The Distribution of the Slope Coefficient

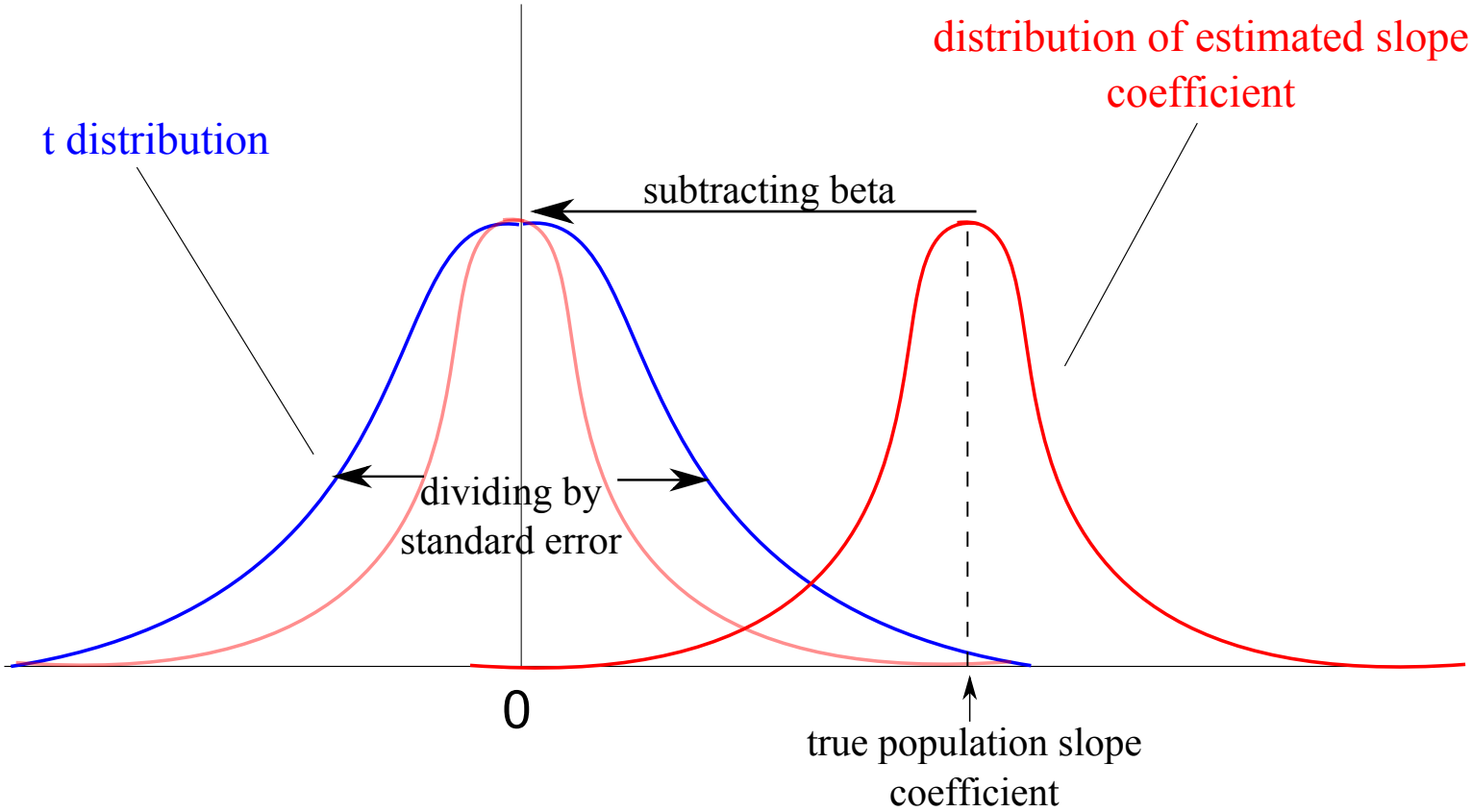
Given our population assumptions:

- $E(b_2) = \beta_2$ (so b_2 is an unbiased estimator)
- The standard error of b_2 is:

$$s_{b_2} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- The test statistic $t^* = \frac{b_2 - \beta_2}{s_{b_2}}$ is t distributed with $(n-2)$ degrees of freedom

The Distribution of the Slope Coefficient



The Distribution of the Intercept

Given our population assumptions:

- $E(b_1) = \beta_1$ (so b_1 is an unbiased estimator)
- The standard error of b_1 is:

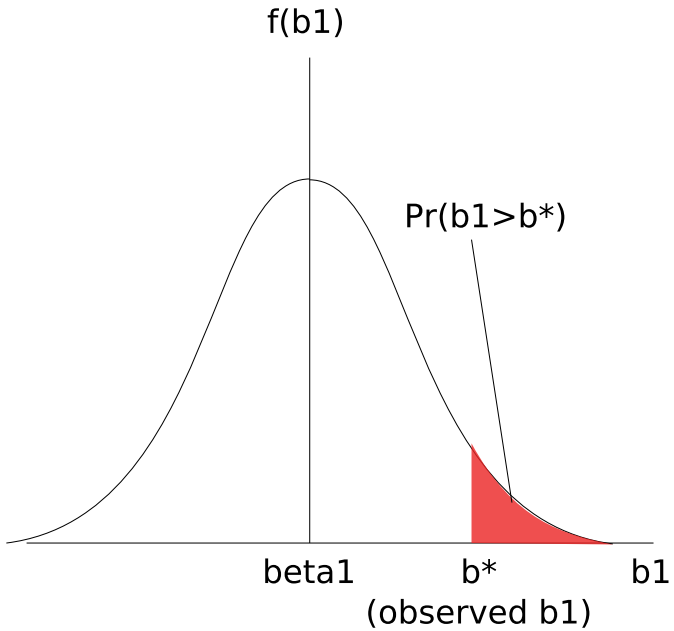
$$s_{b_1} = \sqrt{\frac{s_e^2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

