ECOM20001
Econometrics 1

Lecture Note 5
Single Linear Regression - Testing

A/Prof Victoria Baranov
Department of Economics
University of Melbourne

Stock and Watson: Chapter 5, Sections 5.1-5.4 and part of 5.5

Summary of Key Concepts

- ► Visual evidence
- ► Hypothesis testing with the regression model
  - ► 3 steps to hypothesis tests
  - ► testing and statistical software
  - ► class size and test score example
- ► Confidence intervals for regression model slope β_1 and intercept β_0
- ► Dummy variables
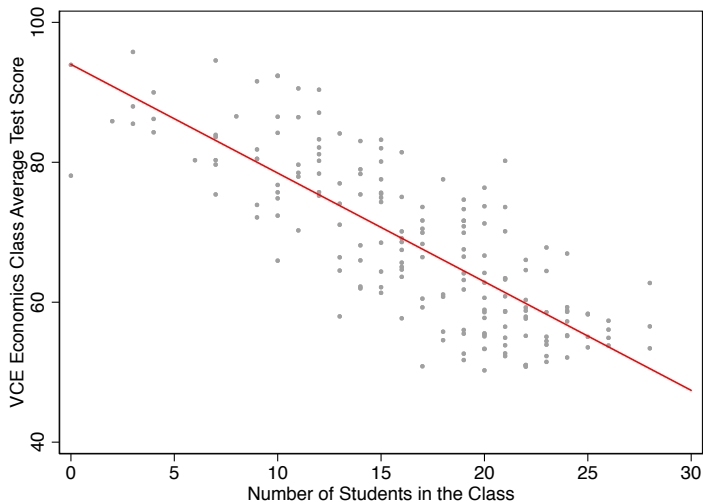- ► Heteroskedasticity and Homoskedasticity

Using Data to Evaluate Claims

▶ Politician says:
  "There's no problem with schools with unequal class sizes! Class size doesn't affect students' test scores!"

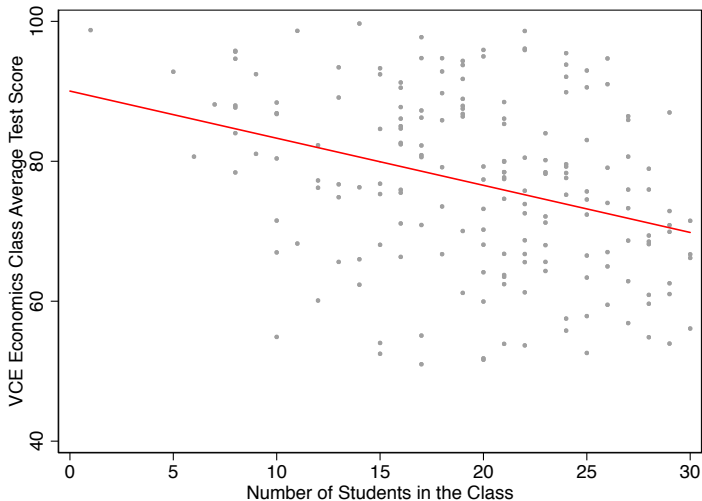▶ How do we use data and regression model to evaluate this claim scientifically?

Insanely compelling visual evidence
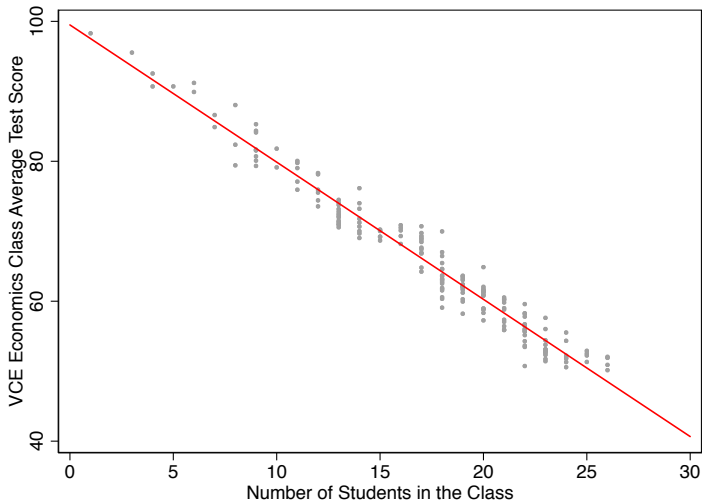
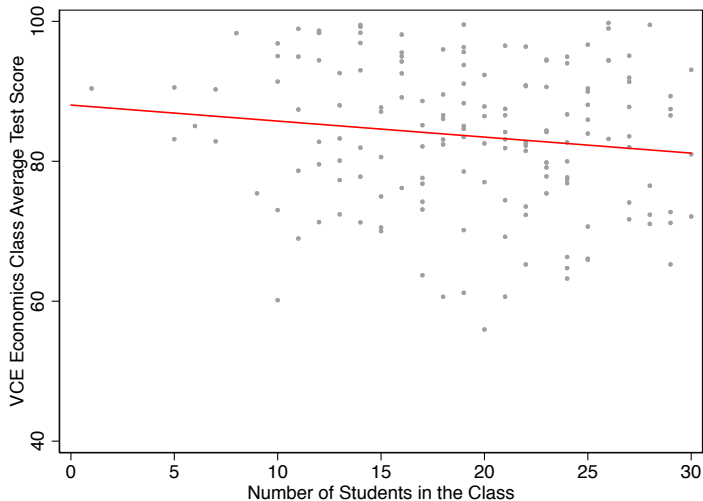# Compelling visual evidence

