Regression: Crash Course Statistics #32

Crash Course: Statistics
https://youtube.com/watch?v=WWqE7YHR4Jc
https://nerdfighteria.info/v/WWqE7YHR4Jc

Hi, I'm Adriene Hill and welcome back to Crash Course Statistics.

There's something to be said for flexibility. It allows you to adapt to new circumstances.

Like a Transformer is a truck, but it can also be an awesome fighting robot. Today we'll introduce you to one of the most flexible statistical tools--the General Linear Model, or GLM. The GLM will allow us to create many different models to help describe the world.

The first we'll talk about is The Regression Model. INTRO General Linear Models say that your data can be explained by two things: your model, and some error: First, the model. It usually takes the form Y = mx + b, or rather, Y = b + mx in most cases.

Say I want to predict the number of trick-or-treaters I'll get this Halloween by using enrollment numbers from the local middle school. I have to make sure I have enough candy on hand. I expect a baseline of 25 trick-or-treaters.

And then for every middle school student, I'll increase the number of trick-or-treaters I expect by 0.01. So this would be my model: There were about 1,000 middle school students nearby last year, so based on my model, I predicted that I'd get 35 trick-or-treaters. But reality doesn't always match predictions.

When Halloween came around, I got 42, which means that the error in this case was 7. Now, error doesn't mean that something's WRONG, per se. We call it error because it's a deviation from our model.

So the data isn't wrong, the model is. And these errors can come from many sources: like variables we didn't account for in our model-- including the candy-crazed kindergartners from the elementary school--or just random variation Models allow us to make inferences --whether it's the number of kids on my doorstep at Halloween, or the number of credit card frauds committed in a year. General Linear Models take the information that data give us and portion it out into two major parts: information that can be accounted for by our model, and information that can't be.

There's many types of GLMS, one is Linear Regression. Which can also provide a prediction for our data. But instead of predicting our data using a categorical variable like we do in a t-test, we use a continuous one.

For example, we can predict the number of likes a trending YouTube video gets based on the number of comments that it has. Here, the number of comments would be our input variable and the number of likes our output variable. Our model will look something like this: The first thing we want to do is plot our data from 100 videos: This allows us to check whether we think that the data is best fit by a straight line, and look for outliers--those are points that are really extreme compared to the rest of our data.

These two points look pretty far away from our data. So we need to decide how to handle them. We covered outliers in a previous episode, and the same rules apply here.

We're trying to catch data that doesn't belong. Since we can't always tell when that happened, we set a criteria for what an outlier is, and stick to it. One reason that we're concerned with outliers in regression is that values that are really far away from the rest of our data can have an undue influence on the regression line.

Without this extreme point, our line would look like this. But with it, like this. That's a lot of difference for one little point!

There's a lot of different ways to decide, but in this case we're gonna leave them in. One of the assumptions that we make when using linear regression, is that the relationship is linear. So if there's some other shape our data takes, we may want to look into some other models.

This plot looks linear, so we'll go ahead and fit our regression model. Usually a computer is going to do this part for us, but we want to show you how this line fits. A regression line is the straight line that's as close as possible to all the data points at once.

That means that it's the one straight line that minimizes the sum of the squared distance of each point to the line. The blue line is our regression line. Its equation looks like this: This number--the y-intercept--tells us how many likes we'd expect a trending video with zero comments to have.

Often, the intercept might not make much sense. In this model, it's possible that you could have a video with 0 comments, but a video with 0 comments and 9104 likes does seem to conflict with our experience on YouTube. The slope, aka, the coefficient--tells us how much our likes are determined by the number of comments.

Our coefficient here is about 6.5, which means that on average, an increase in 1 comment is associated with an increase of about 6.5 likes. But There's another part of the General Linear Model: the error. Before we go any further, let's take a look at these errors--also called residuals.

The residual plot looks like this: And we can tell a lot by looking at its shape. We want a pretty evenly spaced cloud of residuals. Ideally, we don't want them to be extreme in some areas and close to 0 in others. It's especially concerning if you can see a weird pattern in your residuals like this: Which would indicate that the error of your predictions is dependent on how big your predictor variable value is. That would be like if our YouTube model was pretty accurate at predicting the number of likes for videos with very few comments, but was wildly inaccurate on videos with a lot of comments.

So, now that we've looked at this error, This is where Statistical tests come in. There are actually two common ways to do a Null Hypothesis Significance test on a regression coefficient. Today we'll cover the F-test. The F-test, like the t-test, helps us quantify how well we think our data fit a distribution, like the null distribution. Remember, the general form of many test statistics is this: But I'm going to make one small tweak to the wording of our general formula to help us understand F-tests a little better. The null hypothesis here is that there's NO relationship between the number of comments on a trending YouTube video and the number of likes. IF that were true, we'd expect a kind of blob-y, amorphous-cloud-looking scatter plot and a regression line with a slope of 0. It would mean that the number of comments wouldn't help us predict the number of likes. We'd just predict the mean number of likes no matter how many comments there were.

