Univariate Regression: Hypothesis Tests and Confidence Intervals (SW Ch. 5)

Part 3

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Spring 2022

Brooklyn College
Mar 28, 2022

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**

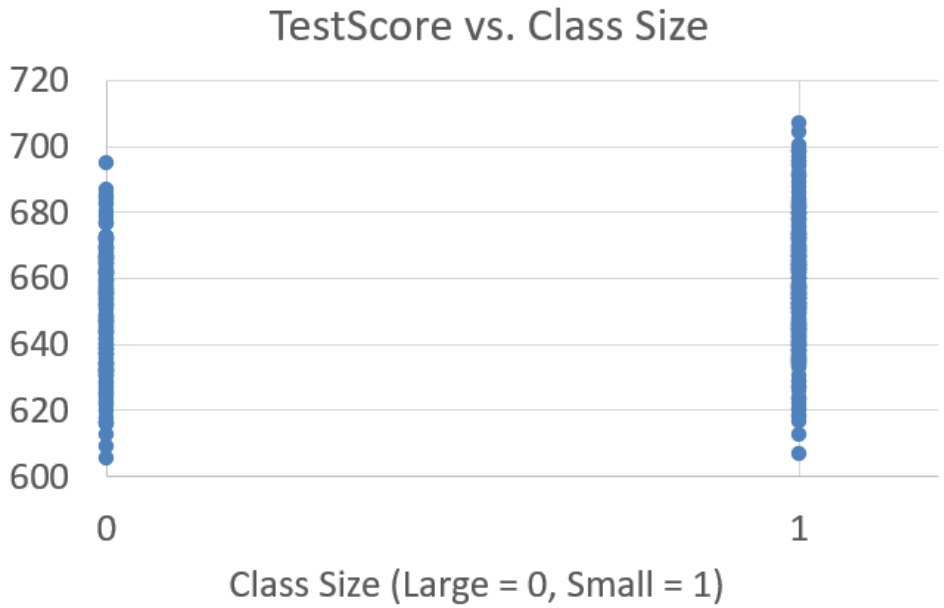
Heteroskedasticity and Homoskedasticity, and Homoskedasticity-Only Standard Errors (Section 5.4)

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

What do these two terms mean?

