Linear Regression

AGEC 317: Economic Analysis for Agribusiness Management

Instructor: Michael Black

TEXAS A&M UNIVERSITY
Agricultural
Economics

A tale of two variables

If we have two variables, we may be interested in the relationship between them:

- Amount of time spent studying and grade on an exam
- Moon phase and Aggie football score

A tale of two variables

We can calculate the correlation and covariance between the variables, but correlation doesn't tell us anything about the *magnitude* of the relationship, and covariance is sensitive to the variance of each variable.

Instead of coming up with a single number describing the relationship, let's build a **model** that helps us define the relationship as an equation.

Linear regression

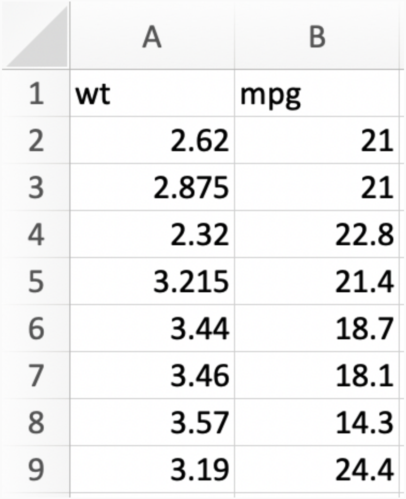
Linear regression is a way to represent the relationship between two variables using a straight line:

$$Grade = \beta_0 + \beta_1(StudyTime)$$

(Nothing different between this and the classic "$y = mx + b$". It is just a line.)

What is linear regression?

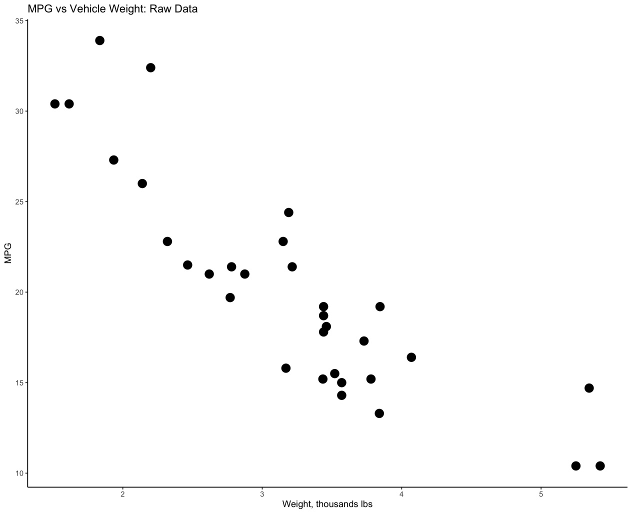
We begin with raw data:



	A	B
1	wt	mpg
2	2.62	21
3	2.875	21
4	2.32	22.8
5	3.215	21.4
6	3.44	18.7
7	3.46	18.1
8	3.57	14.3
9	3.19	24.4

What is linear regression?

Then we make a scatterplot:



What is linear regression?

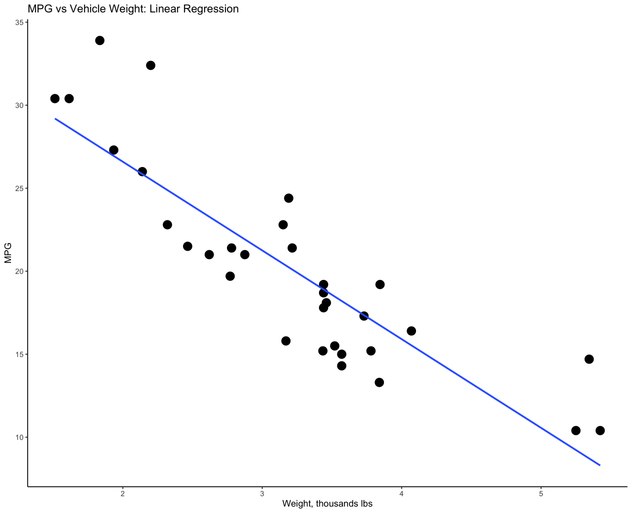
A single observation, in this case, is a car. The car has a weight and an MPG. However, not all cars weigh the same and have the same MPG; there is much variety on the market. We are interested in the *relationship* between the two variables. That is, for an increase in weight of the car, how much does MPG typically change? The answer is surprisingly simple:

- Look at raw data
- Draw a line
- Slope of line = relationship

No math needed! All you need is an eye (or two), a pencil, and a scatterplot of the data.

