ECON 7310 Elements of Econometrics
Week 3: Linear Regression with One Regressor – Hypothesis Tests and Confidence Intervals

David Du[1]

[1] University of Queensland

Draft

Outline

- Hypothesis tests concerning β_1
- Confidence intervals for β_1
- Regression when X is binary
- Heteroskedasticity and homoskedasticity
- 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

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

where $\beta_1 = \Delta Y / \Delta X$, for an autonomous change in X (causal effect)

- **Estimator:** the OLS estimator β_1.
- **The Sampling Distribution of β_1:**
  To derive the large-sample distribution of β_1, we make the following assumptions:
- **The 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 ($E(X^4) < \infty$, $E(Y^4) < \infty$).

Hypothesis Testing of β_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:

    $$H_0 : \beta_1 = \beta_1^0 \quad vs. \quad H_1 : \beta_1 \neq \beta_1^0$$

    where β_1^0 is the hypothesized value under the null.
  - Null hypothesis and **one-sided** alternative:

    $$H_0 : \beta_1 = \beta_1^0 \quad vs. \quad H_1 : \beta_1 < \beta_1^0$$

- We are going to use asymptotic distribution of $\widehat{\beta}_1$, i.e., Under the Least Squares Assumptions, for *n* large,

$$\frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} \overset{approx}{\sim} \mathcal{N}(0, 1)$$

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

- **In general:**
$$t = \frac{\text{estimator} - \text{hypothesised value}}{\text{standard error of the estimator}}$$
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:
$$t = \frac{\overline{Y} - \mu_Y^0}{SE(\overline{Y})}$$
- For testing β_1:
$$t = \frac{\widehat{\beta}_1 - \beta_1^0}{SE(\widehat{\beta}_1)}$$

Asymptotic Distribution of t

- Recall that $SE(\widehat{\beta}_1)$ is a consistent estimate of $\sqrt{V(\widehat{\beta}_1)}$. We discussed about $V(\widehat{\beta}_1)$ in the previous lecture (but did not derive it).
- Econometrics softwares such as Stata and R report standard errors. (We will not derive $SE(\widehat{\beta}_1)$.)
- We use the fact that when n is large,

$$t = \frac{\widehat{\beta}_1 - \beta_1^0}{SE(\widehat{\beta}_1)} \overset{approx}{\sim} \mathcal{N}(0,1)$$

  under H_0, i.e., if the true value of β_1 is really β_1^0.

To test $H_0 : \beta_1 = \beta_1^0$ *vs.* $H_0 : \beta_1 \neq \beta_1^0$

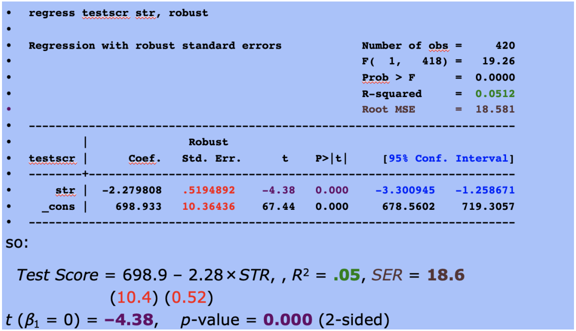
- Construct the *t*-statistic, i.e., $t = (\widehat{\beta}_1 - \beta_1^0)/SE(\widehat{\beta}_1)$ and choose the level of significance α. Suppose we choose $\alpha = 0.05$. Then, make a decision on whether to reject H_0, using any of the following criteria
  1. Reject H_0 if $|t| > 1.96$
  2. Reject H_0 if *p*-value < α.
  3. Reject H_0 if β_1^0 is outside the 95% confidence interval,

$$\widehat{\beta}_1 \pm 1.96 \times SE(\widehat{\beta}_1) = [\widehat{\beta}_1 - 1.96SE(\widehat{\beta}_1), \widehat{\beta}_1 + 1.96SE(\widehat{\beta}_1)]$$

- This procedure relies on the large *n* approximation that *t* is normally distributed under H_0; typically $n = 50$ is large enough for the approximation to be adequate
- If $\alpha = 0.01$, use 2.576 instead of 1.96