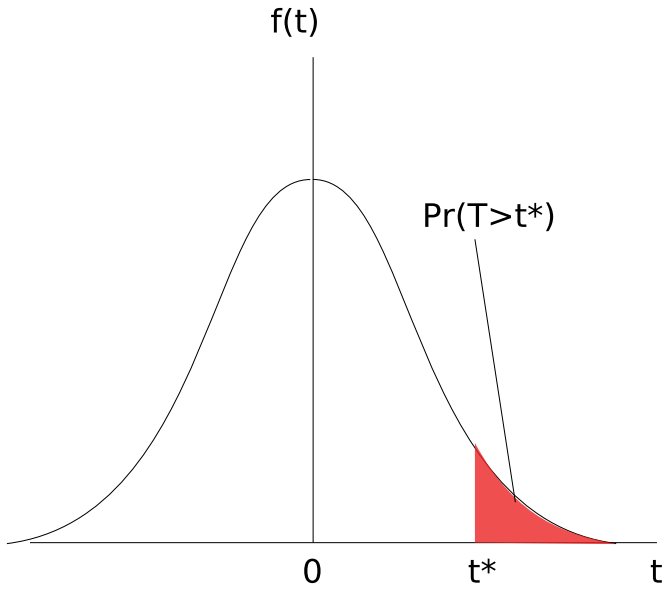
- The test statistic $t^* = \frac{b_1 - \beta_1}{s_{b_1}}$ is t distributed with $(n-2)$ degrees of freedom

Bivariate Hypothesis Testing

- The basic idea behind our hypothesis testing is the same as with univariate data
- We will choose a value for β_1 (or β_2)
- Based on the distribution of b_1 (or b_2), we will determine the probability of observing our b_1 (or b_2) if the true value of the population coefficient is what we guessed
- If this probability is high, we don't reject our initial guess
- It this probability is very low, we reject our intitial guess in favor of whatever we have specified as the alternative

Bivariate Hypothesis Testing





Bivariate Hypothesis Testing

- Once we have calculated our test statistic, everything works the same as it did for univariate hypothesis testing
- The basic steps:
  1. Formulate the null hypothesis (say $\beta_2 = \beta_2^*$) and alternative hypothesis and choose a significance level
  2. Calculate b_2 and s_{b_2}
  3. Use the values of b_2 and s_{b_2} to calculate t^*
  4. Either compare t^* to your critical value or calculate the p-value and compare to α
- The one difference from univariate hypothesis testing is that we are now using $(n-2)$ degrees of freedom (important when you are calculating critical values or p-values)

Bivariate Hypothesis Testing

For a two-tailed test:

- Setting up the hypothesis:

$$H_o: \beta_2 = \beta_2^*$$
$$H_a: \beta_2 \neq \beta_2^*$$

- The test statistic:

$$t^* = \frac{b_2 - \beta_2^*}{s_{b_2}}$$

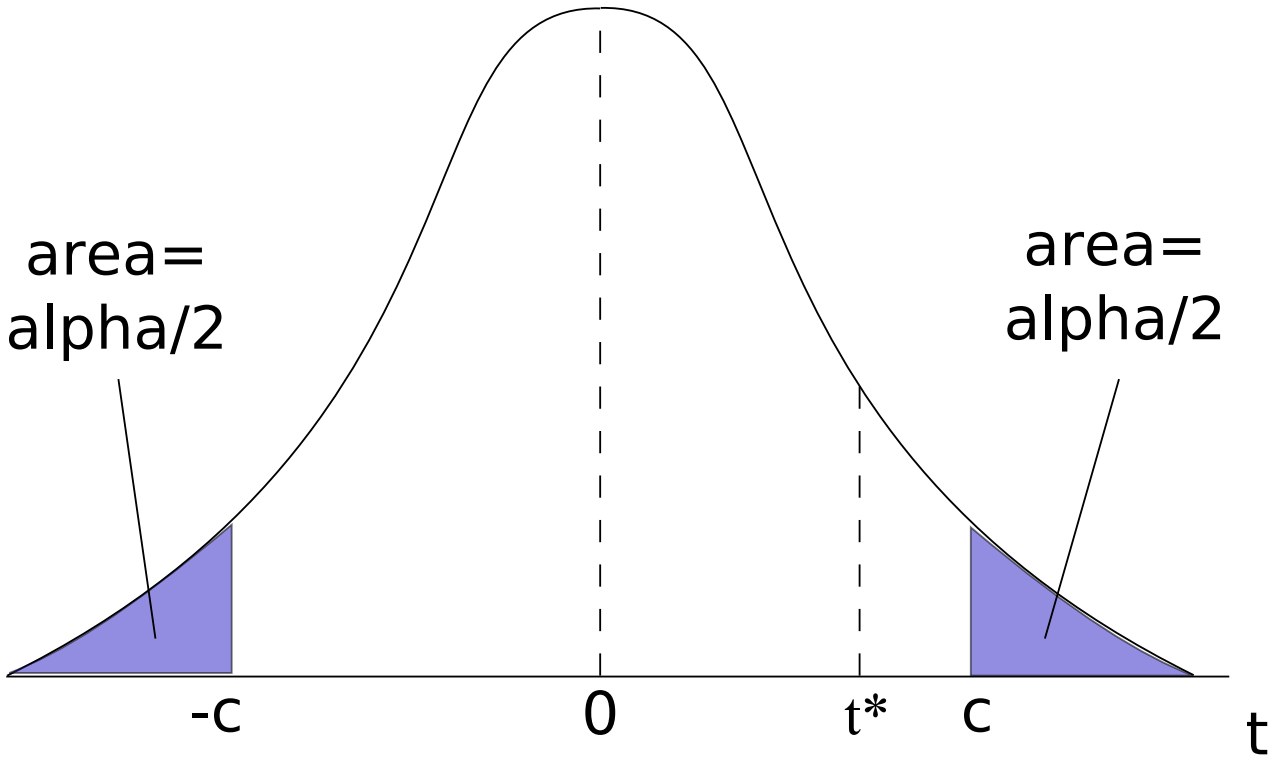
- Evaluating the test statistic

$$p = Prob(|T_{n-2}| \geq |t^*|)$$

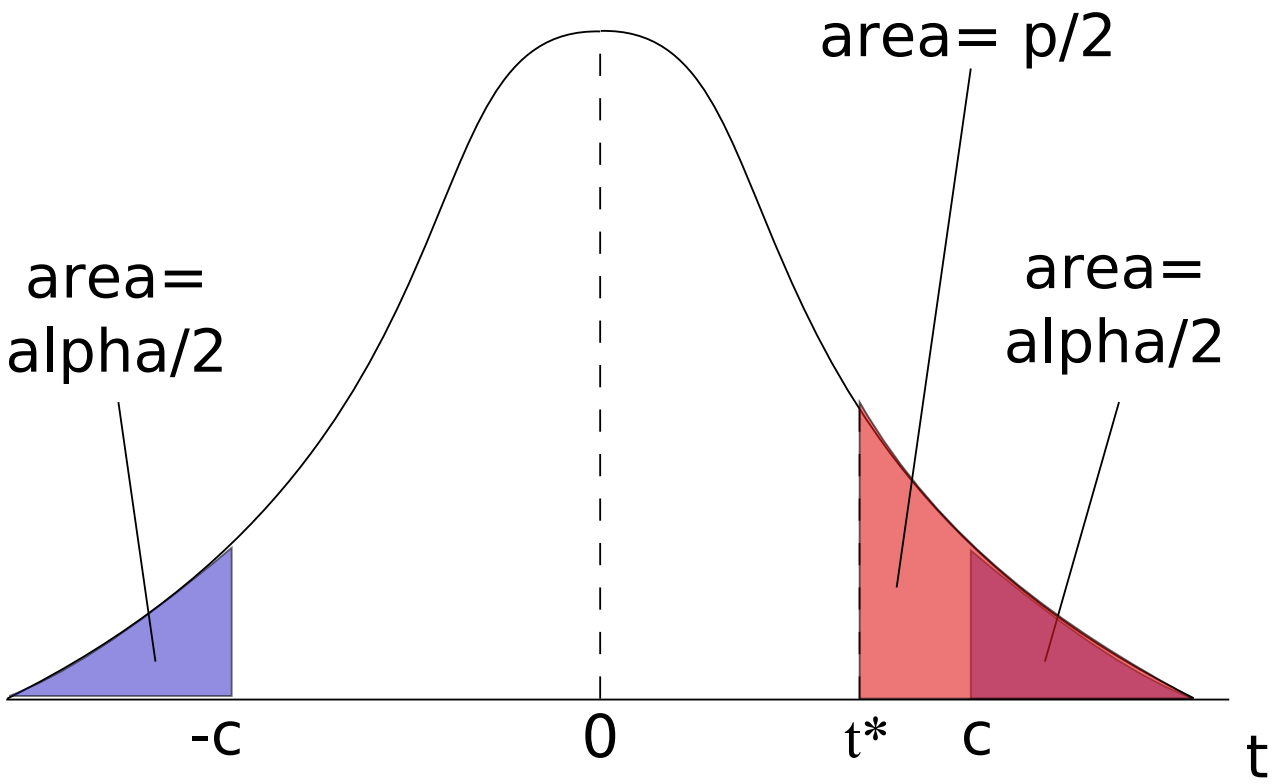
$$c = t_{\frac{\alpha}{2},n-2}$$

- Reject H_o if $p < \alpha$ or if $|t^*| > c$

Bivariate Hypothesis Testing



Bivariate Hypothesis Testing



Bivariate Hypothesis Testing

For an upper one-tailed test:

- Setting up the hypothesis:

$$H_o: \beta_2 \leq \beta_2^*$$
$$H_a: \beta_2 > \beta_2^*$$

- The test statistic:

$$t^* = \frac{b_2 - \beta_2^*}{s_{b_2}}$$

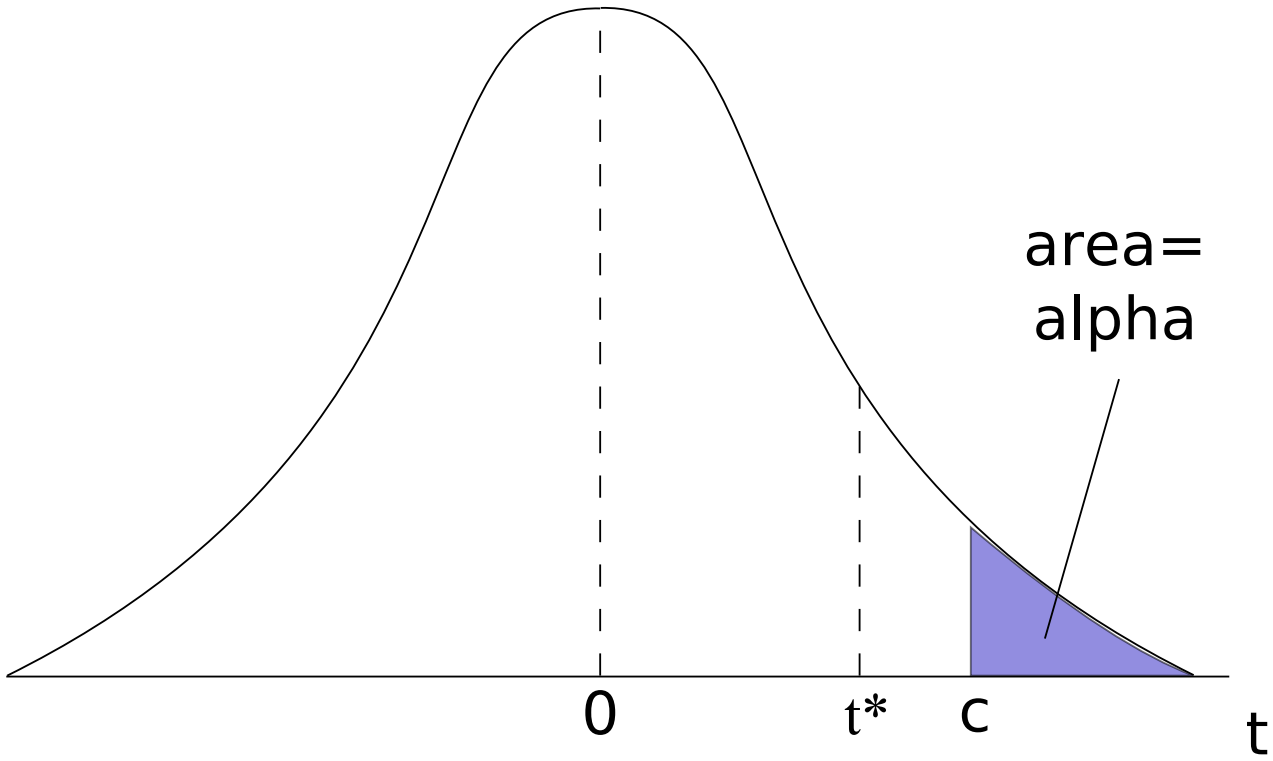
- Evaluating the test statistic

$$p = Prob(T_{n-2} \geq t^*)$$

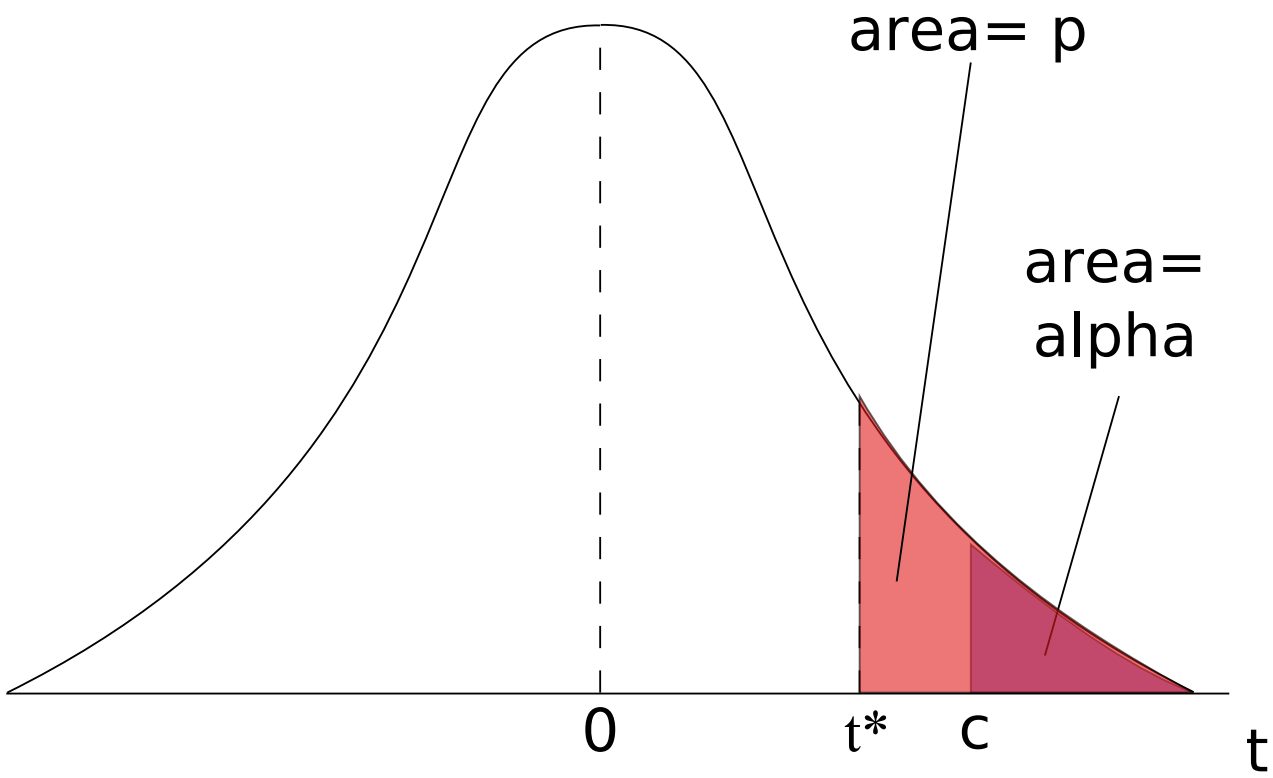
$$c = t_{\alpha,n-2}$$

- Reject H_o if $p < \alpha$ or if $t^* > c$

Bivariate Hypothesis Testing



Bivariate Hypothesis Testing



Bivariate Hypothesis Testing

For a lower one-tailed test:

- Setting up the hypothesis:

$$H_o: \beta_2 \geq \beta_2^*$$
$$H_a: \beta_2 < \beta_2^*$$

- The test statistic:

$$t^* = \frac{b_2 - \beta_2^*}{s_{b_2}}$$

- Evaluating the test statistic

$$p = Prob(T_{n-2} \leq t^*)$$

$$c = -t_{\alpha,n-2}$$

- Reject H_o if $p < \alpha$ or if $t^* < c$

Excel Commands for Hypothesis Testing

- p-value for a two-tailed test: $TDIST(|t^*|, n-2, 2)$
- p-value for an upper one-tailed test (assuming $t^* > 0$): $TDIST(t^*, n-2, 1)$
- p-value for a lower one-tailed test (assuming $t^* < 0$): $TDIST(-t^*, n-2, 1)$
- critical value for a two-tailed test: $TINV(\alpha, n-2)$
- critical value for an upper one-tailed test: $TINV(2\alpha, n-2)$
- critical value for a lower one-tailed test: $-TINV(2\alpha, n-2)$

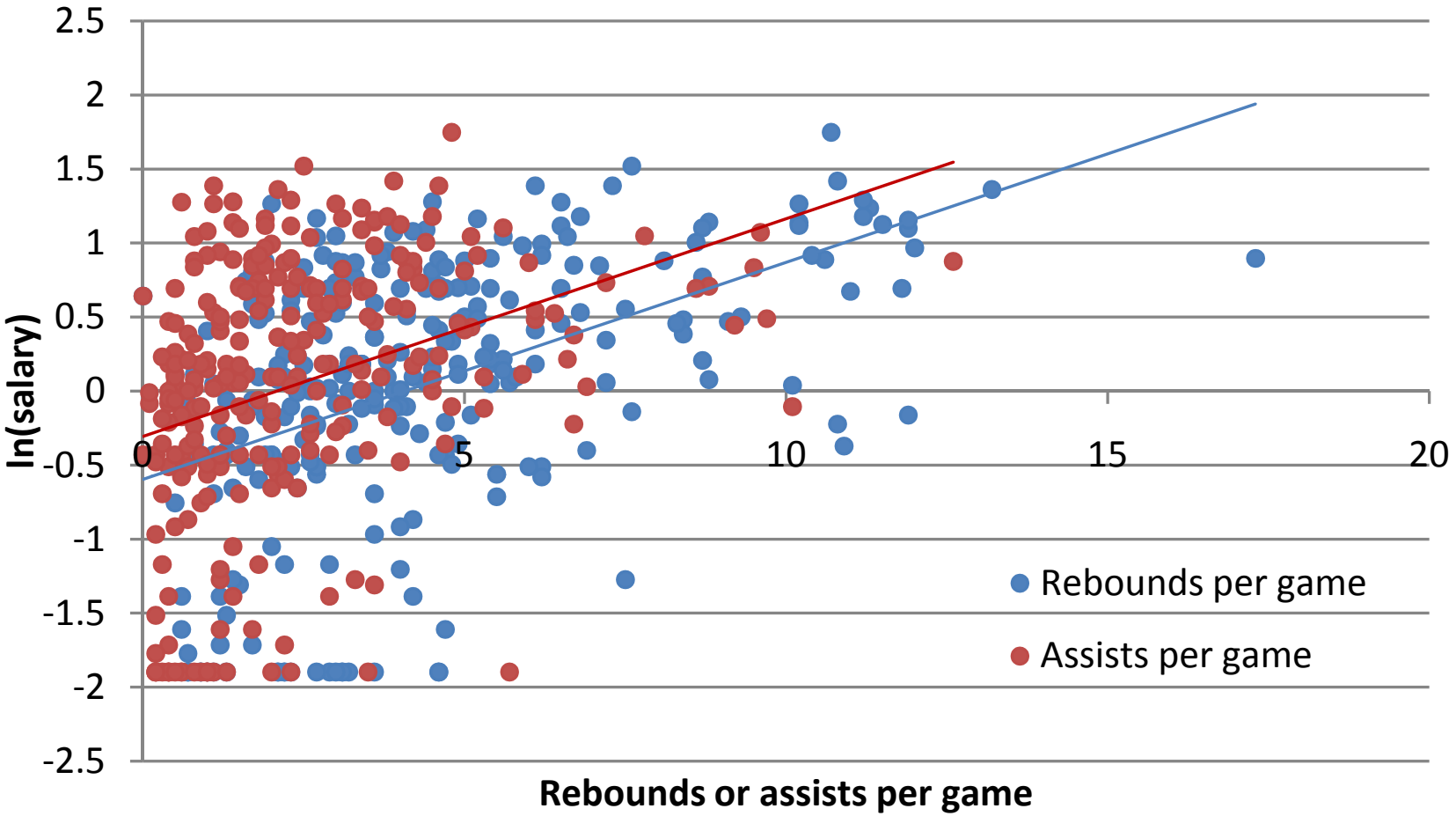
Hypothesis Testing Example

- Let's go back to our NBA salary data and try out bivariate hypothesis testing
- It turns out that when ln(salary) is regressed on assists per game, the slope coefficient is .147, so an extra assist per game is associated with a 14.7% increase in salary
- Suppose we want to know whether an extra rebound per game is less valuable than an extra assist per game
- The hypothesis we want to test is whether an extra rebound per game is associated with an increase in salary of less than 14.7%:

$$H_0: \beta_2 \geq 14.7$$
$$H_a: \beta_2 < 14.7$$

- Let's go to Excel to test this (nba-data.csv) ...

