ECO220Y

Simple Regression: Testing the Slope

Readings: Chapter 18 (Sections 18.3-18.5)

Winter 2012

Lecture 19

Simple Regression Model

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

$$\Uparrow$$

Model

- OLS produces b_0 and b_1
- Estimated regression line: $\boxed{\hat{y} = b_0 + b_1 x}$
- Are b_0 and b_1 point or interval estimators of the population parameters?
- What do we need to obtain interval estimators?
- $\varepsilon_i \sim N(0, \sigma^2)$

Standard Error of the Slope Estimate

Standard error of the slope estimate also tells us how precisely we are able to estimate the slope. (Why?)

$$b_1 = \frac{s_{xy}}{s_x^2} \Rightarrow V(b_1) = V\left[\frac{s_{xy}}{s_x^2}\right]$$

$$\sigma_{b_1}^2 = \frac{\sigma^2}{(n-1)s_x^2} \qquad \sigma_{b_1} = \sqrt{\frac{\sigma^2}{(n-1)s_x^2}}$$

$$b_1 \sim N(\beta, \sigma_{b_1}^2)$$

But: We do not know σ!

Variance of ε_i

- $\varepsilon_i \sim N(0, \sigma^2)$
- Do we observe ε_i?
- We cannot compute σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})^2}{N} = \frac{\sum_{i=1}^{N}(\varepsilon_i - 0)^2}{N}$$

What implicit assumption do we make when setting $\bar{\varepsilon} = 0$?

Estimate of ε_i

- Residual e_i is an estimate of the error term ε_i
- Residual e_i is the difference between the estimated model and y_i
- Error ε_i is the difference between the true model and y_i
- We will use variability of residuals to estimate the variance of the errors.

Variance of e_i

$$\sigma^2 = \frac{\sum_{i=1}^{N} (\varepsilon_i - 0)^2}{N} \qquad s_e^2 = \frac{\sum_{i=1}^{n} (e_i - 0)^2}{n-2}$$

- s_e^2 is an unbiased estimate of σ^2
- $n - 2$ in the denominator because we used up two degrees of freedom to compute estimates of β_0 and β_1 to find e_i: $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$

Standard Error of Estimate: s_e

Standard error of estimate, s_e, is an estimate of σ, where σ comes from $\varepsilon_i \sim N(0, \sigma^2)$

$$s_e^2 = \frac{\sum_{i=1}^{n}(e_i - 0)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2} = \frac{SSE}{n-2}$$

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

Variance of the Slope Estimate

$$b_1 = \frac{s_{xy}}{s_x^2} \Rightarrow V(b_1) = V\left[\frac{s_{xy}}{s_x^2}\right]$$

$$\sigma_{b_1}^2 = \frac{\sigma^2}{(n-1)s_x^2} \qquad \sigma_{b_1} = \sqrt{\frac{\sigma^2}{(n-1)s_x^2}}$$

$$b_1 \sim N(\beta, \sigma_{b_1}^2)$$

We do not know σ, but we can estimate s_e!

Solution: Replace σ with s

Standard Error of the slope estimate (b_1):

$$\boxed{s_{b_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}}$$

Consequences of uncertainty: For hypothesis testing use Student t (not z)
Example of standard notation:

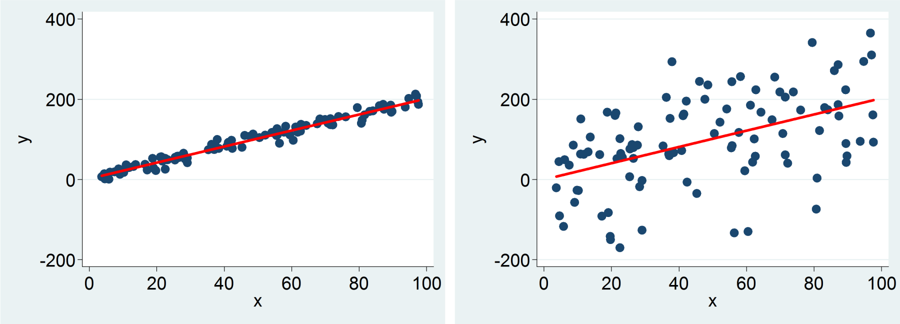
$$\begin{array}{lccc} \hat{y} = & 38.25 & - & 2.68x \\ & (5.8) & & (0.05) \\ & \uparrow & & \uparrow \\ & s_{b_0} & & s_{b_1} \end{array}$$

