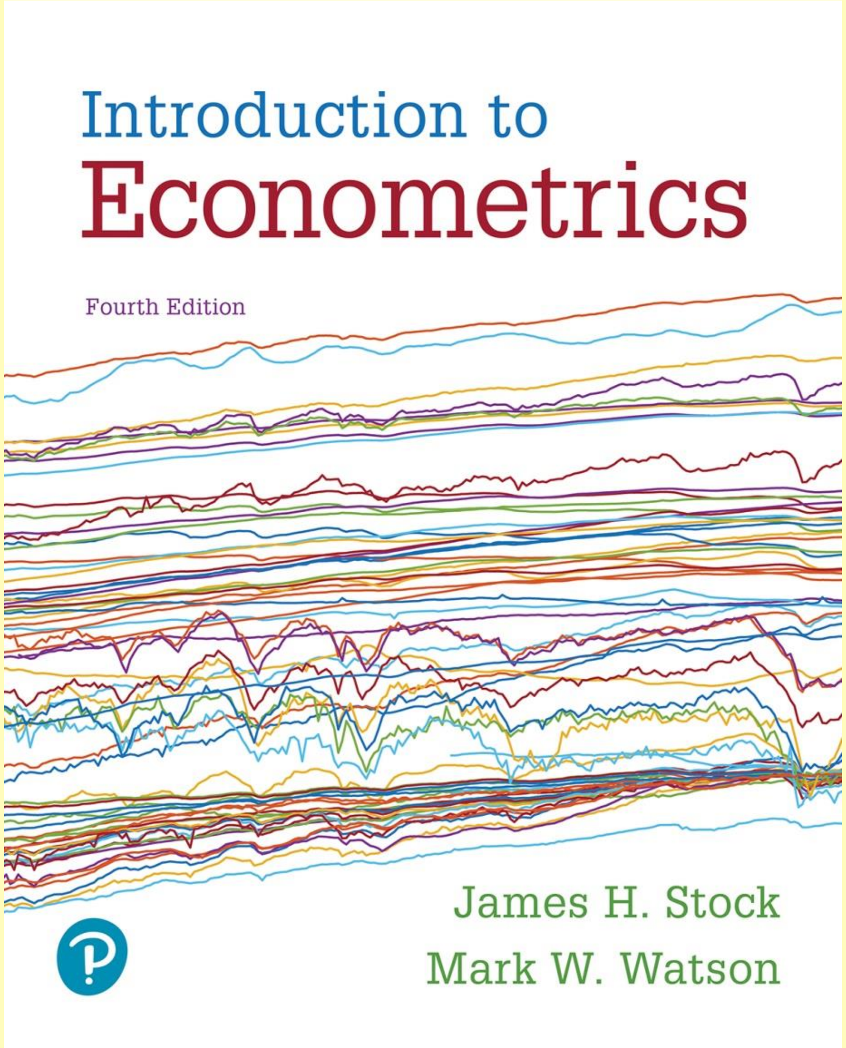
Introduction to Econometrics

Fourth Edition



Chapter 5

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

Outline

1. The standard error of $\hat{\beta}_1$
2. Hypothesis tests concerning β_1
3. Confidence intervals for β_1
4. Regression when X is binary
5. Heteroskedasticity and homoskedasticity
6. Efficiency of OLS and the Student t distribution

A big picture review of where we are going

We want to learn about the slope of the population regression line. We have data from a sample, so there is sampling uncertainty. There are five steps towards this goal:

1. State the population object of interest
2. Provide an estimator of this population object
3. Derive the sampling distribution of the estimator (this requires certain assumptions). In large samples this sampling distribution will be normal by the CLT.
4. The square root of the estimated variance of the sampling distribution is the standard error (SE) of the estimator
5. Use the *SE* to construct *t*-statistics (for hypothesis tests) and confidence intervals.

Object of interest: β_1 (1 of 2)

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots, n$$

β_1 = slope of population regression line

Estimator: the OLS estimator $\hat{\beta}_1$.

The Sampling Distribution of $\hat{\beta}_1$:

Because the population regression line is $E(Y|X) = \beta_0 + \beta_1 X$, $E(u_i|X_i) = 0$.

To derive the large-sample distribution of $\hat{\beta}_1$ assume:

- (X_i, Y_i), $i = 1,\ldots,n$ are i.i.d.
- Large outliers in X and/or Y are rare (X and Y have four moments)

These are the second and third least squares assumptions.

Object of interest: β_1 (2 of 2)

The Sampling Distribution of $\hat{\beta}_1$:

For n large, $\hat{\beta}_1$ is approximately distributed,

$$\boxed{\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ where } v_i = (X_i - \mu_X)u_i}$$

Hypothesis Testing and the Standard Error of $\hat{\beta}_1$ (Section 5.1)

The objective is to test a hypothesis, like $\beta_1 = 0$, using data – to reach a tentative conclusion whether the (null) hypothesis is correct or incorrect.

General setup

Null hypothesis and **two-sided** alternative:

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

where $\beta_{1,0}$ is the hypothesized value under the null.

Null hypothesis and **one-sided** alternative:

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

General approach: construct *t*-statistic, and compute *p*-value (or compare to the *N*(0,1) critical value)

- ***In general***:

$$\boxed{t = \frac{\text{estimator - hypothesized value}}{\text{standard error of the estimator}}}$$

where the *SE* of the estimator is the square root of an estimator of the variance of the estimator.

- ***For testing the mean of Y***:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{S_Y/\sqrt{n}}$$

- ***For testing*** $\boldsymbol{\beta_1}$,

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

where $SE(\hat{\beta}_1) =$ the square root of an estimator of the variance of the sampling distribution of $\hat{\beta}_1$

RSS is the residual sum of squares, $\hat{\sigma}^2 = \frac{RSS}{n-2}$ is an unbiased estimator of σ^2.

Also $\frac{RSS}{\sigma^2}$ has a χ^2–distribution with degree of freedom (n-2).

Further the distribution of RSS is independent of the distribution of $\hat{\alpha}$ and $\hat{\beta}$.

Formula for $SE(\hat{\beta}_1)$ (1 of 2)

Recall the expression for the variance of (large *n*):

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n(\sigma_X^2)^2}, \text{ where } v_i = (X_i - \mu_X)u_i.$$

The estimator of the variance of $\hat{\beta}_1$ replaces the unknown population values of σ_v^2 and σ_X^2 by estimators constructed from the data:

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

where $\hat{v}_i = (X_i - \bar{X})\hat{u}_i$.