Rebounds per Game and Salary



Testing the Value of Rebounds

$$H_0: \beta_2 \geq 14.7$$
$$H_a: \beta_2 < 14.7$$

- From Excel, b_2 was 0.146474 and s_{b_2} was 0.016257

$$t^* = \frac{0.146474 - .147}{0.016257} = -0.03235$$

$$p = TDIST(0.03235, 270 - 2, 1) = 0.487$$

- So we fail to reject the null hypothesis that the value of a rebound is greater than or equal to the value of an assist

When to Use a Two-Tailed Test

- The most common time to use a two-tailed test is when we want to know whether x has any statistically significant relationship with y
- In this case, we are testing the following hypotheses:

$$H_o: \beta_2 = 0$$
$$H_a: \beta_2 \neq 0$$

- This is the test that the t-stats and p values given in Excel's regression output correspond to

When to Use a Two-Tailed Test

- There are situtations that call for a two-sided test and a value for β_2^* other than zero
- Consider a regression of GDP on government spending, we would want to know whether an extra dollar spent by the government leads to simply an extra dollar of GDP or some other amount
- Another case would be estimating whether demand is unit elastic by regressing log of demand on the log of the price

When to Use a One-Tailed Test

- One-sided tests are useful when we want to test for the sign of a coefficient or when we want to test whether a coefficient is greater than an economically important cutoff
- Consider the government spending example again, we may want to test for a multiplier effect (so we would test whether the coefficient on government spending was greater than 1)
- When testing for a sign of a coefficient, people often use a two-tailed test to see if there is any significant relationship and then just check the sign of the coefficient (be careful about interpreting p-values)

Confidence Intervals for Regression Coefficients

- Our hypothesis testing techniques let us test whether the coefficient is equal to a particular value
- Another useful way to do statistical inference is to construct a confidence interval for the coefficient
- Recall confidence intervals from univariate inference: the population mean fell within the $(1-\alpha)$ confidence interval with a probability of $(1-\alpha)\%$
- The confidence interval for the slope coefficient is justing telling us the range of values β_2 is likely to be in

Confidence Intervals for Regression Coefficients

- We calculate the $(1-\alpha)$ confidence interval as:

$$b_2 \pm t_{\frac{\alpha}{2},n-2} \cdot s_{b_2}$$

- Excel will give the 95% confidence interval in the regression output...for other confidence intervals, you need to do the calculation yourself
- The confidence interval will be centered around our estimated slope coefficient
- The interval will be wider if the standard error is larger
- The interval will be wider if we choose a smaller α

An Example: Drinking and Obesity

- Suppose we are interested in whether drinking is associated with obesity (think of beer guts)
- First, we need to decide which data is most useful to our research question
- We could regress weight on amount of drinking but this has some problems
- A lot of people weigh more because they are taller, not because they are overweight
- An alternative is to use the body mass index (bmi):

$$bmi = \frac{weight}{height^2} \cdot 703$$

- To Excel (alcohol-bmi.csv) ...

An Example: Drinking and Obesity

- The coefficient on days of alcohol consumption is very statistically significant (the p-value was 0.0084)
- We had a fairly narrow confidence interval which means we have a good idea of how big the true coefficient is
- But all of this information still doesn't tell us whether we should really care
- Given our regression results, let's predict the bmi's for a person who doesn't drink and a person who drinks every day:

$$\widehat{bmi}(0) = 27.69 - .25 \cdot 0 = 27.69$$

$$\widehat{bmi}(7) = 27.69 - .25 \cdot 7 = 25.94$$

- Is this an important difference?

An Example: Drinking and Obesity

BMI cutoff values	
<16.5	severely underweight
16.5 to 18.5	underweight
18.5 to 25	normal
25 to 30	overweight
30 to 35	obese I
35 to 40	obese II
40 to 45	severely obese
45 to 50	morbidly obese
50 to 60	super obese
>60	hyper obese

An Example: Drinking and Obesity

- So the association between drinking and bmi is very statistically significant
- But the size of the coefficient actually isn't all that impressive (and there are issues with interpreting what it means)
- An extra day of drinking a week is associated with a very small change in bmi relative to our various bmi cutoffs
- This is a case where we say that the coefficient is *statistically significant* but not *economically significant*

Economic vs. Statistical Significance

- Statistical significance is just telling us whether a coefficient is different than zero (or whatever we chose as β_2^*)
- This doesn't mean we should care about the coefficient
- We also need to think about whether the magnitude of coefficient is large
- When we consider whether the magnitude is large enough to be an important effect, we are consider economic significance
- We don't have any formal tests for economic significance, it is left to our judgement

Economic Significance and the Confidence Interval

- We can have a coefficient that is statistically significant but that has a confidence interval too wide to make conclusions about economic significance
- For example, suppose that the coefficient on education when ln(wage) is regressed on years of education has the following 95% confidence interval:

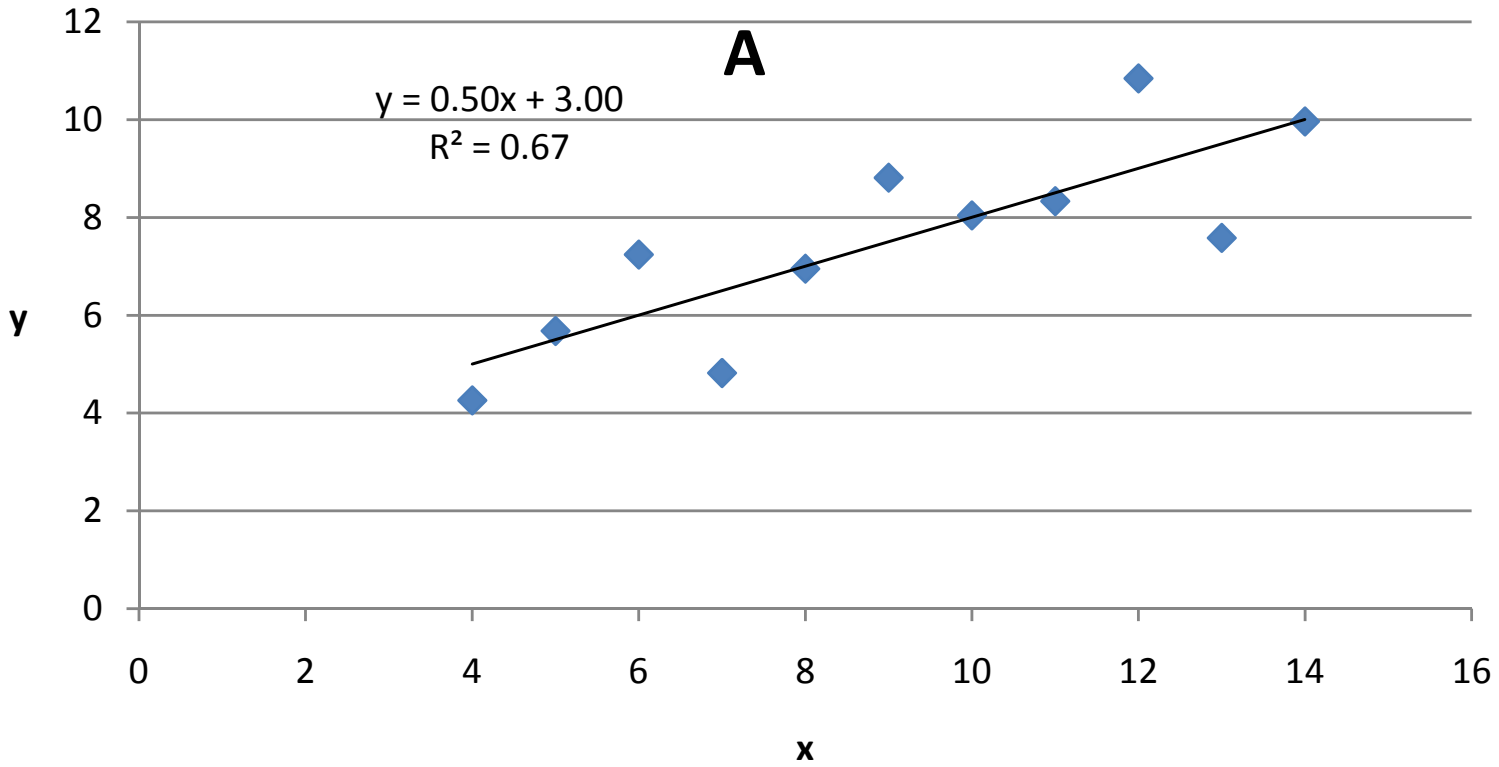
$$\beta_{edu} = .05 \pm .04$$

- At the lower end of this interval, the effect of education on wage is not all that impressive while at the other end it is quite large
- We can say that education is statistically significant but to say whether it is really important would require narrowing the confidence interval (say by using better data and more observations)