What affects precision of the slope estimate?

$$s_{b_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}$$

- Sample size
- Variance of independent variable
- Standard Error of estimate, s_e

Difference Between Two Graphs?



For which b_1 will be a more precise estimate of β_1?

Testing the Slope

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

$$\Uparrow$$

Model

Set of Statistical Hypotheses:

Two-tailed	One-tailed	One-tailed
$H_0 : \beta_1 = \beta_1^0$	$H_0 : \beta_1 = \beta_1^0$	$H_0 : \beta_1 = \beta_1^0$
$H_1 : \beta_1 \neq \beta_1^0$	$H_1 : \beta_1 > \beta_1^0$	$H_1 : \beta_1 < \beta_1^0$

where β_1^0 is any number, does not have to be 0.

test-statistics: $\boxed{t = \frac{b_1 - \beta_1^0}{s_{b_1}}}$

$t \sim$ Student t, with $n - 2$ degrees of freedom

Statistical Significance

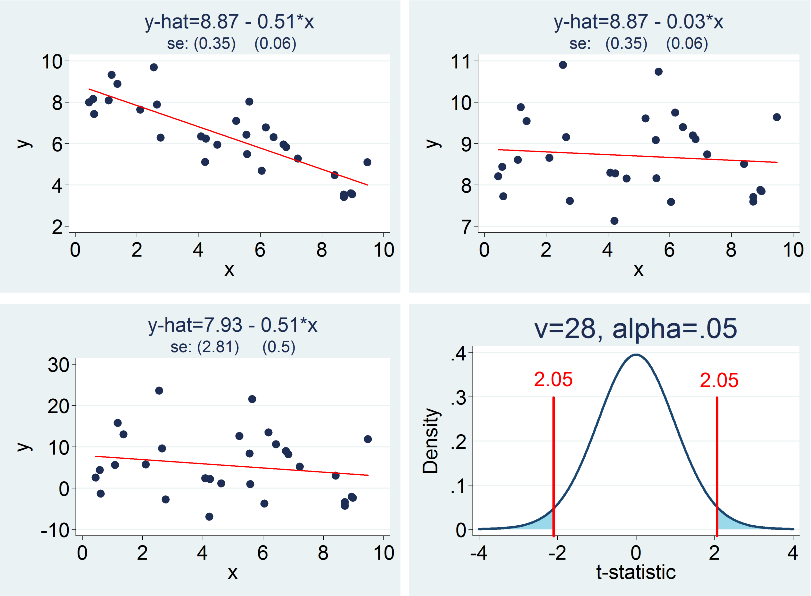
Test of statistical significance for the slope coefficient:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

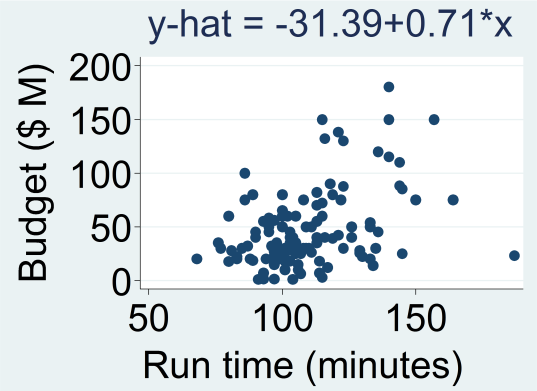
t-statistic: $\boxed{t_{(\nu=n-2)} = \frac{b_1}{s_{b_1}}}$

Can use rejection region approach or p-value approach

Rejection Region Approach



Example: Movie Budget



Standard error of the slope estimate is 0.154

Sample size, n = 120

Can we infer that the movie budget and the running time are linearly related?