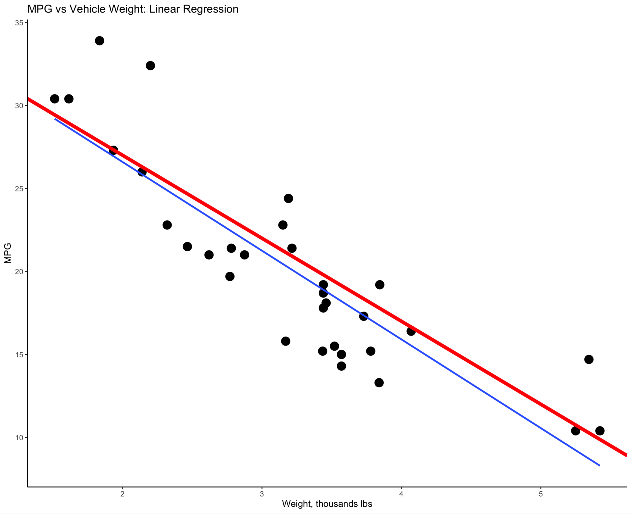
What is linear regression?

Look! A line!



What is linear regression?

But wait! I think the red line is better!



What is linear regression?

We can argue all day about the red line vs the blue line, but the *correct* answer is whichever line uses the following math:

1. Draw a line
2. Calculate the distance from each point to the line (call it a "residual")
3. Square each of those residuals, and add all the residuals together
4. Repeat steps 1 - 3 until you have found the line that **minimizes the sum of squared residuals**

This approach take a while, but you just performed **ordinary least squares**. Least squares = minimized sum of squared residual.

What is linear regression?

So a linear regression is the estimation of a line that best fits our data, and that line describes the relationship between the variables. Some notes:

- Correlation measures the association between variables, not the relationship. That is, correlation can tell us *something* about the relationship, but not everything. If the correlation coefficient is negative, the regression slope should also be negative. But the regression gives us the magnitude of the slope, while correlation cannot.
- Regression, in general, is used to investigate the effect of an independent variable, X, on a dependent variable, Y.

What is linear regression?

Yet again, economists have multiple names for the same thing:

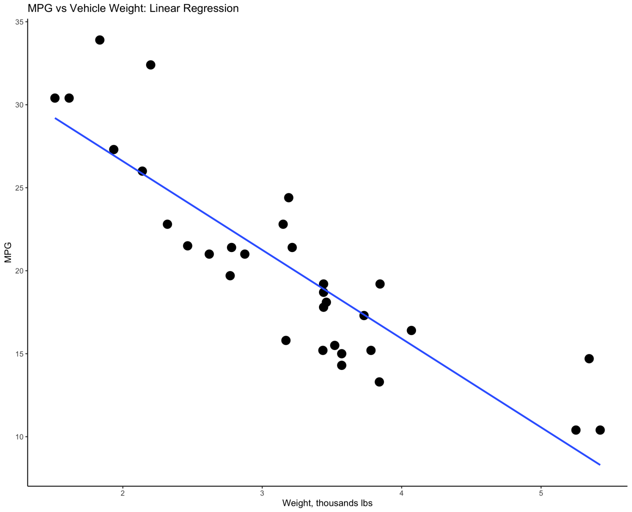
- Y = f(X)
- **Dependent variable = f(Independent variable)**
- Regressand = f(Regressor)
- LHS = f(RHS)

What is linear regression?