Formula for $SE(\hat{\beta}_1)$ (2 of 2)

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

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \text{the standard error of } \hat{\beta}_1$$

This is a bit nasty, but:

- It is less complicated than it seems. The numerator estimates var(v), the denominator estimates $[\text{var}(X)]^2$.
- Why the degrees-of-freedom adjustment n – 2? Because two coefficients have been estimated (β_0 and β_1).
- $SE(\hat{\beta}_1)$ is computed by regression software
- Your regression software has memorized this formula so you don't need to.

Summary:
To test $H_0: \beta_1 = \beta_{1,0}$ v. $H_1: \beta_1 \neq \beta_{1,0}$

- Construct the *t*-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$

- Reject at 5% significance level if $|t| > 1.96$
- The *p*-value is $p = \Pr[|t| > |t^{act}|] =$ probability in tails of normal outside $|t^{act}|$; you reject at the 5% significance level if the *p*-value is < 5%.

- This procedure relies on the large-*n* approximation that $\hat{\beta}_1$ is normally distributed; typically $n = 50$ is large enough for the approximation to be excellent.

Example: *Test Scores* and *STR*, California data (1 of 2)

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

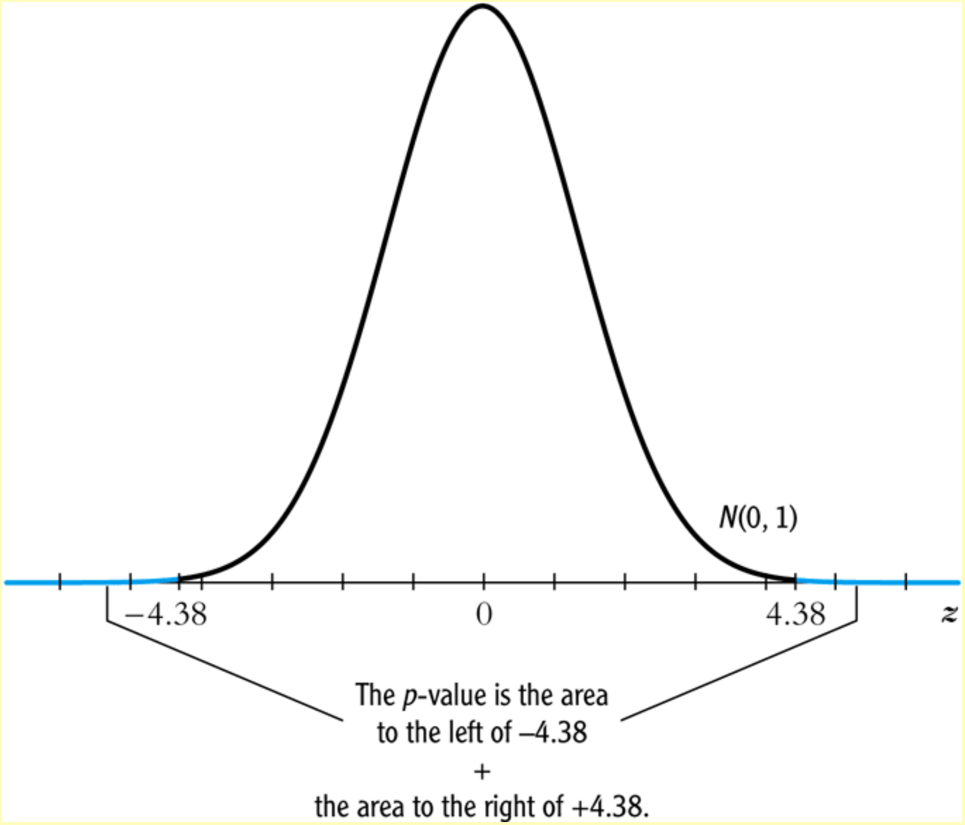
Regression software reports the standard errors:

$$SE(\hat{\beta}_0) = 10.4 \qquad SE(\hat{\beta}_1) = 0.52$$

$$t\text{-statistic testing } \beta_{1,0} = 0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.52} = -4.38$$

- The 1% 2-sided significance level is 2.58, so we reject the null at the 1% significance level.
- Alternatively, we can compute the *p*-value…

Example: *Test Scores* and *STR*, California data (2 of 2)



The p-value based on the large-n standard normal approximation to the t-statistic is 0.00001 (10^{-5})

Confidence Intervals for β_1 (Section 5.2)

Recall that a 95% confidence is, equivalently:

- The set of points that cannot be rejected at the 5% significance level;
- A set-valued function of the data (an interval that is a function of the data) that contains the true parameter value 95% of the time in repeated samples.

Because the *t*-statistic for β_1 is N(0,1) in large samples, construction of a 95% confidence for β_1 is just like the case of the sample mean:

$$\text{95\% confidence interval for } \beta_1 = \{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$$

Confidence interval example: Test Scores and *STR*

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

$$SE(\hat{\beta}_0) = 10.4 \qquad\qquad SE(\hat{\beta}_1) = 0.52$$

95% confidence interval for $\hat{\beta}_1$:

$$\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\} = \{-2.28 \pm 1.96 \times 0.52\}$$
$$= (-3.30,\ -1.26)$$

The following two statements are equivalent *(why?)*

- The 95% confidence interval does not include zero;
- The hypothesis $\beta_1 = 0$ is rejected at the 5% level

A concise (and conventional) way to report regressions: Put standard errors in parentheses below the estimated coefficients to which they apply

$$\widehat{TestScore} = 698.9 - 2.28 \times STR,\ R^2 = .05,\ SER = 18.6$$
$$\qquad (10.4) \quad (0.52)$$

<u>This expression gives a lot of information</u>

- The estimated regression line is "$Testscore = 698.9 - 2.28 \times STR$"
- The standard error of $\hat{\beta}_0$ is 10.4.
- The standard error of $\hat{\beta}_1$ is 0.52.
- The R^2 is .05; the standard error of the regression is 18.6

OLS regression: reading STATA output

```
regress testscr str, robust
Regression with robust standard errors                 Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581
   -------------------------------------------------------------------------
           |               Robust
   testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
   --------+----------------------------------------------------------------
       str |  -2.279808   .5194892    -4.38   0.000    -3.300945   -1.258671
     _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
   -------------------------------------------------------------------------
```

so:

$$\widehat{TestScore} = 698.9 - 2.28\times STR, \; , \; R^2 = .05, \; SER = 18.6$$

$$(10.4) \quad (0.52)$$

$t(\beta_1 = 0) = -4.38$, p-value $= 0.000$ (2-sided)

95% 2-sided conf. interval for β_1 is $(-3.30, \; -1.26)$

Summary of statistical inference about β_0 and β_1

Estimation:

- OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have approximately normal sampling distributions in large samples

Testing:

- H_0: $\beta_1 = \beta_{1,0}$ v. $\beta_1 \neq \beta_{1,0}$ ($\beta_{1,0}$ is the value of β_1 under H_0)
- $t = (\hat{\beta}_1 - \hat{\beta}_{1,0})/SE(\hat{\beta}_1)$
- p-value = area under standard normal outside t^{act} (large n)

Confidence Intervals:

- This is the set of β_1 that is not rejected at the 5% level
- The 95% CI contains the true β_1 in 95% of all samples.
- 95% confidence interval for β_1 is $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$

Regression when X is Binary (Section 5.3)

Sometimes a regressor is binary:

- $X = 1$ if small class size, $= 0$ if not
- $X = 1$ if female, $= 0$ if male
- $X = 1$ if treated (experimental drug), $= 0$ if not

Binary regressors are sometimes called "dummy" variables.