P-value

P-value is probability of obtaining a slope estimate as extreme as the one estimated in the direction of H_1 if H_0 is true.

In general,

- for two tailed test, p-value is equal to

$$P[t < -\frac{b_1 - \beta_1^0}{s_{b_1}}, t > \frac{b_1 - \beta_1^0}{s_{b_1}} | H_0 \text{ is true}]$$

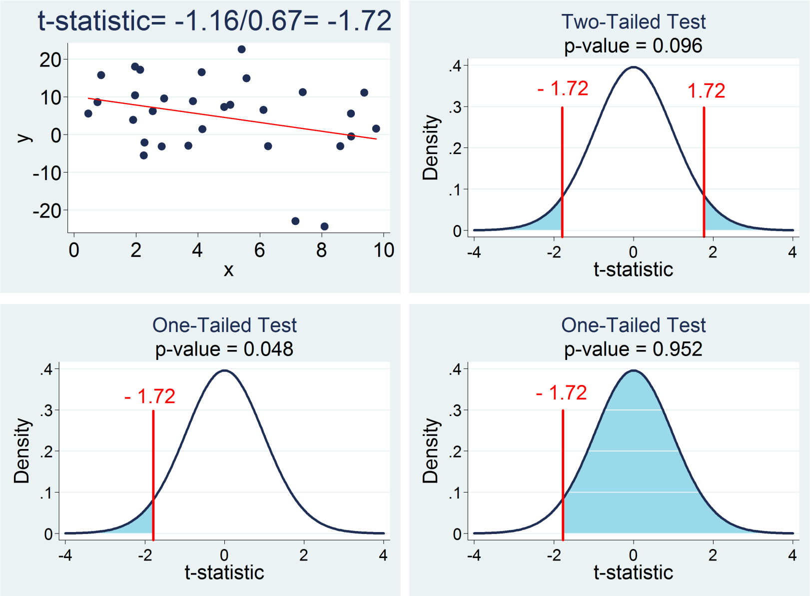
- For one-tailed left-sided test p-value is equal to

$$P[t < \frac{b_1 - \beta_1^0}{s_{b_1}} | H_0 \text{ is true}]$$

- For one-tailed right-sided test p-value is equal to

$$P[t > \frac{b_1 - \beta_1^0}{s_{b_1}} | H_0 \text{ is true}]$$

P-Value Approach



Movie Example

- Set up hypotheses:
- Compute test-statistic:
- Find the value of t-stats in t-table
- Do not forget about the degrees of freedom!
- Conclusion?

Statistical vs Economic Significance

- Statistically Significant: The degree of correlation between explanatory and dependent variables that is not likely observed due to mere chance. The statistical significance of a variable is entirely determined by the size of t_{b_1}. Statistical significance improves as sample size increases. Why?
- **Statistical significance** does not automatically imply that you have found something **important.**
- Economically Significant: An effect large enough in magnitude that decision makers would consider it important. The economical importance is related to the magnitude and sign of b_1 and indicates whether the explanatory variable has a meaningful and plausible influence on dependent variable.