Back to our actual data. This blue line is our observed model. And the red is the model we'd expect if the null hypothesis were true. Let's add some notation so it's easier to read our formulas. Y-hat looks like this, and it represents the predicted value for our outcome variable--here it's the predicted number of likes. Y-bar looks like this, and it represents the mean value of likes in this sample. Taking the squared difference between each data point and the mean line tells us the total variation in our data set. This might look similar to how we calculated variance, because it is. Variance is just this sum of squared deviations--called the Sum of Squares Total--divided by N. And we want to know how much of that total Variation is accounted for by our regression model, and how much is just error.

That would allow us to follow the General Linear Model framework

and explain our data with two things: the model's prediction, and error. We can look at the difference between our observed slope coefficient--6.468--and the one we'd expect if there were no relationship--0, for each point. And we'll start here with this point: The green line represents the difference between our observed model--which is the blue line--and the model that would occur if the null were true--which is the red line. And we can do this for EVERY point in the data set. We want negative differences and positive differences to count equally, so we square each difference so that they're all positive. Then we add them all up to get part of the numerator of our F-statistic: The numerator has a special name in statistics.

It's called the Sums of Squares for Regression, or SSR for short. Like the name suggests, this is the sum of the squared distances between our regression model and the null model. Now we just need a measure of average variation. We already found a measure of the total variation in our sample data, the Total Sums of Squares. And we calculated the variation that's explained by our model. The other portion of the variation should then represent the error, the variation of data points around our model. Shown here in Orange. The sum of these squared distances are called the Sums of Squares for Error (SSE). If data points are close to the regression line, then our model is pretty good at predicting outcome values like likes on trending YouTube Videos. And so our SSE will be small. If the data are far from the regression line, then our model isn't too good at predicting outcome values. And our SSE is going to be big. Alright, so now we have all the pieces of our puzzle. Total Sums of Squares, Sums of Squares for Regression, and Sums of Squares for Error: Total Sums of Squares represents ALL the information that we have from our Data on YouTube likes. Sums of Squares for Regression represents the proportion of that information that we can explain using the model we created. And Sums of Squares for Error represents the leftover information--the portion of Total Sums of Squares that the model can't explain. So the Total Sums of Squares is the Sum of SSR and SSE.

Now we've followed the General Linear Model framework and taken our data and portioned it into two categories: Regression Model, and Error. And now that we have the SSE, our measurement of error, we can finally start to fill in the Bottom of our F-statistic. But we're not quite done yet. The last and final step to getting our F-statistic is to divide each Sums of Squares by their respective Degrees of freedom. Remember degrees of freedom represent the amount of independent information that we have. The sums of square error has n--the sample size--minus 2 degrees of freedom. We had 100 pieces of independent information from our data, and we used 1 to calculate the y-intercept and 1 to calculate the regression coefficient. So the Sums of Squares for Error has 98 degrees of freedom. The Sums of Squares for Regression has one degree of freedom, because we're using one piece of independent information to estimate our coefficient our slope. We have to divide each sums of squares by its degrees of freedom because we want to weight each one appropriately. More degrees of freedom mean more information. It's like how you wouldn't be surprised that Katie Mack who has a PhD in astrophysics can explain more about the planets than someone taking a high school Physics class. Of course she can she has way more information. Similarly, we want to make sure to scale the Sums of Squares based on the amount of independent information each have. So we're finally left with this: And using an F-distribution, we can find our p-value: the probability that we'd get a F statistic as big or bigger than 59.613. Our p-value is super tiny. It's about 0.000-000-000-000-99. With an alpha level of 0.05, we reject the null that there is NO relationship between likes and YouTube comments on trending videos. So we reject that true coefficient for the relationship between likes and comments on YouTube is 0. The F-statistic allows us to directly compare the amount of variation that our model can and cannot explain. When our model explains a lot of variation, we consider it statistically significant. And it turns out, if we did a t-test on this coefficient, we'd get the exact same p-value.

That's because these two methods of hypothesis testing are equivalent, in fact if you square our t-statistic, you'll get our F-statistic! And we're going to talk more about why F-tests are important later.

Regression is a really useful tool to understand. Scientists, economists, and political scientists use it to make discoveries and communicate those discoveries to the public. Regression can be used to model the relationship between increased taxes on cigarettes and the average number of cigarettes people buy. Or to show the relationship between peak-heart-rate-during-exercise and blood pressure. Not that we're able to use regression alone to determine if it causes changes. But more abstractly, we learned today about the General Linear Model framework. What happens in life can be explained by two things: what we know about how the world works, and error--or deviations--from that model. Like say you budgeted $30 for gas and only ended up needing $28 last week. The reality deviated from your guess and now you get to to go to The Blend Den again! Or just how angry your roommate is that you left dishes in the sink can be explained by how many days you left them out with a little wiggle room for error depending on how your roommate's day was. Alright, thanks for watching, I'll see you next time.