Less compelling visual evidence

No visual evidence

Hypothesis Testing

▶ While visual evidence is useful, we would like to have a more systematic way of drawing conclusions based on data

▶ Whenever claims are made of the sort that "X affects Y", we can use regressions, hypothesis testing, and confidence intervals to formally evaluate a claim using data

Hypothesis Testing

- Recall our regression model is:

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

  and we estimate it using data (X_i, Y_i)'s $i = 1, 2, \ldots, n$, which is an n observation random sample from the population

- Determining whether "X affects Y" or not boils down to determining whether $\hat{\beta}_1$, is statistically different from 0

- Everything we covered regarding hypothesis testing, statistical significance, and confidence intervals about sample means $\bar{Y}$ carries over directly to testing slopes of regression lines $\hat{\beta}_1$

Hypothesis Testing

▶ More formally, we wish to test the following null hypothesis H_0 against an alternative H_1:

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_{1,0}$$

▶ Notice that we continue to focus strictly on two-sided hypothesis tests: $\beta_1 \neq \beta_{1,0}$ can mean that β_1 is either much smaller or much larger than $\beta_{1,0}$

▶ In practice, we conduct the hypothesis test in 3 steps, like we did with testing hypotheses about the sample mean

3 Steps for Testing Hypotheses About β_1

1. Compute the OLS estimate $\hat{\beta}_1$ and its standard error, $SE(\hat{\beta}_1)$, which has the following formula:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$$

where

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

2. Compute the t-statistic:

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

3a. Compute the p-value

$$\text{p-value} = 2\Phi(-|t^{act}|)$$

where Φ is the cumulative density of the normal distribution.

3 Steps for Testing Hypotheses About β_1

(3a continued) Letting α be the level of significance of the test, we reject the null $H_0 : \beta_1 = \beta_{1,0}$ if

$$\text{p-value} < \alpha$$

where typical values of α are 0.10, 0.05, 0.01

3b. We can equivalently use our t-statistic and critical values from the normal distribution to conduct the hypothesis test.

We reject the null $H_0 : \beta_1 = \beta_{1,0}$ in favour of $H_0 : \beta_1 \neq \beta_{1,0}$ if

$$|t^{act}| > t^{\alpha}_{crit}$$

where recall:

- $t^{\alpha}_{crit} = 1.65$ if $\alpha = 0.10$
- $t^{\alpha}_{crit} = 1.96$ if $\alpha = 0.05$
- $t^{\alpha}_{crit} = 2.58$ if $\alpha = 0.01$

Hypothesis Testing and Statistical Software

▶ In its regression package, R automatically reports the $\hat{\beta}_1$, $SE(\hat{\beta}_1)$, and the t-statistic and p-value for the test of:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

▶ That is, the test of the null hypothesis of no relationship between X and Y versus the alternative hypothesis that there exists a relationship (either positive or negative) between X and Y

▶ This is by far the most popular hypothesis test employed in practice: testing whether a relationship exists or not

▶ So in practice, we use statistics software like R to compute standard errors and do hypothesis testing of this sort

How Regression Output Looks in R

```
Call:
lm(formula = earnings ~ height, data = mydata1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7972 -2.1909 -0.7923  3.4421  5.0579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051174   0.338050  -0.151     0.88
height       0.027859   0.001984  14.042   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.678 on 17868 degrees of freedom
Multiple R-squared:  0.01092,    Adjusted R-squared:  0.01086
F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

How Regression Output Looks in R

Y_i X_i

```
Call:
lm(formula = earnings ~ height, data = mydata1)   Regression

Residuals:
    Min      1Q  Median      3Q     Max
-4.7972 -2.1909 -0.7923  3.4421  5.0579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051174   0.338050  -0.151     0.88
height       0.027859   0.001984  14.042   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.678 on 17868 degrees of freedom
Multiple R-squared:  0.01092,   Adjusted R-squared:  0.01086
F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