Statistical vs Economic Significance – Example

Consider hypothetical regression of test scores on class size and assume we have found:

$$\text{Testscore\_hat} = \underset{(15.45)}{675} - \underset{(1.12)}{5} * \text{ClassSize}$$

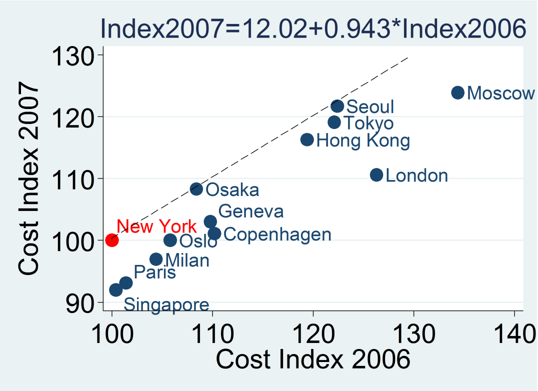
Is b_1 statistically significant? Can you give an answer just by eyeballing?

Interpretation: Consider average test score of 650. Reduction in class size by 1 student is associated with improvement in average test score by 5 points, or only 0.8%! Is this result economically significant?

β_1^0 does not have to be 0!

- Can test for other values of β_1, not only 0!
- For instance, $H_0 : \beta_1 = 3$ vs. $H_A : \beta_1 > 3$
- The same procedure as for the standard t-test for statistical significance
  1. Set the hypotheses
  2. Compute t-statistic: $t_{n-2} = \frac{b_1 - \beta_1^0}{se_{b_1}}$
  3. Use p-value or rejection region approach
- Example: Cost of living index

Example: Cost of Living Index



Sample size: 15

Standard error of the slope estimate: 0.115

Do we have enough evidence to infer that $\beta_1 \neq 1$?

Confidence Interval Estimator of β_1

Confidence interval (CI) estimator for β_1:

$$b_1 \pm t_{(\alpha/2, n-2)} se_{b_1}$$

Confidence level: $1 - \alpha$

Degrees of freedom: $\nu = n - 2$

Stata Output - Read It!

```
regress income education

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  113.82
       Model |  28364.7755     1  28364.7755           Prob > F      =  0.0000
    Residual |  36881.8178   148  249.201472           R-squared     =  0.4347
-------------+------------------------------           Adj R-squared =  0.4309
       Total |  65246.5933   149    437.8966           Root MSE      =  15.786

------------------------------------------------------------------------------
   income    |      Coef.   Std. Err.      t     P>|t|      [95% Conf. nterval]
------------------------------------------------------------------------------
education    |  4.136793   .3877479    10.67   0.000      3.370556     4.90303
    _cons    |  23.63131   5.268046     4.49   0.000      13.22101    34.04162
------------------------------------------------------------------------------
```

Red box – coefficients' estimates

Blue box – standard errors of coefficients

Purple box – Red box / Blue box = t-statistic

Green box – p-value

Orange box - 95% confidence interval of the slope estimate

```
regress income education

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  113.82
       Model |  28364.7755     1  28364.7755           Prob > F      =  0.0000
    Residual |  36881.8178   148  249.201472           R-squared     =  0.4347
-------------+------------------------------           Adj R-squared =  0.4309
       Total |  65246.5933   149     437.8             Root MSE      =  15.786

------------------------------------------------------------------------------
   income    |      Coef.   Std. Err.      t    P>|t|     [95% Conf. nterval]
------------------------------------------------------------------------------
education    |   4.136793   .3877479    10.67   0.000     3.370556    4.90303
    _cons    |   23.63131   5.268046     4.49   0.000     13.22101   34.04162
------------------------------------------------------------------------------
```

$\boxed{\text{SST}} = \boxed{\text{SSR}} + \boxed{\text{SSE}}$

$\boxed{\text{MSE}} = \boxed{\text{SSE}} / \boxed{\text{df}}$, df=n-2

$\boxed{\text{Root MSE}}$= Square root of $\boxed{\text{MSE}}$ =s.e. of estimate

Stata output provides estimates of the coefficients, results of the tests for significance, confidence intervals, and reports the measure of fit - R-squared.

