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:

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

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.

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

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

- The estimated regression line is

$$\widehat{TestScore} = 698.9 - 2.28\times STR$$

- 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)
- $t = (\hat{\beta}_1 - \beta_{1,0})/SE(\hat{\beta}_1)$
- p-value = area under standard normal outside t^{act} (large n)

Confidence Intervals:

- 95% confidence interval for β_1 is $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$
- 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.

Exercise 5.1

Suppose that a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression,

$$\widehat{TestScore} = \underset{(20.4)}{520.4} - \underset{(2.21)}{5.82} \times CS, \quad R^2 = 0.08, \quad SER = 11.5.$$

a. Construct a 95% confidence interval for β_1, the regression slope coefficient.
b. Calculate the *p*-value for the two-sided test of the null hypothesis H_0: $\beta_1 = 0$. Do you reject the null hypothesis at the 5% level? At the 1% level?
c. Calculate the *p*-value for the two-sided test of the null hypothesis H_0: $\beta_1 = -5.6$. Without doing any additional calculations, determine whether -5.6 is contained in the 95% confidence interval for β_1.
d. Construct a 99% confidence interval for β_0.

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

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ where } X \text{ is binary } (X_i = 0 \text{ or } 1):$$

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:

$$\beta_1 = E(Y_i|X_i=1) - E(Y_i|X_i=0)$$
$$= \text{population difference in group means}$$

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

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

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

Tabulation of group means:

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

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

Standard error: $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = \mathbf{1.8}$

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 (*as we will very soon*)

Exercise 5.2

Suppose that a researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression,

$$\widehat{Wage} = \underset{(0.23)}{12.52} - \underset{(0.36)}{2.12} \times Male, \;\; R^2 = 0.06, \;\; SER = 4.2,$$

Where *Wage* is measured in \$/hour and *Male* is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the wage gender gap as the difference in mean earnings between men and women.

a. What is the estimated gender gap?
b. Is the estimated gender gap significantly different from zero? (Compute the *p*-value for testing the null hypothesis that there is no gender gap.)
c. Construct a 95% confidence interval for the gender gap.
d. In the sample, what is the mean wage of women? Of men?
e. Another researcher uses these same data, but regresses *Wages* on *Female*, a variable that is equal to 1 if the person is female and 0 if the person is a male. What are the regression estimates calculated from this regression?