How Regression Output Looks in R

```
Call:
lm(formula = earnings ~ height, data = mydata1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7972 -2.1909 -0.7923  3.4421  5.0579

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051174   0.338050  -0.151     0.88
height       0.027859   0.001984  14.042   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.678 on 17868 degrees of freedom
Multiple R-squared:  0.01092,   Adjusted R-squared:  0.01086
F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$

Test Results for Null that Intercept=0

Intercept

How Regression Output Looks in R

```
Call:
lm(formula = earnings ~ height, data = mydata1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7972 -2.1909 -0.7923  3.4421  5.0579
```

Test Results for Null that Slope=0

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051174   0.338050  -0.151     0.88
height       0.027859   0.001984  14.042   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.678 on 17868 degrees of freedom
Multiple R-squared:  0.01092,   Adjusted R-squared:  0.01086
F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_1$

Slope

How Regression Output Looks in R

```
Call:
lm(formula = earnings ~ height, data = mydata1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7972 -2.1909 -0.7923  3.4421  5.0579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051174   0.338050  -0.151     0.88
height       0.027859   0.001984  14.042   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.678 on 17868 degrees of freedom
Multiple R-squared:  0.01092,   Adjusted R-squared:  0.01086
F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

Regression Fit: R-Squared and SER

Additional Comments on Hypothesis Testing

▶ Remember, fundamental to conducting hypothesis testing in this way is having a large enough n so we can apply the LLN and CLT to compute p-values and critical values for a given significance level α

▶ We can also conduct hypothesis tests for the intercept of the regression model of the form:

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{vs.} \quad H_1 : \beta_0 \neq \beta_{0,0}$$

where R similar reports standard errors, t-statistics, and p-values for the OLS estimate $\hat{\beta}_0$

▶ The testing procedure is identical as it is for testing β_1

One-sided Alternatives

▶ We can conduct one-sided hypothesis tests for the slope (and intercept) coefficients exactly as we did with sample means

▶ For the following test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 > 0$$

we would compute p-value$=1 - \Phi(t^{act})$

▶ For the following test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 < 0$$
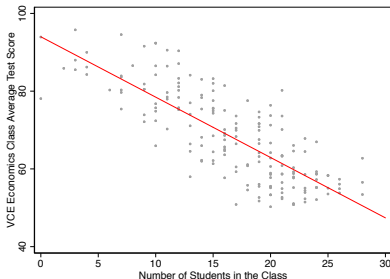
we would compute p-value$=\Phi(t^{act})$

▶ One-sided tests are rarely used with regression because we want to the data to tell us the sign of the slope of the regression line, not make an assumption about from the outset

Hypothesis Testing with Class Size and Test Score Example

Original Class Size and Test Score Dataset

- ▶ Original scatter plot of class size and test scores:

1. We can report the corresponding regression results as follows:

$$\widehat{TestScore}_i = \underset{(1.55)}{93.99} - \underset{(0.09)}{1.55}\, ClassSize_i, \quad R^2 = 0.63, SER = 7.24$$

where the numbers in parantheses () are $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$

Hypothesis Testing with Class Size and Test Score Example

Original Class Size and Test Score Dataset

▶ With our results in hand, we are ready to conduct steps 2 and 3 of our hypothesis test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

2. The t-statistic for the test is computed as:

$$t^{act} = \frac{-1.55 - 0}{0.09} = -17.22$$

3a. The corresponding p-value for the test statistic is:

$$\text{p-value} = 2\Phi(-|-17.22|) = 2\Phi(-17.22) = 0.00001$$

We reject the null as our p-value is less than our level of significance: $0.00001 < 0.05 = \alpha$

- ▶ there is a $0.00001 \times 100 = 0.001\%$ chance that we obtain $\hat{\beta}_1 = -1.55$ or smaller if the null $H_0 : \beta_1 = 0$ is true
- ▶ extremely unlikely that $\beta_1 = 0$ given our estimate $\hat{\beta}_1$

Hypothesis Testing with Class Size and Test Score Example

Original Class Size and Test Score Dataset

3b. (Alternative test using t^{act} and critical values)
Using $\alpha = 0.05$, we compare t^{act} to the critical value of $t_{crit}^{\alpha=0.05} = 1.96$ and find:

$$|t^{act}| = 17.22 > 1.96$$

which leads us to reject the null hypothesis of $H_0 : \beta_1 = 0$ in favour of the the alternative $H_1 : \beta_1 \neq 0$.

▶ The evidence (or hypothesis tests) does not support the politician's claim of no test score – class size relationship!