Suppose you work in sales for Biarte Coffee (a coffee grower/roaster founded by a recent graduate of our MAB program!). What are some variable pairs you could use regression to explore? What kind of data would you need?

Linear regression model

Recall the line we drew earlier



Linear regression model

A regression model between Y (MPG) and X (weight) is:

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

where y_i is the MPG of some car i, x_i is the weight of the car, β_0 is the intercept, β_1 is the slope, and u_i is the distance from the observation to the line. We **observe** y_i and x_i, **estimate** β_0 and β_1, and **calculate** u_i.

β_0 and β_1 are the *structural* parts of the model, and u_i is the *random* component.

Linear regression model

When we perform a linear regression, it is important to note that u_i (the error or residual) captures *everything else* about i. Is MPG a function of car weight *and* engine size? If we don't observe engine size, then that effect is captured by the error term. It is the catch-all for our function.

If there is something important hidden in the catch-all, then our simple model is not good. We'll come back to how to recognize this problem, known as omitted variable bias.

Linear regression model

Recall our first try at ordinary least squares (OLS):

1. Draw a line
2. Calculate the distance from each point to the line (call it a "residual")
3. Square each of those residuals, and add all the residuals together
4. Repeat steps 1 - 3 until you have found the line that **minimizes the sum of squared residuals**

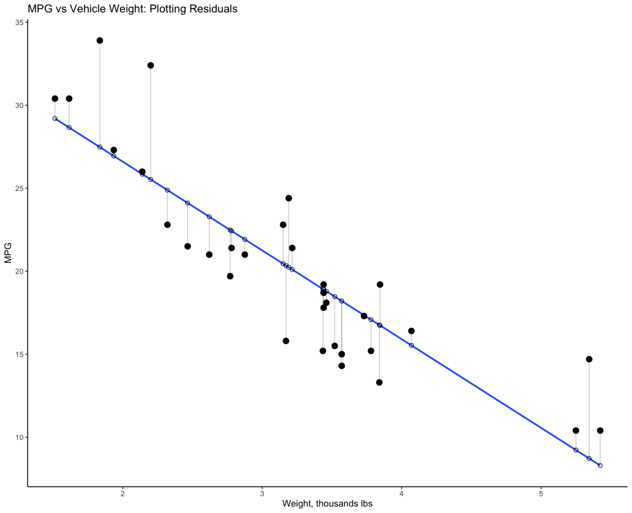
Linear regression model

The steps again, but using our model terminology:

1. Draw a line ⇒ Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ (the "hat" represents a guess)
2. Calculate the distance from each point to the line (call it a "residual") ⇒ Calculate $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
3. Square each of those residuals, and add all the residuals together ⇒ Calculate $\sum_i^N \hat{u}_i{}^2$
4. Repeat steps 1 - 3 until you have found the line that **minimizes the sum of squared residuals**

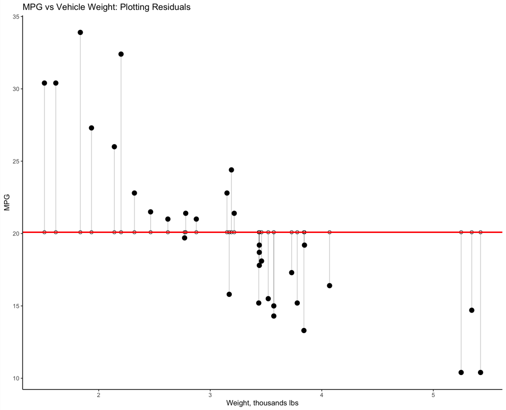
Note that $\hat{u}_i$ is a residual, and u_i is an error. We calculate the residuals, and thus that's what we will work with.