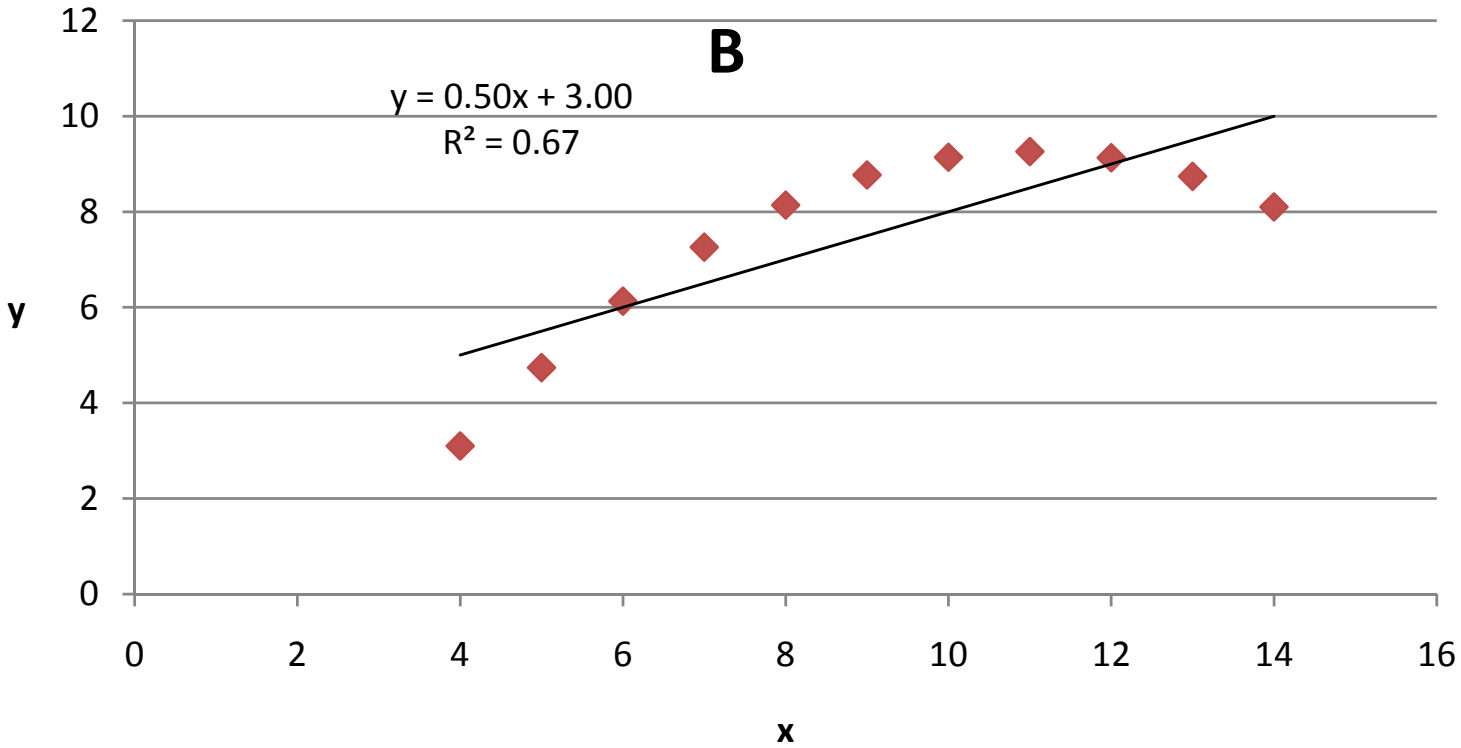
A Cautionary Note

- A regression still doesn't tell you everything about a bivariate relationship
- We've already talked about how a basic regression does not establish causality
- Beyond that, there are many other features of data that won't be apparent in regression results
- We'll take a look at just how important this can be with a set of data called the Anscombe's Quartet (anscombe-example.xlsx)

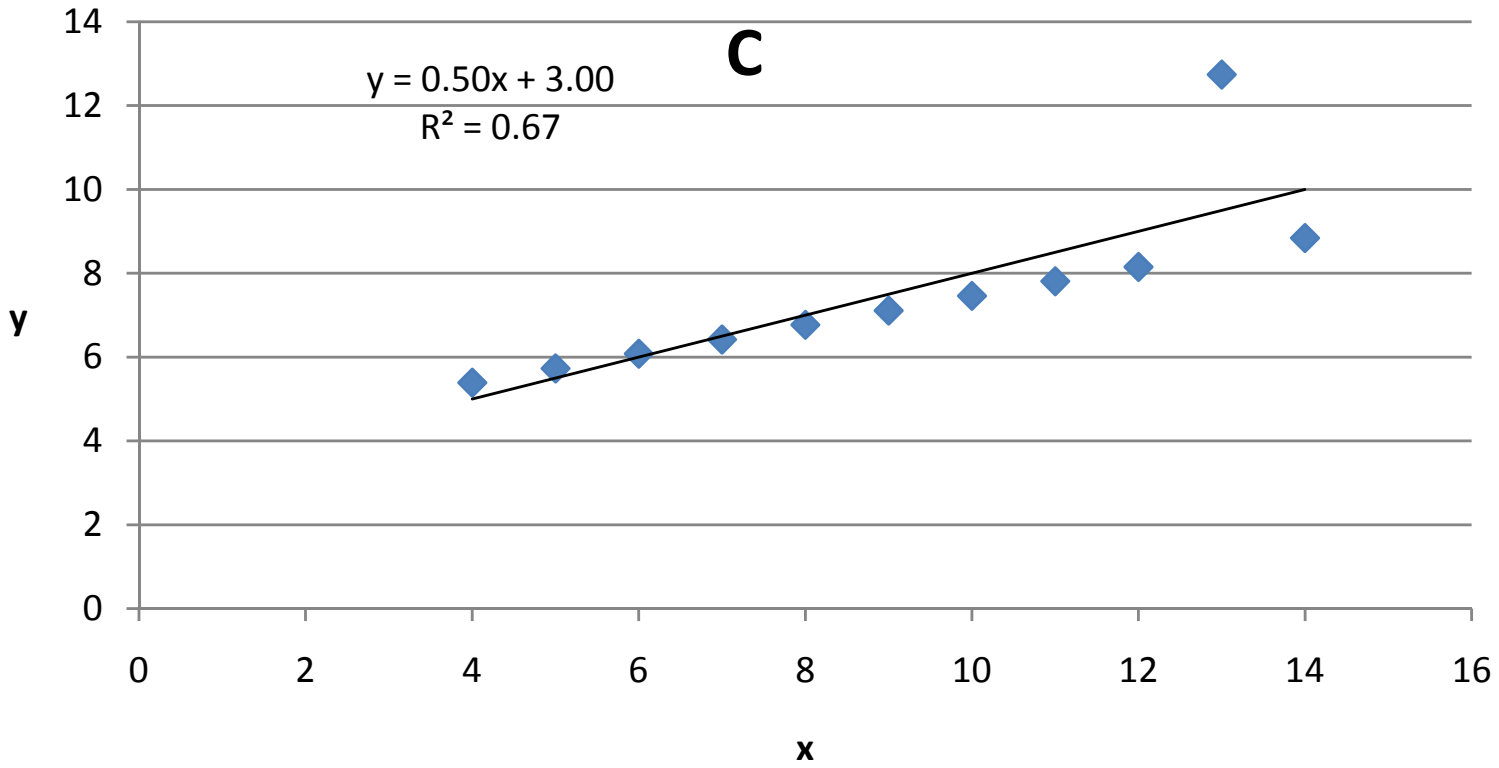
The Anscombe Quartet



The Anscombe Quartet



The Anscombe Quartet



The Anscombe Quartet