So far, β_1 has been called a "slope," but that doesn't make sense if X is binary.

How do we interpret regression with a binary regressor?

Interpreting regressions with a binary regressor $Y_i = \beta_0 + \beta_1 X_i + u_i$, where X is binary (X_i = 0 or 1) (1 of 2)

When $X_i = 0$, $Y_i = \beta_0 + u_i$

- the mean of Y_i is β_0
- that is, $E(Y_i | X_i = 0) = \beta_0$

When $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$

- the mean of Y_i is $\beta_0 + \beta_1$
- that is, $E(Y_i | X_i = 1) = \beta_0 + \beta_1$

so:

$$\beta_1 = E(Y_i | X_i = 1) - E(Y_i | X_i = 0)$$

= population difference in group means

Interpreting regressions with a binary regressor $Y_i = \beta_0 + \beta_1 X_i + u_i$, where X is binary (X_i = 0 or 1) (2 of 2)

$$\textbf{\textit{Example}}: \quad \text{Let} \quad D_i = \begin{cases} 1 \text{ if } STR_i \le 20 \\ 0 \text{ if } STR_i > 20 \end{cases}$$

$$\textbf{\textit{OLS regression}}: \quad \widehat{TestScore} = \mathbf{650.0} + \mathbf{7.4} \times D$$

$$(1.3) \quad (\mathbf{1.8})$$

Tabulation of group means:

Class Size	Average score ($\bar{Y}$)	Std. dev. (s_Y)	*N*
Small (*STR* > 20)	657.4	19.4	238
Large (*STR* ≥ 20)	**650.0**	17.9	182

$$\textbf{Difference in means}: \quad \bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = 657.4 - 650.0 = \mathbf{7.4}$$

$$\textbf{Standard error} \quad SE = \sqrt{\frac{S_s^2}{n_s} + \frac{S_l^2}{n_1}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = \mathbf{1.8}$$

Summary: regression when X_i is binary (0/1)

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

- β_0 = mean of Y when X = 0
- $\beta_0 + \beta_1$ = mean of Y when X = 1
- β_1 = difference in group means, X =1 minus X = 0
- $SE(\hat{\beta}_1)$ has the usual interpretation
- *t*-statistics, confidence intervals constructed as usual
- This is another way (an easy way) to do difference-in-means analysis
- The regression formulation is especially useful when we have additional regressors

Dummy Variables When There ARE More Than Two Groups

It is straightforward to extend our model to the case where the qualitative variable can take on more than two values.

Consider a model relating the earnings of a child to the education of the child's parents.

Here we assume the education of the parents is classified into four groups:

Parents' education = (1) less than high school, (2) high school, (3)some college, (4) college graduate or higher.

We incorporate the parents' education into the model by creating three dummy variables.

We create three dummy variables despite the fact **that** we have four categories **because** this approach allows us to estimate an intercept for each group, as shown below.

In general, **if** a qualitative variable assumes J outcomes, J-1 dummy variables are included into the model.

Equivalently, one category is always dropped.

This is because knowing one person is not in the J-1 categories tells us they must be in the J^{th} category.

Hence, the J^{th} category is redundant.

More technically, including J dummy variables would create perfect multi-collinearity.

Now, we can structure the model as follows:

$$\ln(Y)_i = \beta_0 + \beta_1 Edu_{1i} + \beta_2 Edu_{2i} + \beta_3 Edu_{3i} + \varepsilon_i$$

where

$$Edu_1 = \begin{cases} 1 \; if\ your\ education\ is\ less\ than\ high\ school \\ 0, otherwise \end{cases}$$

$$Edu_2 = \begin{cases} 1 \; if\ your\ education\ is\ exactly\ high\ school \\ 0, otherwise \end{cases}$$

$$Edu_3 = \begin{cases} 1 \; if\ your\ education\ is\ some\ college \\ 0, otherwise \end{cases}$$

Notice that we did not include a dummy variable indicating their parents with college education or higher.

In effect, we left out

$$Edu_4 = \begin{cases} 1 \; if\ your\ education\ is\ college\ graduate\ or\ higher \\ 0, otherwise \end{cases}$$

We might have left out any group — the choice is ours—but the interpretation of the regression coefficients is affected by the group left out.

To see this, consider the following conditional expectation:

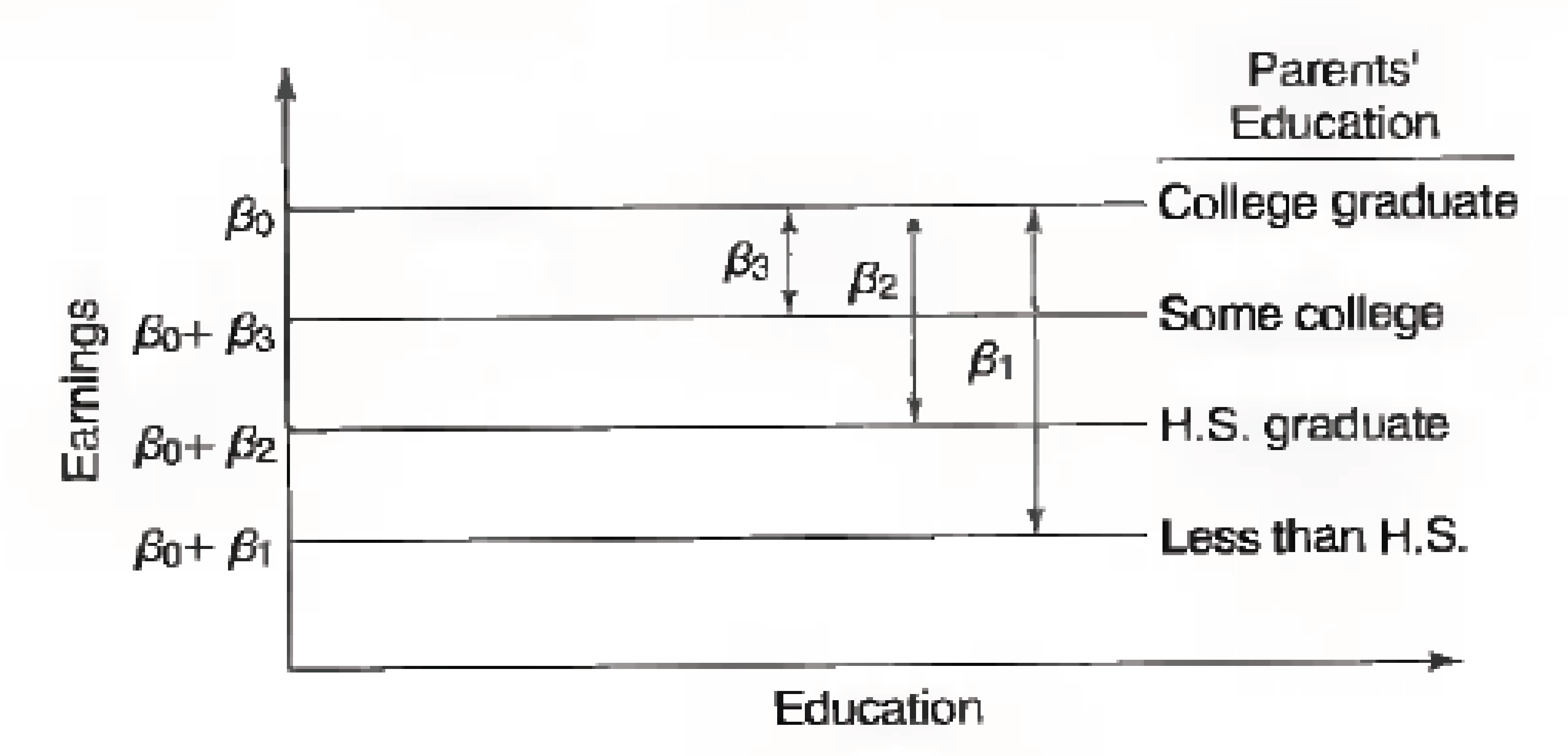
$E(Y_i|Educ_1=1) = \beta_0+\beta_1,$

$E(Y_i|Educ_2=1) = \beta_0+\beta_2,$

$E(Y_i|Educ_3=1) = \beta_0+\beta_3,$

$E(Y_i|Educ_4=1) = \beta_0,$

The interpretation of these relationship is facilitated by the figure below.



Parents' Education
College graduate
Some college
H.S. graduate
Less than H.S.
Earnings
Education
β_0
$\beta_0+\beta_3$
$\beta_0+\beta_2$
$\beta_0+\beta_1$
β_3
β_2
β_1

Notice that the intercept measures the mean value of the dependent variable for the excluded group (i.e., Educ4).

Notice further that

$$E(Y_i|Educ_1=1) - E(Y_i|Educ_4=1) = \beta_1,$$

$$E(Y_i|Educ_2=1) - E(Y_i|Educ_4=1) = \beta_2,$$

$$E(Y_i|Educ_3=1) - E(Y_i|Educ_4=1) = \beta_3,$$

That is, the coefficients on the included dummy variables measure the impact of the corresponding explanatory variable compared to the excluded category.

Thus, in our example, β_1 measures the difference in mean earnings between children whose parents had less than a high school education and those who had completed a college education (i.e., the excluded category).

Similarly, β3 measures the difference in mean earnings between children whose parents had some college education and those who had completed a college education.

We might have excluded some other category.

For example, had we excluded Educ1 and included Educ4, then effects would be compared to those whose parents' education was less than high school.

The results are always compared to the one category that is left out.

Finally, if additional explanatory variables were included in this model, we would simply change the interpretation given above by adding the phrase: other things equal.

Interactive Variables

As we have seen, dummy variables allow the intercept of regression line to vary between different groups as characterized by some qualitative variable.

It is sometimes useful, however, to allow the slope coefficient to vary across the groups as well.

This is accomplished by incorporation by using an interaction variable (or interaction term).

Consider the model again:

$$Earnings_i = \beta_0 + \beta_1 Gender_i + \beta_2 Education_i + \varepsilon_i, (2)$$

As we have seen, the coefficient of Gender tells us the difference between male and female earnings, holding education constant.

$$Earnings_i = \beta_0 + \beta_1 Gender_i + \beta_2 Edu_i + \varepsilon_i, (2)$$

In this model β_1 measures the difference in earnings between male and female, holding education constant.

The slope coefficient on education, β_2, measures the increment in earnings resulting from an additional year of schooling,

$$\frac{\partial Earnings}{\partial Edu} = \beta_2$$

In this specification, this "return" to education is assumed to be equal for men and women.

But it is possible **that** it differs by gender **as** represented in the figure below.

To allow for this possibility we include a new variable in the model is the education variable multiplied by the gender variable.

Such a variable is called an interactive variable or interaction term.

It will capture any interaction between gender and the impact of educations on earnings.

The model is now

$$Earning_i = \beta_0 + \beta_1 Gender_i + \beta_2 Edu_i + \beta_3 Gender_i \times Edu_i + \varepsilon_i$$

Now, to see how the inclusion of an interaction broadens the model, consider the following partial derivative

$$\frac{\Delta Earning}{\Delta Edu} = \beta_2 + \beta_3 Gender = \begin{cases} \beta_2, Gender = 0 \\ \beta_2 + \beta_3, Gender = 1 \end{cases}$$

The return to earnings depends on gender.

For men it's β_2 and for women it's $\beta_2 + \beta_3$.

Thus, β_3 represents the difference in slope between men and women.

Heteroskedasticity and Homoskedasticity, and Homoskedasticity-Only Standard Errors (Section 5.4)

1. What…?
2. Consequences of homoskedasticity
3. Implication for computing standard errors

What do these two terms mean?

If $var(u|X = x)$ is constant – that is, if the variance of the conditional distribution of u given X does not depend on X – then u is said to be ***homoskedastic***.

Otherwise, u is ***heteroskedastic***.