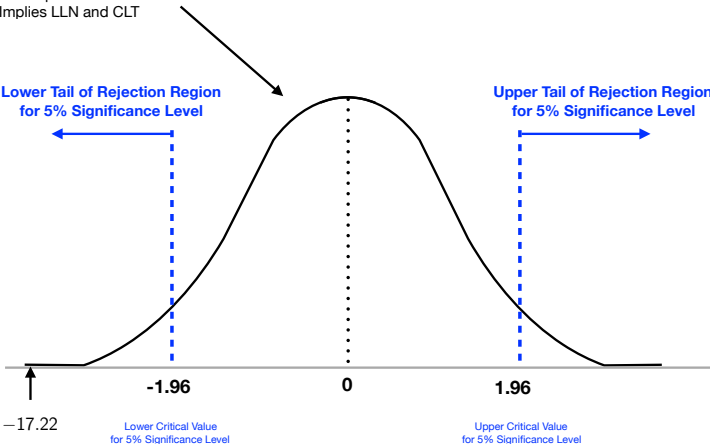# Hypothesis Testing with Class Size and Test Score Example

Class Size and Test Score Dataset

Distribution of t-statistics
- N(0,1) if sample size is large enough and 3 Least Squares Assumptions hold
- Implies LLN and CLT

Lower Tail of Rejection Region for 5% Significance Level

Upper Tail of Rejection Region for 5% Significance Level
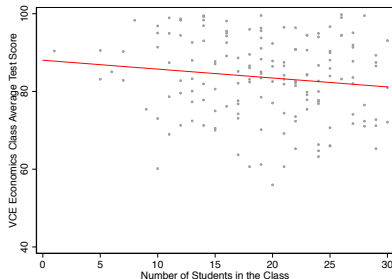
-1.96

0

1.96

$t^{act} = -17.22$

(t-statistic for H0 vs H1 from the regression in the lower tail of the rejection region)

Lower Critical Value for 5% Significance Level

Upper Critical Value for 5% Significance Level

Hypothesis Testing with Class Size and Test Score Example

▶ Consider the "no visual evidence" scatter plot of class size and test scores:

1. We can report the corresponding regression results as follows:

$$\widehat{TestScore}_i = 88.04 - 0.23 \; ClassSize_i, \quad R^2 = 0.018, SER = 10.34$$
$$\;\;\;\;\;\;\;\;\;\;(2.71) \;\;\;\; (0.133)$$

where the numbers in parantheses are $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$

Hypothesis Testing with Class Size and Test Score Example

Noisier Class Size and Test Score Dataset

▶ With our results in hand, we are ready to conduct steps 2 and 3 of our hypothesis test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

2. The t-statistic for the test is computed as:

$$t^{act} = \frac{-0.23 - 0}{0.13} = -1.77$$

3b. The corresponding p-value for the test statistic is:

$$\text{p-value} = 2\Phi(-|-1.77|) = 2\Phi(-1.77) = 0.089$$

We fail to reject the null as our p-value is greater than our level of significance: $0.089 > 0.05 = \alpha$

▶ there is a decent (8.9%) chance of obtaining a value $\hat{\beta}_1 = -0.22$ if the null $\beta_1 = 0$ is true

Hypothesis Testing with Class Size and Test Score Example

Noisier Class Size and Test Score Dataset

3b. (Alternative test using t^{act} and critical values)
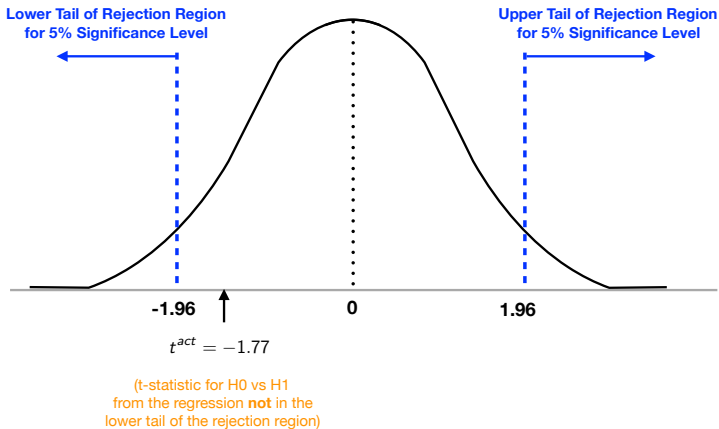Using $\alpha = 0.05$, we compare t^{act} to the critical value of $t_{crit}^{\alpha=0.05} = 1.96$ and find:

$$|t^{act}| = 1.77 < 1.96$$

which leads us to fail to reject the null hypothesis of $H_0 : \beta_1 = 0$ in favour of the the alternative $H_1 : \beta_1 \neq 0$.

▶ The evidence in this example is too noisy, and it <u>does</u> support the politician's claim of no test score – class size relationship!