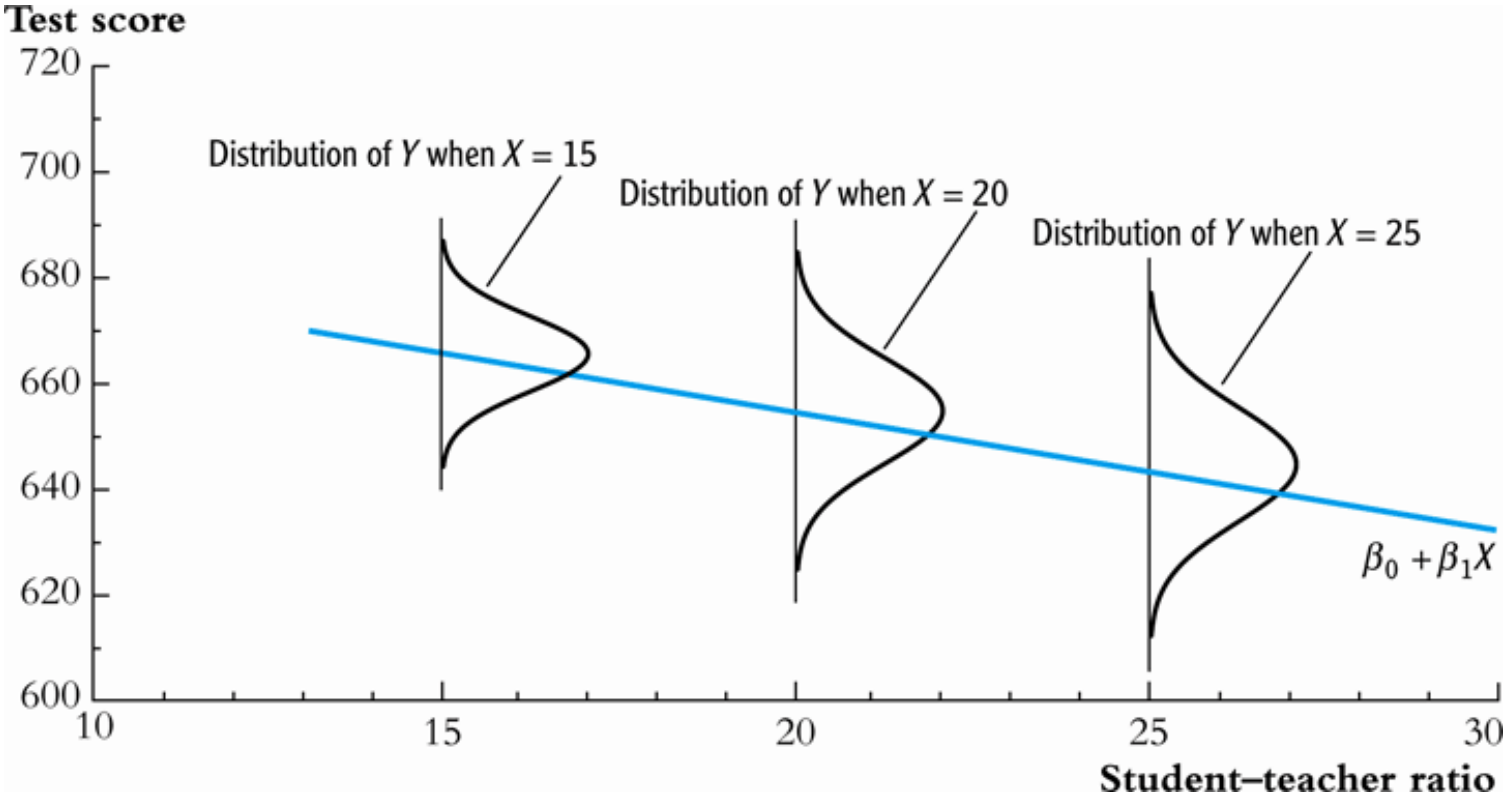
If var($u|X=x$) is constant – that is, if the variance of the conditional distribution of u given X does not depend on X – then u is said to be ***homoskedastic***. Otherwise, u is ***heteroskedastic***.