Example: *TestScores* and *STR*, California data

- We test $H_0 : \beta_1 = 0$ at $\alpha = 0.05$. Stata output is given as



```
•   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:

$$\widehat{Test\ Score} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, \ , R^2 = .05, \ SER = 18.6$$

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

- We reject H_0 because
  1. $|t| = |(-2.28 - 0)/0.52| = 4.38 > 1.96$
  2. p-value = 0.000 < α=0.05
  3. The 95% CI = $[-3.30, -1.26]$ does not include the hypothesised value 0.
- We could do similar test for the intercept β_0 using the Stata output.

Some comments on *p*-values and *CI*s

- For a given *t*-statistic, the *p*-value measures $\Pr(|Z| > |t|)$ where $Z \sim \mathcal{N}(0, 1)$. We can consider the *p*-value as the largest α at which we reject H_0.
- The $(1 - \alpha) \times 100\%$ confidence interval is the set of parameter values that cannot be rejected by two-sided test at significance level of α.
- Consider a repeated sampling, i.e., every day we collect a sample of the same size. Then, each day, we will have sample (data) so that we have different estimates, different SE, and different a 95% CI.
- In the repeated sampling, the 95% CI would contain the true parameter (β_1) in 95% of times.

Regression when X is Binary (Section 5.3)

- Sometimes a regressor is binary:
  - $X = 1$ if small class size, $X = 0$ if not
  - $X = 1$ if female, $X = 0$ if male
  - $X = 1$ if treated (experimental drug), $X = 0$ 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

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

where X_i is binary ($X_i = 0$ or $X_i = 1$)

- Linear regression = we are estimating the conditional expectation function $E[Y|X]$ under the assumption that $E[Y|X] = \beta_0 + \beta_1 X$
- When $X_i = 0$, we have $Y_i = \beta_0 + u_i$. That is, the mean of Y_i is β_0,

$$E[Y|X = 0] = \beta_0$$

- When $X_i = 1$, we have $Y_i = \beta_0 + \beta_1 + u_i$. That is, the mean of Y_i is $\beta_0 + \beta_1$,

$$E[Y|X = 1] = \beta_0 + \beta_1$$

- So

$$\begin{aligned} \beta_1 &= E[Y|X = 1] - E[Y|X = 0] \\ &= \text{population difference in group means} \end{aligned}$$

Example: *TestScore* and *STR*

- But, suppose we observe $D_i = 1$ if $STR_i < 20$ and $D_i = 0$ if $STR_i \geq 20$
- **OLS regression:**

$$\begin{array}{rcrcrl} TestScore_i = & 650.0 & + & 7.4 & D_i \\ & (1.3) & & (1.8) & \end{array}$$

- **Tabulation of group means:**

Class Size	Average score $(\overline{Y})$	Standard deviation $(SDev(\overline{Y}))$	n
Small ($STR_i < 20$)	657.4	19.4	238
Large ($STR_i \geq 20$)	650.0	17.9	182

- **Difference in means:** $\overline{Y}_{small} - \overline{Y}_{large} = 657.4 - 650.0 = 7.4$
- **Standard error:** $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_\ell^2}{n_\ell}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = 1.8$

regression when X_i is binary (0/1)

- β_0 = mean of Y when $X = 0$
- $\beta_0 + \beta_1$ = mean of Y when $X = 1$
- β_1 = difference in means between those two groups.
- We can construct *t*-statistics and confidence intervals 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 (as we will see next week)

Heteroskedasticity and Homoskedasticity (Section 5.4)

1. What are heteroskedasticity and homoskedasticity?
   - The error u is said to be homoskedasticity if $V(u|X = x)$ is constant.
   - The error u is said to be heteroskedasticity, otherwise.
2. Consequences of homoskedasticity
3. Implication for computing standard errors

Example: hetero/homoskedasticity in the case of a binary regressor

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

$$SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_\ell^2}{n_\ell}}$$

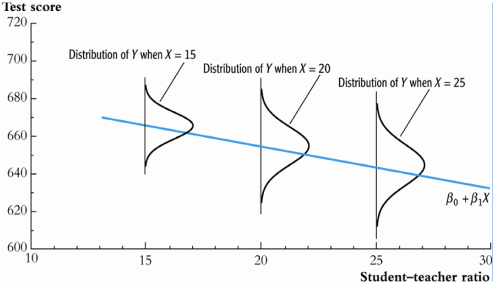
- Standard error when group variances are **equal**:

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

  where $s_p =$ "pooled estimator of the standard deviation" (Section 3.6)