Example: hetero/homoskedasticity in the case of a binary regressor (that is, the comparison of means)

- Standard error when group variances are ***unequal***:

$$SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

- Standard error when group variances are ***equal***:

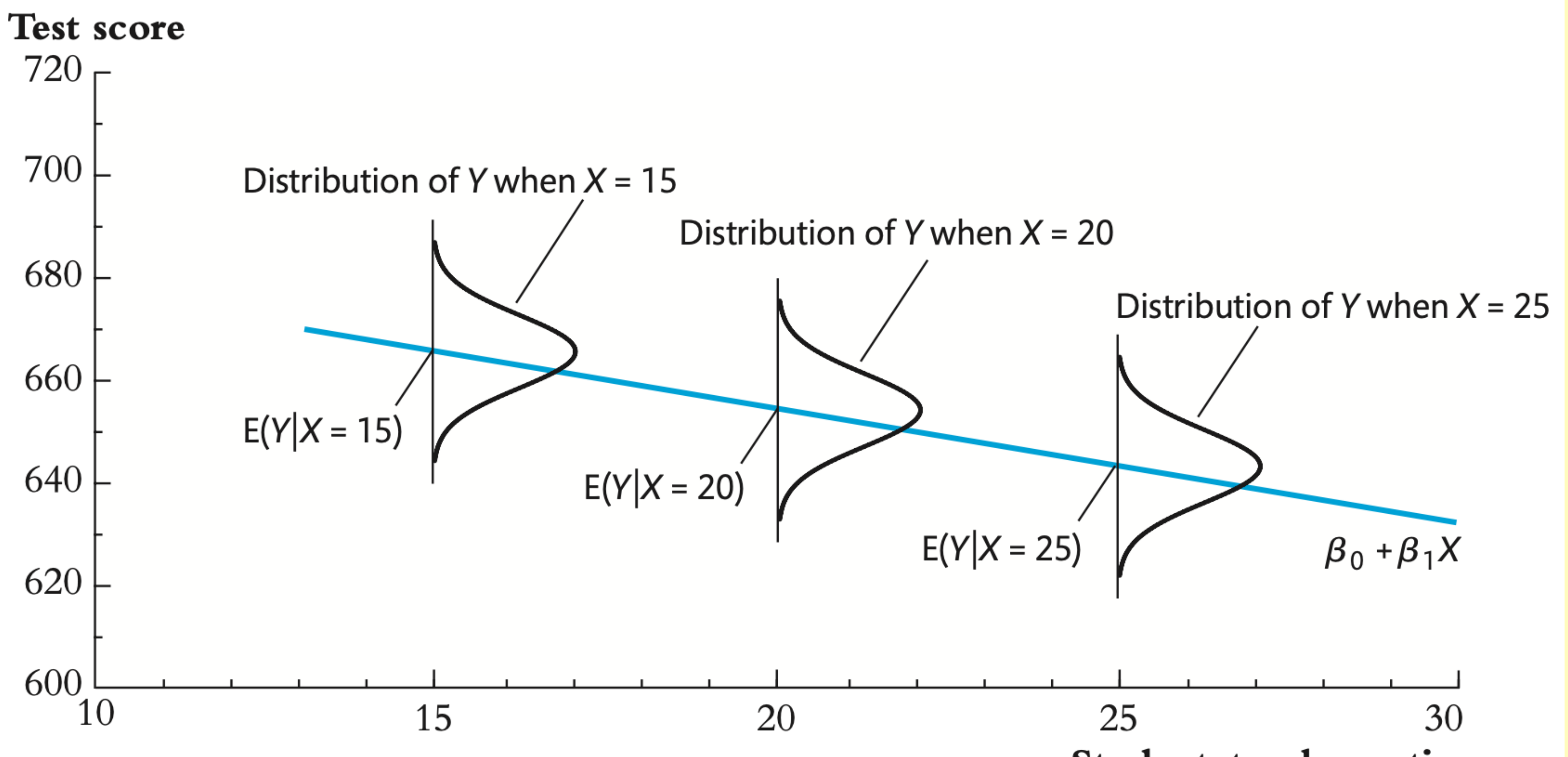
$$SE = s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_l}}$$

$$\text{Where } S_p^2 = \frac{(n_s - 1)S_s^2 + (n_l - 1)S_l^2}{n_s + n_l - 2} \quad \text{(SW, Sect 3.6)}$$

$$S_p = \text{“pooled estimator of } \sigma^2\text{” when } \sigma_l^2 = \sigma_s^2$$

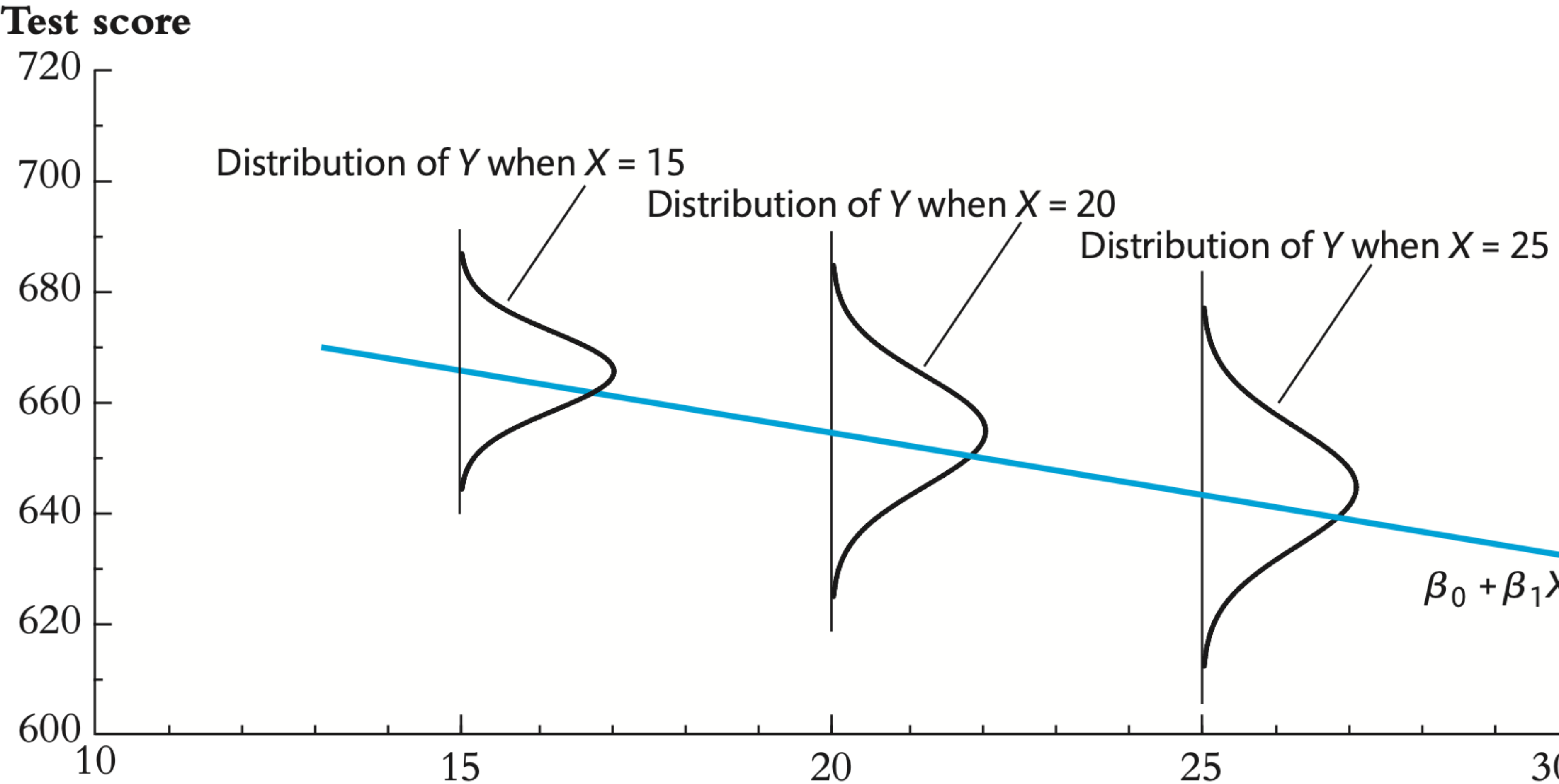
- **Equal** group variances = **homo**skedasticity
- **Unequal** group variances = **hetero**skedasticity