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



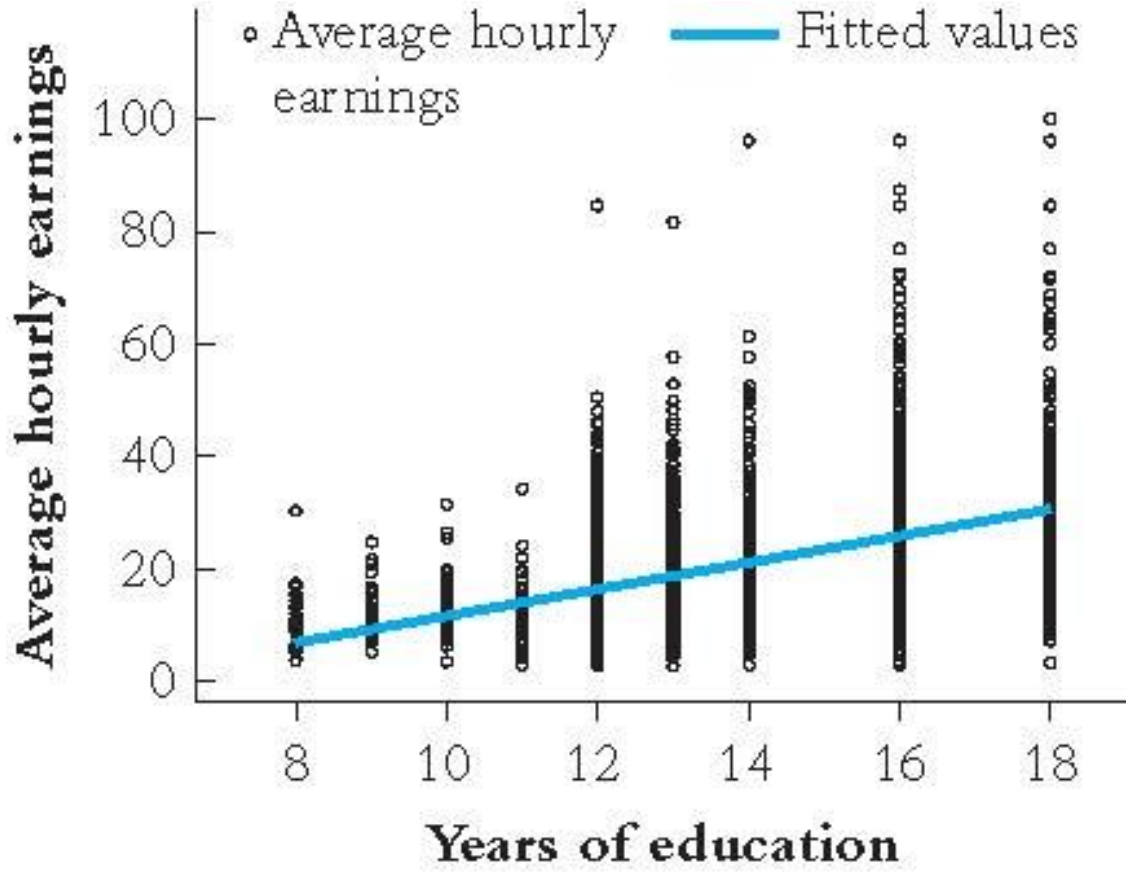
- Standard error when group variances are ***unequal***: $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$
- Standard error when group variances are ***equal***: $SE = s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_l}}$ (SW, Sect 3.6)
  S_p = "pooled estimator of σ^2" when $\sigma_l^2 = \sigma_s^2$
- **Equal** group variances = **homo**skedasticity
- **Unequal** group variances = **hetero**skedasticity

Heteroskedasticity in a picture:



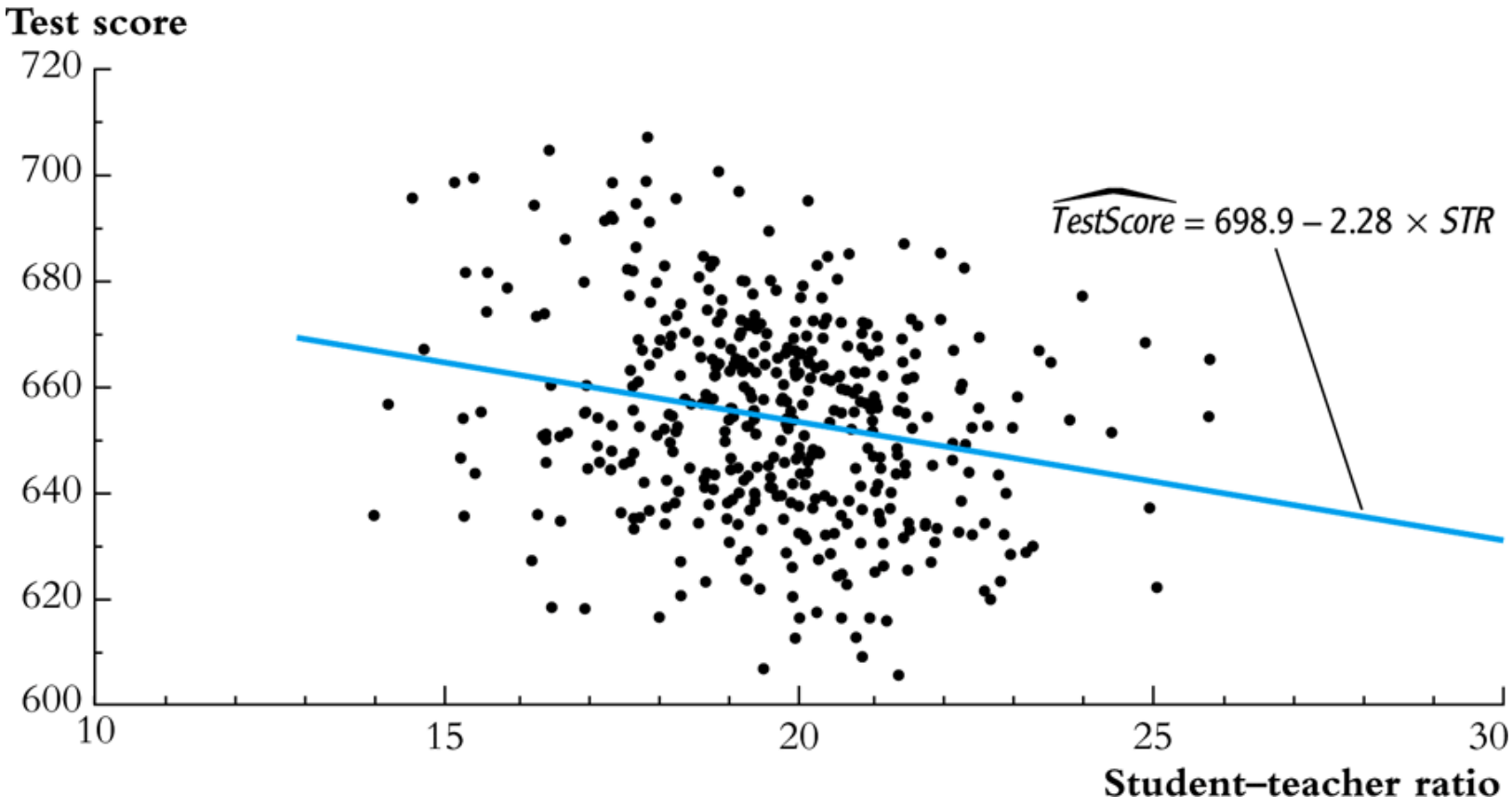
- $E(u|X=x) = 0$ ($\beta_0 + \beta_1 X$ is the population regression line)
- The variance of u ***does not*** depend on x

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



Heteroskedastic or homoskedastic?

The class size data:



Heteroskedastic or homoskedastic?

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

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 $\text{var}(u_i|X_i = x) =$, then

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

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

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:

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{(homoskedasticity-only) and} \qquad (5.29)$$

$$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) s^2_{\hat{u}}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{(homoskedasticity-only),} \qquad (5.30)$$

where $s^2_{\hat{u}}$ is given in Equation (4.19). The homoskedasticity-only standard errors are the square roots of $\tilde{\sigma}^2_{\hat{\beta}_0}$ and $\tilde{\sigma}^2_{\hat{\beta}_1}$.

$$SER = s_{\hat{u}} = \sqrt{s^2_{\hat{u}}}, \text{ where } s^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2}, \qquad (4.19)$$

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.

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 there is in fact heteroskedasticity, your standard errors (and *t*-statistics and confidence intervals) will be wrong – typically, homoskedasticity-only *SE*s are smaller**

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 $\hat{\beta}_1$ inconsistent if there is heteroskedasticity).
- The two formulas coincide (when *n* is large) in the special case of homoskedasticity

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 $Y_1,...,Y_n$:

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

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

- 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 all five homoskedastic normal regression assumptions – including normally distributed errors – $\hat{\beta}_1$ has the smallest variance of *all* 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 (2 of 2)

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 $\widehat{\beta}_1 - \beta_1$ under assumptions 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 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…