \omega_i$
- If ω_i is purely random, then $corr(\omega_i, X_i) = 0$ and $corr(\omega_i, u_i) = 0$.
- Under this **classical measurement error model**

$$\hat{\beta}_1 \xrightarrow{plim} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\omega^2} \beta_1$$

- Bias in $\hat{\beta}_1$ is **towards zero** with random error.
- Non-random errors:
  - For "best guess" measurement error: $E[(\tilde{X}_i - X_i)\tilde{X}_i] = 0$, $\hat{\beta}_1$ is consistent but imprecise, $var(\nu_i) > var(u_i)$.
  - Intentional under(over)-reporting: $\tilde{X}_i = 0.9X_i \Rightarrow$ bias upward.

4. Missing Data and Sample Selection

1. Missing at random: There is no bias, but it reduces our sample size.
2. Missing regressor values: Same as above.
3. Missing Y due to selection process (**sample selection bias**):
   - For example of this type of bias, think about why older, say, baseball players are on average better than younger ones.

5. Simultaneous Causality

This can happen if the causality runs in the opposite direction: $Y_i \Rightarrow X_i$.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad (1)$$

and

$$X_i = \gamma_0 + \gamma_1 Y_i + \nu_i \qquad (2)$$

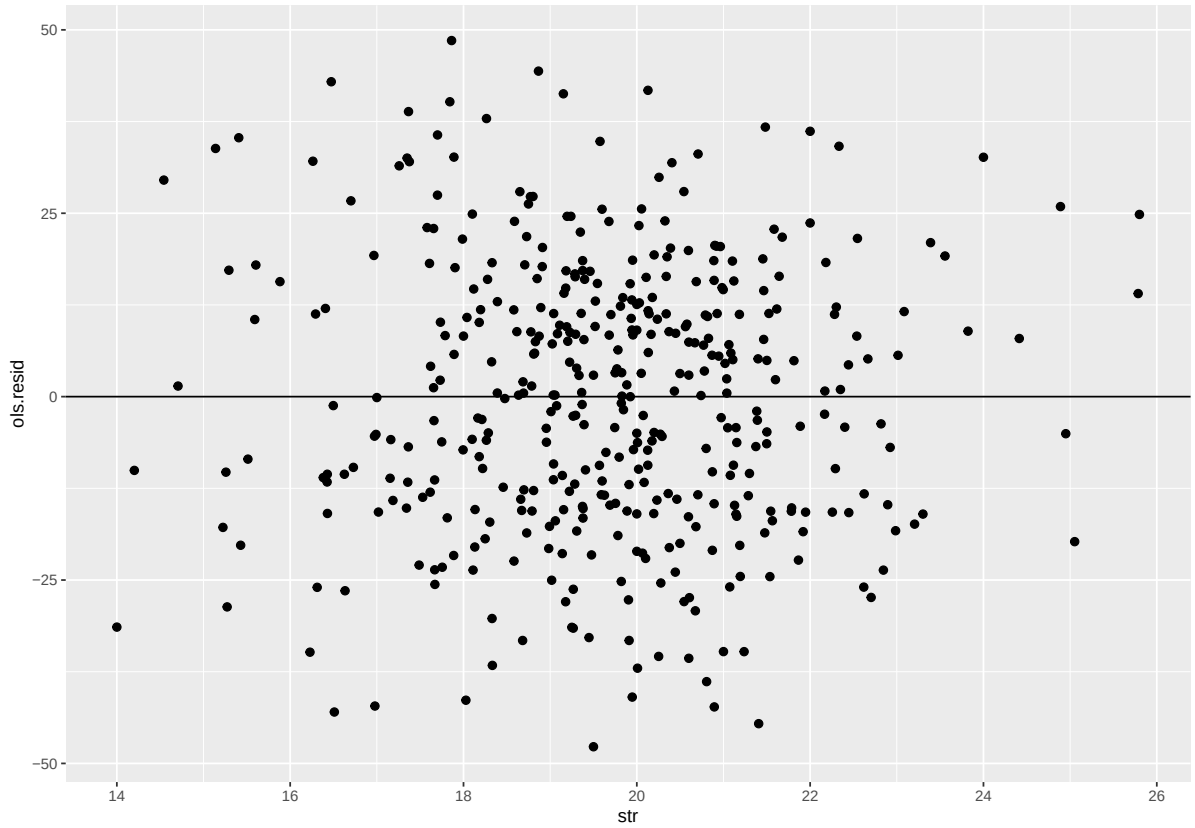
X_i is correlated to u_i through Y_i in equation 2. i.e.

$$cov(X_i, u_i) = \frac{\gamma_1 \sigma^2_{u_i}}{1 - \gamma_1 \beta_1}$$

Inconsistency in the OLS Standard Errors

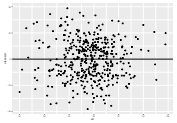
First, we can do a quick visual check with a plot of the residuals of our regression.