Interpretation of the Results

- Slope coefficient
  - significance - t-stat or *p*-value
    - is slope different from zero?
    - is slope different form hypothesized value?
  - interpretation - observational or experimental data
    - observational data - slope has descriptive interpretation
    - experimental data - slope has casual interpretation [not always]
- R-squared
  - How much variation in dependent variable is explained by independent variable?
  - How large is the standard error of estimate?

Drawing Valid Conclusion

Let's take a look at income-education regression again:

```
regress income education

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  113.82
       Model |  28364.7755     1  28364.7755           Prob > F      =  0.0000
    Residual |  36881.8178   148  249.201472           R-squared     =  0.4347
-------------+------------------------------           Adj R-squared =  0.4309
       Total |  65246.5933   149     437.8             Root MSE      =  15.786

---------------------------------------------------------------------------
   income    |      Coef.   Std. Err.      t     P>|t|    [95% Conf. interval]
---------------------------------------------------------------------------
education    |  4.136793   .3877479    10.67   0.000    3.370556    4.90303
    _cons    |  23.63131   5.268046     4.49   0.000    13.22101   34.04162
---------------------------------------------------------------------------
```

Is slope statistically significant? How much variation in income is due to education? Can we conclude that we are 95\% confident that one more year of education produces increase in income from 3.4 thousands to 5 thousands?

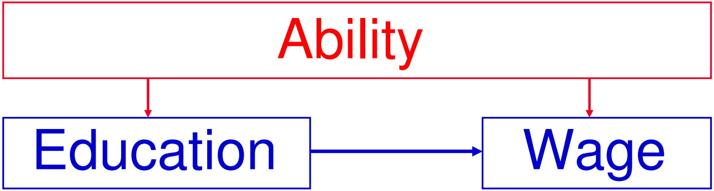
Biased Estimate

- Definition: Bias $\Rightarrow E[b_1] \neq \beta_1$.
- In a simple linear regression model, when we have only one explanatory/independent variable x, we can sign the direction of the bias ("+" or "–").
- Sign of the bias depends on the covariance between independent variable (x) and omitted variable (error term ε).
- As a result, if $COV(x_i, \varepsilon_i) > 0$, then $E[b_1] > \beta_1$.
- If $COV(x_i, \varepsilon_i) < 0$, then $E[b_1] < \beta_1$.
- But: we also need to think about the relationship between dependent variable and omitted variable.[not in this course]

OK

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Assumption # (?) is violated



Which arrow will be missing with experimental data?

Example

Consider again a simple linear regression model of income and education:

$$Income_i = \beta_0 + \beta_1 * Education_i + \varepsilon_i$$

Is b_1 biased? What is the direction of bias in this case?

Assume that we have only one omitted variable - Ability.

$$\text{COV(Education, Ability)} > 0$$

$$\Rightarrow \text{E}[b_1] > \beta_1$$

Intuition behind this result?

More Examples

Try to determine the direction/sign of bias in the following cases:

- Training $\Rightarrow$ employment?
- Lecture attendance $\Rightarrow$ test scores?
- Number of children $\Rightarrow$ hours of work?

Direction of Bias - Big Picture [Aside]

	$Cov(x_i, \varepsilon_i) > 0$	$Cov(x_i, \varepsilon_i) < 0$
$Cov(y_i, \varepsilon_i) > 0$	"+" bias, or OLS estimate b_1 of β_1 will be on average larger than the true value of population parameter	"–" bias, or OLS estimate of β_1 will be on average smaller than the true value of population parameter
$Cov(y_i, \varepsilon_i) < 0$	"–" bias	"+" bias

Here, you have to think of $Cov(y_i, \varepsilon_i)$ as direction of the relationship between dependent variable y and omitted variable which is a part of the error term ε. The correct mathematical expression for the sign of the bias is:

$$Sign\left[Bias(b_1)\right] = Sign\left[\beta_z * \rho(x_1, z)\right]$$

where z is an omitted variable and β_z is the OLS coefficient from regression of y on z: $y_i = \alpha + \beta_z * z_i + \upsilon_i$