Homoskedasticity in a picture



(a) The errors are homoscedastic

- $E(u|X=x)=0$ ($\beta_0+\beta_1 X$ is the population regression line)
- The variance of *u* ***does not*** depend on *x*

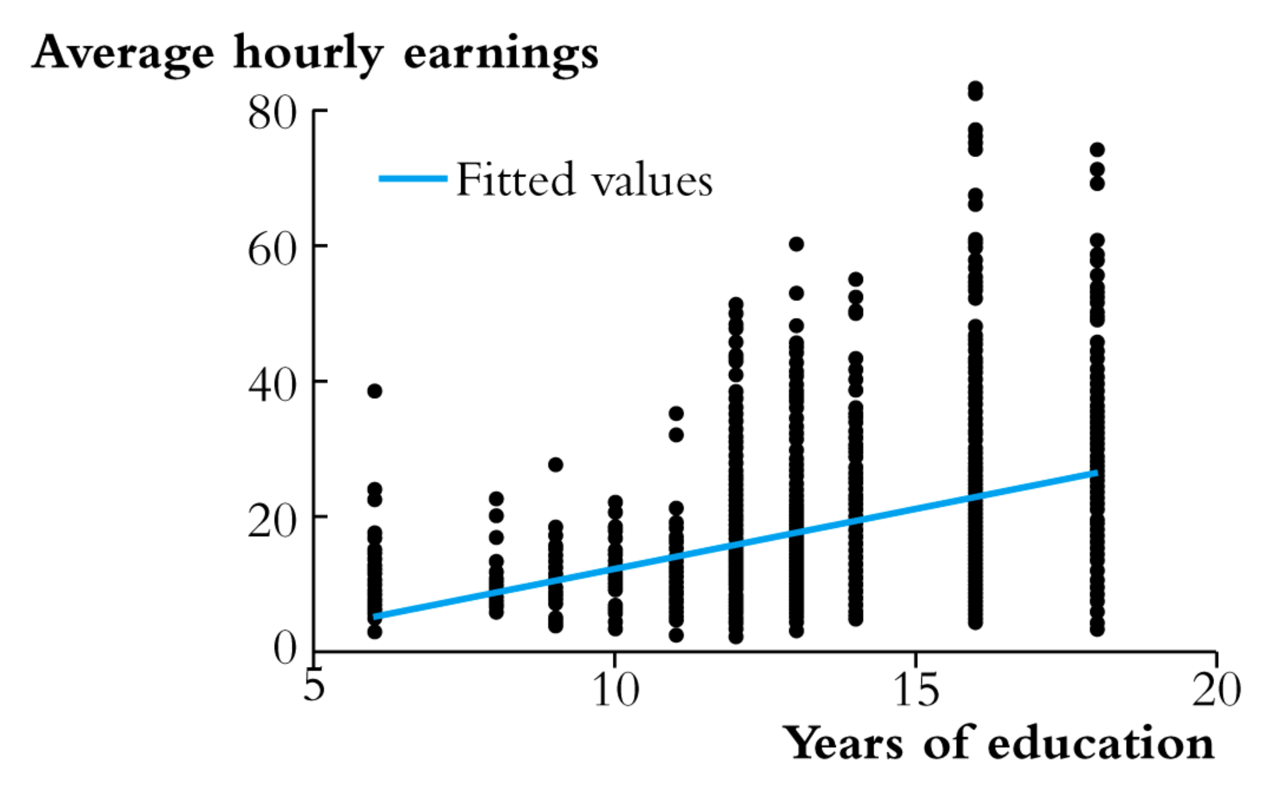


(b) The errors are heteroskedastic

- The variance of *u* ***does*** depend on *x*

A real-data example from labor economics: average hourly earnings vs. years of education (data source: Current Population Survey)