Ordinary least squares (OLS)

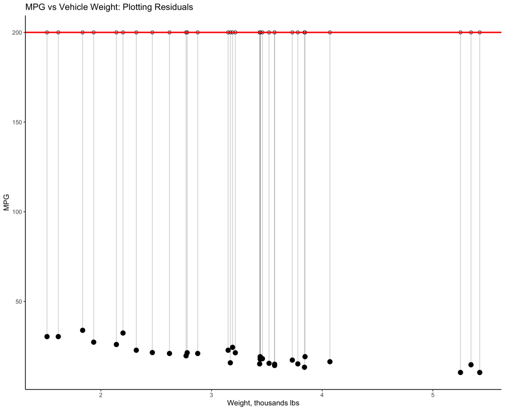


Ordinary least squares (OLS)

So why are we squaring the error term? The error is naturally positive for points above our line, and negative for points below the line.



Ordinary least squares (OLS)



The total error here would be very negative. More negative than the previous horizontal line.

Ordinary least squares (OLS)

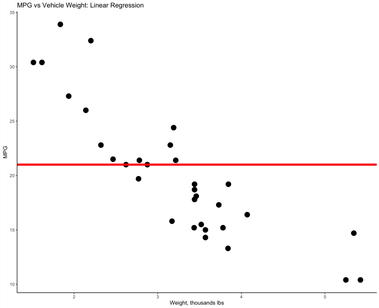
Something is obviously wrong. Our lines do not describe the data well at all. The problem is that negative errors are good and positive errors are bad when we minimize. The solution is to make all errors positive, which we can do by:

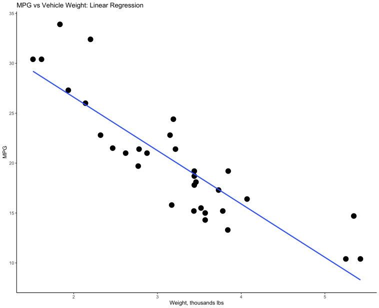
- Taking the absolute value: $|\hat{u}_i|$
- Squaring each error: ${\hat{u}_i}^2$

Both would work, but if we square each error, we get the added bonus that large errors become even larger and are even more undesirable than before we square.

Ordinary least squares (OLS)

Once we square the errors, and add, we see that the right is much much better than the left:





Ordinary least squares (OLS)

Remember the true model is:

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

But we never observe the true model. We **estimate** it as:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + u_i$$

and we can use the model to predict y_i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and finally calculate the residuals:

$$\hat{u}_i = y_i - \hat{y}_i$$

Let's look at the excel sheet associated with this lecture to see what these numbers look like.

Ordinary least squares (OLS)

Quick vocab:

- Total sum of squares (SST): $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum of squared errors (SSE): $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Sum of squared residuals (SSR): $\sum_{i=1}^{n}(\hat{u}_i)^2$

$$SST = SSE + SSR$$

Ordinary least squares (OLS)

Remember the best regression line is the one that minimizes the **sum of squared residuals (SSR)**

Ordinary least squares (OLS)

Okay. Now time to stop guessing, and find the best line directly.

Ordinary least squares (OLS)

We want to minimize the sum of squared-residuals:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\hat{u}_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

How do we minimize a function?

First we take the FOC wrt β_0:

$$
\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
\Rightarrow \sum_{i=1}^{n} (\beta_0) = \sum_{i=1}^{n} (y_i) - \sum_{i=1}^{n} (\hat{\beta}_1 x_i) \\
\Rightarrow N\beta_0 = \sum_{i=1}^{n} (y_i) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i) \\
\Rightarrow \beta_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i) - \frac{1}{n} \hat{\beta}_1 \sum_{i=1}^{n} (x_i) \\
\Rightarrow \beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}
$$

Now we take the FOC wrt β_1:

$$\begin{aligned}
&\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \\
&\Rightarrow \sum_{i=1}^{n} (y_i x_i) - \hat{\beta}_0 \sum_{i=1}^{n} (x_i) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = 0 \\
&\Rightarrow \sum_{i=1}^{n} (y_i x_i) - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} (x_i) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = 0 \\
&\Rightarrow \sum_{i=1}^{n} (y_i x_i) - \bar{y} \sum_{i=1}^{n} (x_i) - \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} (x_i) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = 0
\end{aligned}$$

