Part 4 | Bivariate GLM

- Part 4.1 | Continuous Predictors
- Part 4.2 | Categorical Predictors
- Part 4.3 | Timeseries
- Part 4.4 | Causality

From Brilliant

- r square
- screenshots 1 - 25

Outline

- In Part 3 we refreshed the ideas for how a t-test works and showed that a t-test is the simplest general linear model.
- The idea in Part 3 is that even with unknown populations, we know the distribution of the sample means, so we can calculate the probability of seeing our sample given any guess we could make.
- Then we showed that this idea of hypothesis testing is a very simple version of a much more powerful model: GLM.
- The GLM is essentially just drawing lines through data with the general form: y = mx + b.
- We choose the line that minimizes how wrong the model is, measured using an idea very similar to variance: MSE.
- We started with the simplest GLM, where there is no x (or x = 0).
- We then saw that in the intercept only model, placing the line at the mean minimizes the sum of squared errors.
- You'll remember that we have sampling error in every sample, so if we redraw a sample of wait time differences a few times we can see that the location of b is different each time.
- We plotted these b's on a histogram and it turned out to be the familar normal distribution around the truth.
- You said you through this was intuitive, since b is equal to the sample mean in the GML.
- This means we know the sampling distribution of model parameter, b, making it possible for us to perform hypothesis tests on the model parameters in the GLM!
- If I were to ask you the probability of getting this sample mean under the standard null of b=0, what would you say?
- This was the big idea from last class: we can perform hypothesis tests on the parameters in the equation for the line.
- You've seen these things before, but I hope you don't miss that this is an amazing thing to be able to do: we're

making probabistic claims about the parameters in a statistical model.

- In general we don't ask many questions about vertical incercepts.
- A more common thing for us to do is to ask whether there's a relationship between two variables.
- At the beginning of the semester we talked about descriptive relationships between variables like this one.
- Higher income countries look to be happier than low income countries.
- Lets draw a line through the data with the form: $y = mx + b$.
- How do we choose what m and b are? We minimize the mean squared error.
- Which of these three looks like it minimizes the mean squared error?
- Nice. Just like with univariate data last time, we have sampling error here too.
- But instead of just variability in the intercept, we have variability in the slope coefficient too.
- Here's a picture of a couple of samples and the slope coefficients of their fitted models.
- If I took all the slope coefficients m and plotted them on a histogram, what do you think that would look like?
- It's a normal distribution again!
- Just like with sample means, sample intercepts, the central limit theorem tells us the slope coefficients will be approach a normal distribution as the sample size gets larger.
- Like before, since we know the sampling distribution of slope parameters, we can ask whether there IS a relationship between any two variables.
- What do you think our null hypothesis would be?
- Right! If there is no relationship between variables then $m = 0$. The parameter being zero is the default null hypothesis.
- Then we can quantify the probability of seeing this slope if there actually is no slope: the p-value.
- We call the fitted value our best guess. They are best in a lot of ways. If I pick anything other than the model parameters, I get a smaller p-value than if I used the sample parameter as my null hypothesis. So everything else is less likely.
- Exercise 4.1 |
- We can use this model to make predictions.
- Lets find the average happiness for medium GDP counties.
- Then lets find the average happiness for slighly higher GDP counties.
- We've increased happiness by β_1 as we increase GDP.
- We interpret β_1 as the change in the y variable as we increase the x variable by 1.
- What if we were to predict the happiness of a county with zero GDP?
- Well we'd get a very strange result. Zero GDP is outside of the domain of the model: no data lives here. So the model has nothing to say here. If we force it to say something, it will say something strange.
- Exercise 4.1 |
- GLM isn't always correct: we need some things to be true for the model to not give us biased results.