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```



Residual Plot Code

```
# Regression and storing residuals
caschool.ols = lm(testscr ~ str, data = caschool)
caschool$ols.resid = caschool.ols$residuals
# plot the residuals
library(ggplot2)
ggplot(aes(y = ols.resid, x = str), data = caschool) +
  geom_point() +
  geom_abline(slope = 0)
```



Breusch-Pagan Test

Second, we can do a test formally by running the regression of the squared residuals on the regressors:

$$var(y_i) = \sigma_i^2 = E(u_i^2) = f(\alpha_1 + \alpha_2 X_{2i} + ... + \alpha_s X_{si})$$

where u_i are the residuals and s is the number of regressors including intercept.

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_s = 0$$

We construct the test statistic by multiplying the R^2 from this regression with N, the sample size. This will have a χ^2_{s-1} distribution. We compare this to a critical value from the chi-distribution at some significance level.

$$\chi^2 = N \times R^2 \sim \chi^2_{s-1}$$

Calulating BP test statistic

```
# Regression
caschool.ols = lm(testscr ~ str, data = caschool)
caschool$ols.resid2 = caschool.ols$residuals^2
# run the regression of residuals squared on the regressor
var_y <- lm(ols.resid2 ~ str, data = caschool)
var_y.r2 = summary(var_y)$r.squared
var_y.r2
```

```
## [1] 0.01379428
```

Multiply the R^2 with $N = 420$, we get 5.79. We need to compare with critical value.

```
# critical value at 5% significance
qchisq(0.95,df=1)
```

```
## [1] 3.841459
```

Since, $5.79 > 3.84$, we reject the null of homoskedasticity assumption at 5% significance level.

BP statistic directly with R code

```r
# Regression
caschool.ols = lm(testscr ~ str, data = caschool)

# BP code
library(lmtest)
bptest(caschool.ols)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  caschool.ols
## BP = 5.7936, df = 1, p-value = 0.01608
```

Replication of Table 9.2

Regressors	(1) CA: model 3	(2) CA: model 4	(3) CA: model 5	(4) MA: model 3	(5) MA: model 4	(6) MA: model 5
Student-Teacher Ratio (STR)	-0.508**	55.62**	-0.342	-0.641**	12.43	-1.018***
	(0.254)	(24.14)	(0.361)	(0.268)	(14.01)	(0.370)
STR^2		-2.915**			-0.680	
		(1.221)			(0.737)	
STR^3		0.0499**			0.012	
		(0.0204)			(0.013)	
% English Learners	-0.205***	-0.196***		-0.437	-0.434	
	(0.034)	(0.035)		(0.303)	(0.300)	
% Eligible for Free Lunch	-0.410***	-0.412***	-0.454***	-0.582***	-0.587***	-0.709***
	(0.033)	(0.033)	(0.029)	(0.097)	(0.104)	(0.091)
Average Income of District	-1.063*	-0.913	-0.597	-3.067	-3.382	-3.867
	(0.568)	(0.576)	(0.598)	(2.353)	(2.491)	(2.488)
$Income^2$	0.0733***	0.0674***	0.0509**	0.164*	0.174*	0.184**
	(0.022)	(0.022)	(0.023)	(0.085)	(0.089)	(0.090)
$Income^3$	-0.000887***	-0.000826***	-0.000615**	-0.00218**	-0.00229**	-0.00234**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
HiEL			6.924			-12.56
			(10.06)			(9.793)
HiEL * STR			-0.612			0.799
			(0.507)			(0.555)
Constant	687.0***	330.1**	682.6***	744.0***	665.5***	759.9***
	(7.245)	(158)	(8.743)	(21.32)	(81.33)	(23.23)
Observations	420	420	420	220	220	220
R^2	0.809	0.812	0.802	0.685	0.687	0.686

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1