Hypothesis Testing with Class Size and Test Score Example

Lower Tail of Rejection Region for 5% Significance Level

Upper Tail of Rejection Region for 5% Significance Level

-1.96   0   1.96

$t^{act} = -1.77$

(t-statistic for H0 vs H1 from the regression **not** in the lower tail of the rejection region)

Hypothesis Testing with Class Size and Test Score Example

Testing Other Types of Null Hypotheses

▶ Suppose we wanted to run the following hypothesis test with our original dataset:

$$H_0 : \beta_1 = -1 \quad \text{vs.} \quad H_1 : \beta_1 \neq -1$$

1. Obtain regression results:

$$\widehat{TestScore}_i = 93.99 - \underset{(1.55)}{1.55} ClassSize_i, \quad R^2 = 0.63, SER = 7.24$$
$$\underset{(1.55)}{} \underset{(0.09)}{}$$

2. The t-statistic for the test is computed as:

$$t^{act} = \frac{-1.55 - (-1)}{0.09} = -6.11$$

3b. Using $\alpha = 0.05$, we compare t^{act} to the critical value of $t^{\alpha=0.05}_{crit} = 1.96$ and find:

$$|t^{act}| = 6.11 > 1.96$$

which leads us to reject the null hypothesis of $H_0 : \beta_1 = -1$ in favour of the the alternative $H_1 : \beta_1 \neq -1$.

Confidence Intervals

- We can compute confidence intervals (CI) for β_1 in the exact same way that we computed CIs for the population mean μ_Y in Lecture Note 3

- Remember, the basic idea with CIs: with a random sample, we can never figure out exactly what the true value of β_1 is

- But we can use our OLS estimate $\hat{\beta}_1$, and its sampling distribution $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, to determine a range in which the true value of β_1 is, with confidence $1 - \alpha$

- Also remember that if we choose a level of statistical significance α for hypothesis testing, the corresponding level of confidence for constructing CIs is $1 - \alpha$

- Throughout we will focus on CIs for β_1; CIs for β_0 are computed the exact same way

Confidence Intervals

- Let's work with $\alpha = 0.05$ so we have a $1 - 0.05 = 0.95$ or 95% CI
- CIs have two equivalent definitions:
    - <u>Definition 1</u>: the 95% CI is the set of values for the null hypothesis that <u>cannot</u> be rejected using a two-sided hypothesis test with a 5% significant level
    - <u>Definition 2</u>: the 95% CI is an interval that has a 95% probability of containing the true value of β_1 from the population

Confidence Intervals
Constructing a 95% CI from the t-statistic

- What are the null β_1 values that are <u>not</u> rejected at the 5% level of significance?
  - recall that $t_\alpha = 1.96$ is the critical value for the test corresponding to $\alpha = 5\%$
- Using our rejection rule for a 2-sided hypothesis test. We <u>cannot reject</u> all null values of β_1 if

$$-1.96 < t^{act} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < 1.96$$

- Re-arranging the inequalities, this is equivalent to saying that we cannot reject all null values of β_1 if

$$\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)$$

Confidence Intervals
Constructing a 95% CI from the t-statistic

▶ Hence the 95% CI for β_1 is given by:

$$[\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)]$$

▶ Interpretations:
  ▶ given our data and our OLS estimates of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$, any null hypothesis value for β_1 that falls between $\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1)$ and $\hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)$ would <u>not</u> be rejected at the 5% level of significance
  ▶ given our data and our OLS estimates of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$, there is a 95% chance that the true value of β_1 lies between $\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1)$ and $\hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)$

CIs Application to Class Size and Test Scores

▶ OLS regression results:

$$\widehat{TestScore}_i = \underset{(1.55)}{93.99} - \underset{(0.09)}{1.55}\, ClassSize_i, \quad R^2 = 0.63, SER = 50.46$$

▶ 95% confidence interval is:

$$[-1.55 - 1.96 \times 0.09, -1.55 + 1.96 \times 0.09]$$

which is

$$[-1.73, -1.37]$$

▶ Given our data and OLS estimates, there is a 95% chance that the true value of β_1 lies in the range $[-1.73, -1.37]$
  ▶ with 95% confidence, the true value of β_1 is as low as -1.73 (one additional student reduces test scores by 1.73 out of 100)
  ▶ with 95% confidence, the true value of β_1 is as high as -1.37 (one additional student reduces test scores by 1.37 out of 100)

Class Size and Test Scores - Other Common CIs

- 90% confidence interval is:

  $$[-1.55 - 1.65 \times 0.09, -1.55 + 1.65 \times 0.09] = [-1.70, -1.40]$$

- 95% confidence interval is:

  $$[-1.55 - 1.96 \times 0.09, -1.55 + 1.96 \times 0.09] = [-1.73, -1.37]$$

- 99% confidence interval is:

  $$[-1.55 - 2.58 \times 0.09, -1.55 + 2.58 \times 0.09] = [-1.78, -1.32]$$

- Again, notice how how the confidence interval gets <u>wider</u> as the <u>confidence level goes up</u>
  - You need a wider interval of potential values of β_1 to be more confident that true value of β_1 lies in the interval