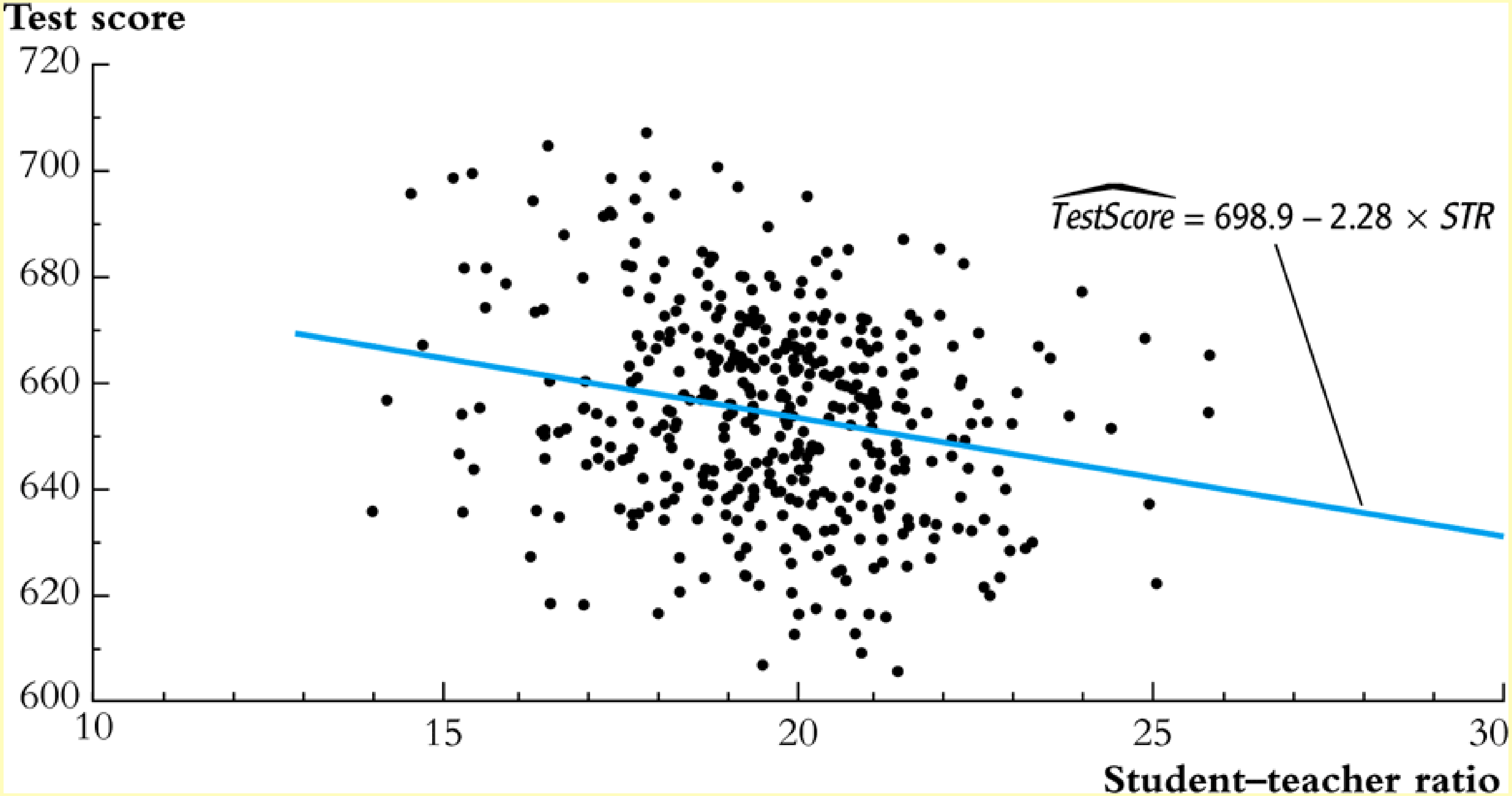
FIGURE 5.3: Scatterplot of Hourly Earnings and Years of Education for 29- to 30-Year-Olds in the United States in 2015



Heteroskedastic or homoskedastic?

Pearson

The class size data



Heteroskedastic or homoskedastic?

So far we have assumed that *u* might be heteroskedastic (without saying so)

Recall the three least squares assumptions:

1. $E(u|X=x)=0$
2. (X_i,Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare

Heteroskedasticity and homoskedasticity concern $var(u|X = x)$. Because we have not explicitly assumed homoskedastic errors, we have implicitly allowed for heteroskedasticity.

What if the errors are in fact homoskedastic? (1 of 2)

- You can prove that OLS has the lowest variance among estimators that are linear in *Y*… a result called the Gauss-Markov theorem that we will return to shortly.
- The formula for the variance of $\hat{\beta}_1$ and the OLS standard error simplifies: If $\text{var}(u_i|X_i = x) = \sigma_u^2$, then

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} \quad \text{(general formula)}$$

$$= \frac{\sigma_u^2}{n\sigma_X^2} \quad \text{(simplification of } u \text{ is homoscedastic)}$$

Note: $\text{var}(\hat{\beta}_1)$ is inversely proportional to $\text{var}(X)$: more spread in X means more information about $\hat{\beta}_1$ – we discussed this earlier but it is clearer from this formula.

What if the errors are in fact homoskedastic? (2 of 2)

- Along with this homoskedasticity-only formula for the variance of $\hat{\beta}_1$, we have homoskedasticity-only standard errors:

Homoskedasticity-only standard error formula:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

Some people (e.g. Excel programmers) find the homoskedasticity-only formula simpler – but it is wrong unless the errors really are homoskedastic.

We now have two formulas for standard errors for $\hat{\beta}_1$

- ***Homoskedasticity-only standard errors*** – these are valid only if the errors are homoskedastic.
- The usual standard errors – to differentiate the two, it is conventional to call these ***heteroskedasticity – robust standard errors***, because they are valid whether or not the errors are heteroskedastic.
- The main advantage of the homoskedasticity-only standard errors is that the formula is simpler. But the disadvantage is that the formula is only correct if the errors are homoskedastic.

In his influential research, the econometrician Halbert White derived appropriate formulas for standard errors under heteroskedasticity.

In the case of simple linear regression we have seen that the standard errors for the slope coefficient are

$$var(\hat{\beta}_1) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} x_i^2 \sigma_i^2$$

where σ_i^2 is the variance of the errors.

In the place of σ_i^2, White proposed using the square of the least squares residual e_i^2.

He has shown that this produces a consistent estimator for the variance of the slope coefficient.

These formulas can be accessed easily in statistical software.

Practical implications

- The homoskedasticity-only formula for the standard error of $\hat{\beta}_1$ and the "heteroskedasticity-robust" formula differ – so in general, *you get different standard errors using the different formulas.*

- Homoskedasticity-only standard errors are the default setting in regression software – sometimes the only setting (e.g. Excel). To get the general "heteroskedasticity-robust" standard errors you must override the default.

- **If you don't override the default and there is in fact heteroskedasticity, your standard errors (and *t*-statistics and confidence intervals) will be wrong – typically, homoskedasticity-only *SE*s are too small.**

Heteroskedasticity-robust standard errors in STATA

```
regress testscr str, robust
Regression with robust standard errors  Number of obs =     420
        F(  1,   418) =   19.26
         Prob > F      =  0.0000
         R-squared     =  0.0512
        Root MSE      =  18.581
-------------------------------------------------------------------------
                        Robust
testscr |      Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
-------------------------------------------------------------------------
qqqqstr |  -2.279808   .5194892    -4.39   0.000    -3.300945  -1.258671
qq_cons |    698.933   10.36436    67.44   0.000     678.5602   719.3057

-------------------------------------------------------------------------
```

- If you use the "**, robust**" option, STATA computes heteroskedasticity-robust standard errors
- Otherwise, STATA computes homoskedasticity-only standard errors

The bottom line

- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity).
- The two formulas coincide (when *n* is large) in the special case of homoskedasticity
- So, *you should always use heteroskedasticity-robust standard errors*.