Continuing:

$$\begin{aligned}
\Rightarrow \sum_{i=1}^{n}(y_i x_i) - \bar{y}\sum_{i=1}^{n}(x_i) - \hat{\beta}_1 \bar{x}\sum_{i=1}^{n}(x_i) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i)^2 = 0 \\
\Rightarrow \hat{\beta}_1 \left[\sum_{i=1}^{n}(x_i)^2 - \bar{x}\sum_{i=1}^{n}(x_i)\right] = \sum_{i=1}^{n}(y_i x_i) - \bar{y}\sum_{i=1}^{n}(x_i) \\
\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i x_i) - \bar{y}\sum_{i=1}^{n}(x_i)}{\sum_{i=1}^{n}(x_i)^2 - \bar{x}\sum_{i=1}^{n}(x_i)} \\
\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

Recall that:

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Does our derivation of β_1 look familiar now?

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ordinary least squares (OLS)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(X, Y)}{var(X)}$$

Ordinary least squares (OLS)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Notice that our estimates for the model come from random variables (X and Y). Thus, the coefficient estimates are random values themselves. So even though we get a single number (point estimate), that estimate is technically a random variable, and thus has variance and standard deviation. We can imagine, then, that estimates with small variances are good (accurate), and estimates with large variances are bad (inaccurate).

Ordinary least squares (OLS)

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$var(\hat{\beta}_0) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where

$$\sigma^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (\hat{u}_i{}^2)$$

Ordinary least squares (OLS)

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$se(\hat{\beta}_0) = \sqrt{var(\hat{\beta}_0)} = \sqrt{\frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where

$$\sigma^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (\hat{u_i}^2)$$

Ordinary least squares (OLS)

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$se(\hat{\beta}_0) = \sqrt{var(\hat{\beta}_0)} = \sqrt{\frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where

$$\sigma^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (\hat{u}_i{}^2)$$

We want **accurate** estimates. Thus, we want estimates with small standard errors. What helps make $se(\hat{\beta}_1)$ small?

Ordinary least squares (OLS)

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

What helps make $se(\hat{\beta}_1)$ small? Well, we could make the numerator tiny, the denominator huge, or both! To make the numerator tiny, we want small SSR. To make the denominator huge, we want lots of variation in our explanatory variables!

Ordinary least squares (OLS)

Performing OLS guarantees that we have the smallest SSR possible, so we can't make the numerator any smaller than it is. Making the denominator big is the greatest desire of economists/data scientists: variation in the data is very, very good.

Hypothesis testing in OLS

So we have two variables, and we estimate a simple model that results in $\hat{\beta}_0$ and $\hat{\beta}_1$. How do we know if our model good? Even if we have minimized the SSR, is the model *helpful*? These are tricky questions. What *is* a good model? What do we mean by good?

Hypothesis testing in OLS

Before drowning in questions, let's start with a simple one:

- Is $\hat{\beta}_0$ or $\hat{\beta}_1$ significantly different from zero?

Since both $\hat{\beta}_0$ and $\hat{\beta}_1$ are **random variables**, we shouldn't be deceived by the point-estimates; we need to see if the **average** values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are different from zero.

The t-test

Hmm... testing the means of random variables. Sound familiar?

We'll use the **t-test**!

The t-test

Let's focus on $\hat{\beta}_1$. Instead of comparing the means of two random variables, we are comparing $\hat{\beta}_1$ to β_1.
In a regression format, we are going to have the following hypotheses:

- $H_0 : \hat{\beta}_1 = \beta_1$
- $H_1 : \hat{\beta}_1 \neq \beta_1$

The t-test

But wait! It gets even *more* boring and depressing! Our assumption is the model we are estimating **doesn't even exist**. So without any evidence about the true population model, we assume that $\beta_1 = 0$.