CIs for Predicting Effects of Changing X

▶ Our general 95% CI formulas

$$[\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)]$$

are intervals for the average change in Y corresponding to a one-unit change in X of $\Delta X = 1$

▶ From our example, they correspond to the change in Y = test scores from adding $\Delta X = 1$ student to a class

▶ The general 95% CI formula for the change in Y for any ΔX value is given by

$$[(\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1)) \times \Delta X, (\hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)) \times \Delta X]$$

▶ That is, you multiply the upper and lower limits of the usual CI formula for $\Delta X = 1$ by ΔX to get the CI for predicted average change in Y corresponding to a given value of ΔX

CIs for Predicting Effects of Changing *X*

▶ From our example, the 95% CI for the change in test scores from changing $\Delta X = 2$ (adding 2 additional students to a classroom) is:

$$[(-1.55 - 1.96 \times 0.09) \times 2, (-1.55 + 1.96 \times 0.09) \times 2]$$

which equals

$$[-3.46, -2.74]$$

▶ <u>Interpretation</u>: Given our data and OLS estimates, there is a 95% chance the true impact on average test scores from adding 2 students to a classroom lies on the interval $[-3.46, -2.74]$

Dummy Variables

- So far we have focused on a regressor X that is continuous
- Regressions can also incorporate regressors that are binary variables, that is variables that take on the value 0 or 1
- Also called dummy variables and indicator variables, and are denoted by D_i where either $D_i = 0$ or $D_i = 1$
- Examples of binary variables abound:
  - female $D_i = 1$, male $D_i = 0$
  - urban $D_i = 1$, rural $D_i = 0$
  - domestic $D_i = 1$, foreign $D_i = 0$

Dummy Variables

▶ Regression model with D_i as the regressor:

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

▶ Same regression model as we have been working with, except we have our dummy variable D_i in where X_i was before

▶ The β_1 coefficient has a very different interpretation, however: it is <u>not</u> a slope coefficient anymore

▶ We instead call β_1 the coefficient on D_i

Dummy Variables - How do They Work?

▶ Regression model with D_i as the regressor:

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

▶ Consider the model when $D_i = 0$ and when $D_i = 1$, that is, when the dummy is "switched off" and "switched on":

$$Y_i = \beta_0 + u_i \quad (D_i = 0)$$

and

$$Y_i = \beta_0 + \beta_1 + u_i \quad (D_i = 1)$$

Dummy Variables - How do They Work?

▶ In terms of conditional expectations, we have:

$$E[Y_i|D_i = 0] = E[\beta_0 + \beta_1 \underbrace{D_i}_{0} + u_i|D_i] = \beta_0$$

and

$$E[Y_i|D_i = 1] = E[\beta_0 + \beta_1 \underbrace{D_i}_{1} + u_i|D_i] = \beta_0 + \beta_1$$

▶ In other words, the population mean is β_0 for the group where $D_i = 0$, and is $\beta_0 + \beta_1$ when $D_i = 1$

▶ Therefore, β_1 is the interpreted as the difference in the population mean between groups for which $D_i = 0$ and $D_i = 1$

Dummy Variables Application to Class Size and Test Scores

THE AGE

NATIONAL VICTORIA

Country students falling behind at school

By Jewel Topsfield Education Editor
27 August 2014 – 3:44pm

▶ Politician says: "That is fake news! There is no gap between the academic performance of urban and regional students"

▶ Let's use our class size, test scores, and (new) urban/regional status dataset to rigorously shed some light on this debate

▶ The dummy variable of interest is $Urban_i$
  ▶ $Urban_i = 1$ if class i is in a location with $> 100,000$ people
 ▶ $Urban_i = 0$ if class i is in a location with $< 100,000$ people

Dummy Variables Application to Class Size and Test Scores

Dummy Variables Data

classid	class_size	grade	urban
1	12	75.23	1
2	28	53.41	0
3	23	67.82	1
4	19	73.31	0
5	10	65.94	1
6	24	55.27	1
7	7	80.31	1
8	24	52.10	0
9	17	71.68	1
10	14	61.98	1
11	24	59.29	0
12	10	74.85	1
13	10	92.38	1
14	16	67.48	0
15	17	68.33	1
16	16	70.13	0
17	12	82.12	1
18	12	83.29	0
19	15	82.01	1
20	16	64.67	0
21	19	64.13	1
22	19	69.93	1
23	15	75.65	0
24	24	55.17	1
25	13	64.53	1