- Equal group variances = homoskedasticity
- Unequal group variances = heteroskedasticity

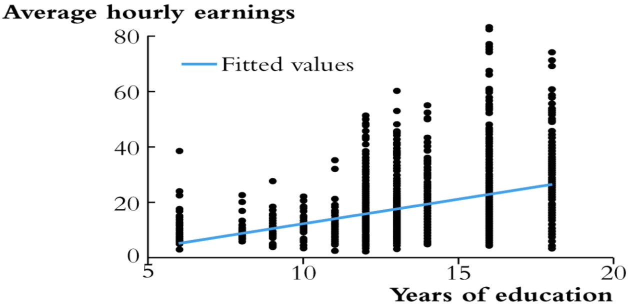
Heteroskedasticity in a picture:



- This shows the conditional distribution of test scores for three different class sizes.
- The distributions become more spread out (have a larger variance) for larger class sizes.
- Because $V(u|X = x)$ depends on x, u is heteroskedastic.

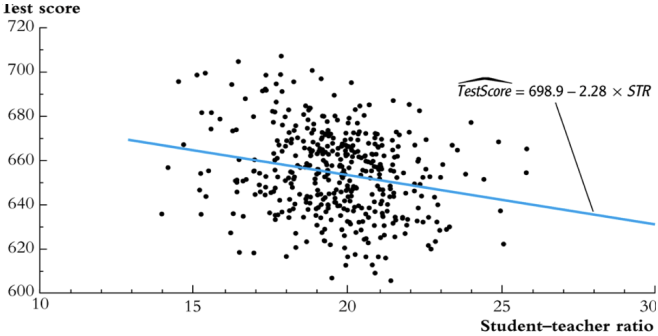
Another example:

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 ($E(X^4) < \infty$, $E(Y^4) < \infty$).

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

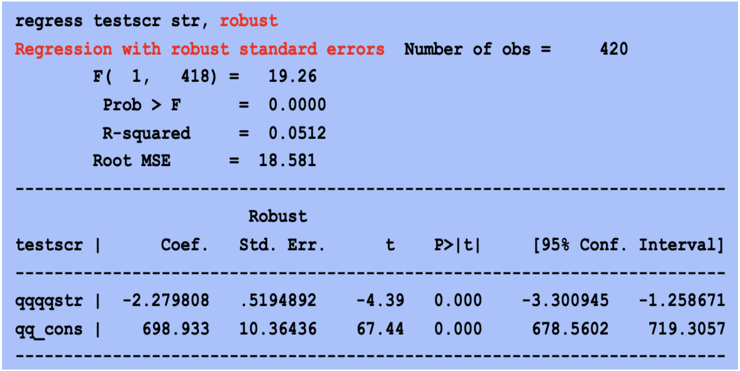
What if the errors are in fact homoskedastic?

- You can prove that OLS has the lowest variance among estimators that are linear in Y; This is a result called the Gauss-Markov theorem that we will return to shortly.
- The formula for $\widehat{\beta}_1$ stays the same. But, the formula for the variance of $\widehat{\beta}_1$ becomes simpler. So does the formula of $SE(\widehat{\beta}_1)$.
- **Homoskedasticity-only standard errors** 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.

Two standard errors? Practical Implications

- 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.
- Errors are likely to be heteroskedastic in almost all economic data. So, if you use homoskedasticity-only standard errors, your inference (testing, confidence intervals) will be very likely to be wrong.
- Even if data are homoskedastic (again, very unlikely), heteroskedasticity robust standard errors are still valid.
- Hence, you are recommended to always use heteroskedasticity robust standard errors, especially for micro-level data.

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

Some Additional Theoretical Foundations of OLS (Section 5.5)

- We have already learned a lot 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.
- 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

Efficiency of OLS estimators

- Let's consider extended least square assumptions;
  1. $E(u|X = x) = 0$.
  2. (X_i, Y_i), $i = 1, \dots, n$, are i.i.d.
  3. Large outliers are rare ($E(X^4) < \infty$, $E(Y^4) < \infty$).
  4. u is homoskedastic
  5. u is normally distributed.
- **Gauss-Markov theorem:** Under extended LS assumptions 1–4 (the basic three, plus homoskedasticity), $\widehat{\beta}_1$ has the smallest variance among all linear unbiased estimators[1] (unbiased estimators that are linear functions of $Y_1, \dots, Y_n$).
- **Optimality of OLS:** Under extended LS assumptions 1–5, $\widehat{\beta}_1$ has the smallest variance of all consistent estimators (linear or nonlinear functions of $Y_1, \dots, Y_n$), as $n \to \infty$.

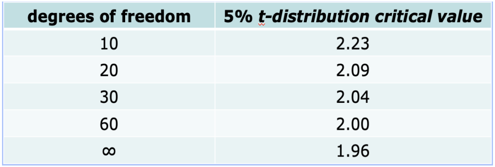
[1] i.e., OLS estimator is the best linear unbiased estimator (BLUE)

Some not-so-good thing about OLS

- 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 is not plausible for many economic data
     - The result is only for linear estimators – only a small subset of estimators (more on this in a moment)
  2. The optimality result requires homoskedastic normal errors – not plausible in applications
  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
- In almost all applied regression analysis, OLS is used – and that is what we will do in this course, too.

How about Student *t* distribution?

- Under the five extended LS assumptions,
  - OLS estimators are also 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
- 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 $\mathcal{N}(0,1)$ is negligible.



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 $\mathcal{N}(0, 1)$ 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 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?
- Does our regression analysis using the California data set answer this convincingly?
  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...