- $H_0 : \hat{\beta}_1 = 0$
- $H_1 : \hat{\beta}_1 \neq 0$

*Same for any β in the model.

The t-test

Our first random variable to perform the t-test is $\hat{\beta}_1$. The second random variable is $\beta_1 = 0$. Wait, is zero a random variable?
No.
It is a scalar with a sample size of one, and a variance of zero. But shhh, don't tell the t-test.

The t-test

Our first random variable to perform the t-test is $\hat{\beta}_1$. The second random variable is $\beta_1 = 0$. We assume (correctly) that these two "random variables" have unequal variances and unequal sample sizes. Recall that the t-test for this situation is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The t-test

Let's start substituting:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{0}{1}}}$$

$$t = \frac{\hat{\beta}_1 - 0}{\frac{s_1}{\sqrt{n_1}}}$$

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

The t-test

Because our hypotheses are:

- $H_0 : \hat{\beta}_1 = 0$
- $H_1 : \hat{\beta}_1 \neq 0$

Our t-test is:

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

The degrees of freedom of the test is $n - k - 1$ where k is the number of explanatory variables and n is the number of observations.

Exercise

Determine the degrees of freedom:

- 400 observations: $y = \beta_0 + \beta_1 x$
- 20 observations: $y = \beta_0 + \beta_1 x$

t-distribution

Your Excel regression estimate give you t-statistics.

$$t_{stat} = \frac{(\hat{\beta}_1 - \beta_1)}{se(\hat{\beta}_1)}$$

To determine significance, we want to see where our t-stat lands on the t-distribution.

Find t-critical value

The t_{crit}, or t-critical value, depends on the t-distribution and thus the degrees of freedom and α. You can use published tables to find the t-critical value, or Excel:

- Two-tailed: "=TINV(α, df)"
- One-tailed: "=TINV($2 \cdot \alpha$, df)"

Make decision

Two-tailed test:

- If $|t_{stat}| > t_{crit}$: Reject null
- Otherwise: Fail to reject null

Make decision

One-tailed test ($H_1 : \beta_1 > 0$):

- If $t_{stat} > t_{crit}$: Reject null
- Otherwise: Fail to reject null

One-tailed test ($H_1 : \beta_1 < 0$):

- If $t_{stat} < -t_{crit}$: Reject null
- Otherwise: Fail to reject null

p-value revisited

The p-value of a coefficient is technically:

Two-tailed: $pvalue = Pr(|t_{stat}| > t_{crit})$

One-tailed, H_1 positive: $pvalue = Pr(t_{stat} > t_{crit})$

One-tailed, H_1 negative: $pvalue = Pr(t_{stat} < -t_{crit})$

We can calculate the p-value in Excel:

“=TDIST(ABS(tstat),df,tails)”

p-value revisited

- If the p-value $< \alpha$, **reject the null**
- If the p-value $> \alpha$, **fail to reject the null**

Confidence Intervals

Once we choose α and calculate the t-critical value, we can construct a confidence interval:

$$\bar{\beta}_1 = \hat{\beta}_1 + (t_{crit} \cdot se(\hat{\beta}_1))$$
$$\underline{\beta}_1 = \hat{\beta}_1 - (t_{crit} \cdot se(\hat{\beta}_1))$$

$\Rightarrow$ "(1-α)% of the data lies between $\underline{\beta}_1$ and $\bar{\beta}_1$"

Decisions

Your hypothesis testing decisions should come from the t-test (and associated p-values), and the confidence interval should back this decision up.

Example

Let’s work through an example in Excel...

Key Skills

In this lecture, we have discussed basic linear regression. At this point, you should be able to:

- Describe what linear regression is
- Derive the slope and intercept for a simple linear regression **by hand**.
- Be able to perform a simple linear regression in Excel
- Be able to determine which variables are statistically significant