Dummy Variables Application to Class Size and Test Scores

Dummy Variables Graphically
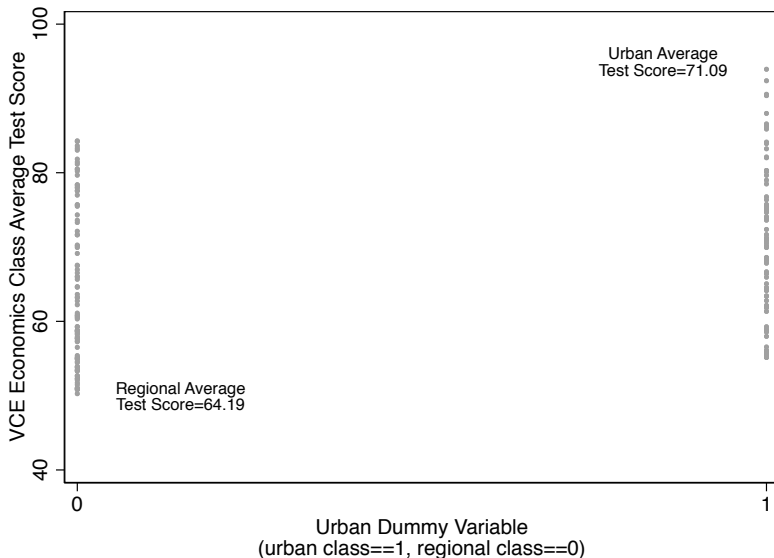
Dummy Variables Application to Class Size and Test Scores
Dummy Variables Regression

▶ We estimate OLS regressions, conduct hypothesis tests, and construct confidence intervals with dummy variables D_i *exactly* as we do with continuous regressors X_i

▶ Regression results based on the urban/rural dummy variable:

$$\widehat{TestScore_i} = \underset{(1.06)}{64.19} + \underset{(1.58)}{6.90}\, Urban_i, \quad R^2 = 0.10, SER = 10.37$$

▶ Hypothesis testing: the 6.90 coefficient on $Urban_i$ has:
  ▶ t-statistic of $t^{act} = 4.38$
  ▶ p-value of 0.000001
  ▶ 95% CI of [3.80,9.99]
  ▶ Strongly reject the null at the $\alpha = 0.05$ level of significance

▶ Evidence here is very worrisome for policy: students in urban markets tend to have 6.90 more points out of 100 on their test scores

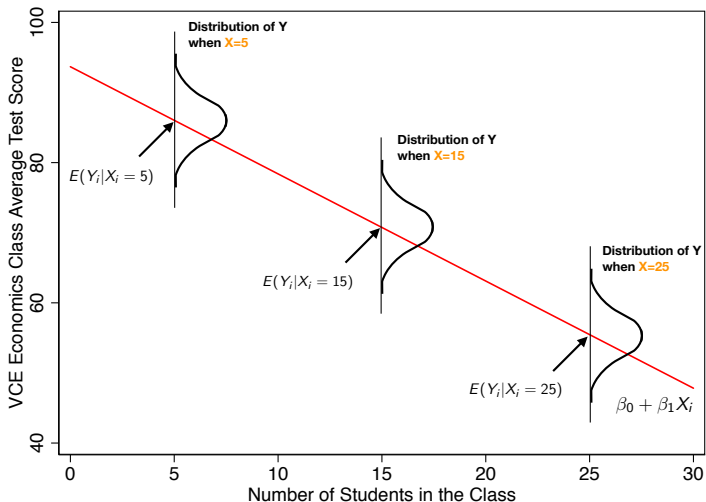# Heteroskedasticity and Homoskedasticity

▶ Let's return to our standard regression model

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

▶ Recall that throughout we have maintained 3 Least Squares Assumptions in estimating and testing our regression model:

1. Independence of u_i and X_i
2. IID Sampling of (X_i, Y_i)
3. No Outliers

▶ We can also characterise the variance of the errors u_i given X_i, $\sigma^2_{u|X}$. We say the errors are . . .

  ▶ homoskedastic if $\sigma^2_{u_i|X_i}$ is <u>the same</u> no matter what X_i is
  ▶ heteroskedastic if $\sigma^2_{u_i|X_i}$ <u>varies</u> with X_i