Some Additional Theoretical Foundations of OLS (Sections 5.5)

We have already learned a very great deal about OLS: OLS is unbiased and consistent; we have a formula for heteroskedasticity-robust standard errors; and we can construct confidence intervals and test statistics.

Also, a very good reason to use OLS is that everyone else does – so by using it, others will understand what you are doing. In effect, OLS is the language of regression analysis, and if you use a different estimator, you will be speaking a different language.

Still, you may wonder

- Is this really a good reason to use OLS? Aren't there other estimators that might be better – in particular, ones that might have a smaller variance?
- Also, what happened to our old friend, the Student *t* distribution?

So we will now answer these questions – but to do so we will need to make some stronger assumptions than the three least squares assumptions already presented.

The Homoskedastic Normal Regression Assumptions

These consist of the three LS assumptions, plus two more:

1. $E(u|X=x)=0.$
2. $(X_i,Y_i),\ i=1,\ldots,n,$ are i.i.d.
3. Large outliers are rare $(E(Y^4)<\infty, E(X^4)<\infty).$
4. *u* is homoskedastic
5. *u* is distributed $N(0,\sigma^2)$

- Assumptions 4 and 5 are more restrictive – so they apply to fewer cases in practice. However, if you make these assumptions, then certain mathematical calculations simplify and you can prove strong results – results that hold if these additional assumptions are true.
- We start with a discussion of the efficiency of OLS

Efficiency of OLS, part I: The Gauss-Markov Theorem (1 of 2)

Under assumptions 1-4 (the basic three, plus homoskedasticity), $\hat{\beta}_1$ has the smallest variance among *all linear estimators* (estimators that are linear functions of $Y_1,\ldots, Y_n$). This is the ***Gauss - Markov theorem***.

Comments

- The GM theorem is proven in SW Appendix 5.2

Efficiency of OLS, part I: The Gauss-Markov Theorem (2 of 2)

- $\hat{\beta}_1$ is a linear estimator, that is, it can be written as a linear function of $Y_1,\ldots,Y_n$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{n}\sum_{i=1}^{n} w_i u_i \ ,$$

$$\text{where } w_i = \frac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

- The G-M theorem says that among all possible choices of $\{w_i\}$, the OLS weights yield the samllest $\text{var}(\hat{\beta}_1)$

Efficiency of OLS, part II

- Under <u>all five</u> homoskedastic normal regression assumptions – including normally distributed errors – $\hat{\beta}_1$ has the smallest variance of <u>*all*</u> consistent estimators (linear *or* nonlinear functions of $Y_1,\ldots,Y_n$), as - $>\infty$.

- This is a pretty amazing result – it says that, if (in addition to LSA 1-3) the errors are homoskedastic and normally distributed, then OLS is a better choice than any other consistent estimator. And because an estimator that isn't consistent is a poor choice, this says that OLS really is the best you can do – if all five extended LS assumptions hold. (The proof of this result is beyond the scope of this course and isn't in SW – it is typically done in graduate courses.)

Some not-so-good thing about OLS (1 of 2)

The foregoing results are impressive, but these results – and the OLS estimator – have important limitations.

1. The GM theorem really isn't that compelling:
   - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
   - The result is only for linear estimators – only a small subset of estimators (more on this in a moment)
2. The strongest optimality result ("part II" above) requires homoskedastic normal errors – not plausible in applications (think about the hourly earnings data!)

Some not-so-good thing about OLS

3. OLS is more sensitive to outliers than some other estimators. In the case of estimating the population mean, if there are big outliers, then the median is preferred to the mean because the median is less sensitive to outliers – it has a smaller variance than OLS when there are outliers. Similarly, in regression, OLS can be sensitive to outliers, and if there are big outliers other estimators can be more efficient (have a smaller variance). One such estimator is the least absolute deviations (LAD) estimator:

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left| Y_i - (b_0 + b_1 X_i) \right|$$

In virtually all applied regression analysis, OLS is used – and that is what we will do in this course too.

Inference if *u* is homoskedastic and normally distributed: the Student *t* distribution (Section 5.6)

Recall the five homoskedastic normal regression assumptions:

1. $E(u|X = x) = 0.$
2. (X_i,Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare $(E(Y^4) < \infty, E(X^4) < \infty)$.
4. *u* is homoskedastic
5. *u* is distributed $N(0,\sigma^2)$

If all five assumptions hold, then:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed *for all n* (!)
- the *t*-statistic has a Student *t* distribution with *n* – 2 degrees of freedom – this holds exactly *for all n* (!)

Normality of the sampling distribution of under 1–5

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} w_i u_i, \text{ where } w_i = \frac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

What is the distribution of a weighted average of normals?

Under assumptions 1 – 5:

$$\hat{\beta}_1 - \beta_1 \sim N\left(0, \frac{1}{n^2}\left(\sum_{i=1}^{n} w_i^2\right)\sigma_u^2\right) \quad (*)$$

Substituting w_i into (*) yields the homoskedasticity-only variance formula.

In addition, under assumptions 1 – 5, under the null hypothesis the *t* statistic has a Student *t* distribution with *n* – 2 degrees of freedom

- Why $n - 2$? because we estimated 2 parameters, β_0 and β_1
- For $n < 30$, the *t* critical values can be a fair bit larger than the $N(0,1)$ critical values
- For $n > 50$ or so, the difference in t_{n-2} and $N(0,1)$ distributions is negligible. Recall the Student *t* table:

degrees of freedom	5% t-distribution critical value
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

Practical implication

- If n < *50* ***and*** you really believe that, for your application, u is homoskedastic and normally distributed, then use the t_{n-2} instead of the $N(0,1)$ critical values for hypothesis tests and confidence intervals.
- In most econometric applications, there is no reason to believe that u is homoskedastic and normal – usually, there are good reasons to believe that neither assumption holds.
- Fortunately, in modern applications, $n > 50$, so we can rely on the large-n results presented earlier, based on the CLT, to perform hypothesis tests and construct confidence intervals using the large-n normal approximation.

Summary and Assessment (Section 5.7)

- The initial policy question:

  Suppose new teachers are hired so the student-teacher ratio falls by one student per class. What is the effect of this policy intervention (“treatment”) on test scores?

- This question requires an estimate of the causal effect on test scores of a change in the *STR*. Does our regression analysis using the California data set provide a compelling estimate of this causal effect?

  Not really – districts with low *STR* tend to be ones with lots of other resources and higher income families, which provide kids with more learning opportunities outside school…this suggests that $\mathrm{corr}(u_i, STR_i) > 0$, so $E(u_i|X_i) \neq 0$.

- It seems that we have omitted some factors, or variables, from our analysis, and this has biased our results…