Homoskedasticity Graphically

Heteroskedasticity Graphically
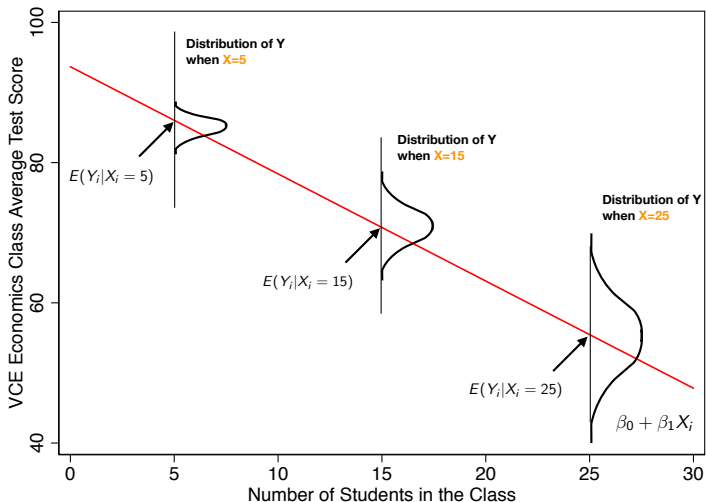
Heteroskedasticity vs Homoskedasticity Example

- A good example from the text that highlights heteroskedasicity is the relation between earnings and gender

- Suppose we had a random sample on $Earnings_i$ for males and females, with a dummy variable $Male_i$ that equals 1 if person i is male and 0 otherwise

- We could run the following regression:

$$Earnings_i = \beta_0 + \beta_1 Male_i + u_i$$

where β_1 is the difference in mean earnings between males and females

- We could also assume homoscedastic errors, then u_i does not depend the regressor, which in this example is $Male_i$.

- This is equivalent to assuming that the variance of earnings is the same for men as it is for women

- <u>Problem</u>: we know that this is just plain wrong! Males tend to have more dispersed earnings than females.

Heteroskedasticity vs Homoskedasticity: Who Cares?

- The variance of u_i conditional on X_i underlies the variance of the residuals $\hat{u}_i$ from an OLS regression
- The variance of the residuals in turn is critical for the calculation of the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$
- Finally, $SE(\hat{\beta}_1)$ directly enters into our calculations of t-statistics and confidence intervals
- So ultimately the variance of u_i is a critical component to hypothesis testing
- <u>Bottom line</u>: if we get the variance of u_i conditional on X_i wrong, our hypothesis tests are wrong!
- <u>Note</u>: regardless of whether we assume heteroskedasticity vs homoskedasticity, OLS is unbiased ($E(\hat{\beta}_1) = \beta_1$)

A Key Result Under Homoskedasticity

- ▶ Homoskedasticity underlies an important theoretical property of OLS estimators:
- ▶ Gauss-Markov Theorem: Under the 3 Least Squares assumptions, and if u_i is homoskedastic, then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is the Best Linear Unbiased Estimator of the linear regression true parameters β_0 and β_1
- ▶ That is, the OLS is BLUE

A Key Result Under Homoskedasticity

- ▶ Let's break down what "OLS is BLUE" means
- ▶ Best: the OLS estimator is the most efficient estimator (i.e., smallest sampling variance, so smallest standard errors) among all the estimators that are linear in $Y_1, \ldots, Y_n$.
- ▶ Linear: OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of $Y_1, \ldots, Y_n$, which they are:
    - ▶ $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})}$
    - ▶ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- ▶ Unbiased: expected values of the OLS estimators are equal to their true values in the population:

$$E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0$$

- ▶ Estimator: OLS values $\hat{\beta}_1$ and $\hat{\beta}_1$ are estimators of population true values β_1 and β_0

BUT.....

- ▶ While OLS is BLUE is a key result in theory, it is virtually never relevant in practice

- ▶ In general, the variance of u_i does vary with X_i, like the earnings and male/female example highlighted

- ▶ With continuous variables like income or age, the variance of u_i almost always varies with X_i

- ▶ Because heteroskedasticity is a prominent practical phenomenon in data, we always do econometrics assuming heteroskedasticity in the errors

- ▶ Therefore, in conducting hypothesis tests and constructing confidence intervals, we compute $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ using heteroskedasticity-robust standard errors, which are often called robust standard errors or White standard errors
  - ▶ in honour of the late Prof. Halbert White who would have won the Nobel Prize before he passed away in 2012

Heteroskedastic Standard Errors

▶ The complex formulas for heteroskedasticity robust standard errors are in the textbook

▶ In practice statistical programs like R readily compute robust standard errors for you

▶ In practice, if you incorrectly assume homoskedasticity, you will compute incorrect $SE(\hat{\hat{\beta}}_0)$ and $SE(\hat{\hat{\beta}}_1)$!

Heteroskedastic Standard Errors

- ▶ In contrast, if you use robust standard errors when the errors are in fact homoskedastic, you will still obtain correct standard errors by using robust standard errors → it's the safe choice!

- ▶ Therefore, for the remainder of the subject, we will assume heteroskedasticity robust standard errors in conducting all hypothesis tests and in computing confidence intervals (and will use R to compute them)

Note: in the text, we stop at the end of Section 5.4; ignore Section 5.5 (Theoretical Foundations of OLS) and Section 5.6 (t-Statistics in Regression When Sample Size is Small)