Interpret measures of uncertainty in regression

Goal of Reading

In this reading, we will continue exploring uncertainty in regression analysis, specifically through confidence intervals, confidence bands, and p-values. Together, we will:

- Review key concepts
- Discuss how to interpret measures of uncertainty
- Review sample graphs

Review of Concepts

Recall that we can represent a simple linear regression line as $y = \beta_0 + \beta_1 X$.

Since regression analysis utilizes **estimation** techniques, there is always a level of uncertainty surrounding the predictions made by regression models. To represent the error, we can actually rewrite the equation to include an error term, represented by the letter ϵ (pronounced "epsilon"): $y = \beta_0 + \beta_1 X + \epsilon$.

There is one residual, also known as the difference between the predicted and actual value, for each data point in the dataset used to construct the model. We can then quantify how uncertain the entire model is through a few measures of uncertainty:

- **Confidence intervals** around beta coefficients
- **P-values** for the beta coefficients
- **Confidence band** around the regression line

You can refer to the glossary of terms to check any key terms and definitions, but we've provided the two key terms here:

- **Confidence interval:** a range of values that describes the uncertainty surrounding an estimate
- **P-value:** the probability of observing results as extreme as those observed when the null hypothesis is true

Interpreting Uncertainty

Let's first revisit the summary of results from the linear regression model we created together in prior videos:

OLS Regression Results

Dep. Variable:	body_mass_g	**R-squared:**	0.769
Model:	OLS	**Adj. R-squared:**	0.768
Method:	Least Squares	**F-statistic:**	874.3
Date:	Mon, 11 Apr 2022	**Prob (F-statistic):**	1.33e-85
Time:	21:11:50	**Log-Likelihood:**	-1965.8
No. Observations:	265	**AIC:**	3936.
Df Residuals:	263	**BIC:**	3943.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-1707.2919	205.640	-8.302	0.000	-2112.202	-1302.382
bill_length_mm	141.1904	4.775	29.569	0.000	131.788	150.592

Omnibus:	2.060	**Durbin-Watson:**	2.067
Prob(Omnibus):	0.357	**Jarque-Bera (JB):**	2.103
Skew:	0.210	**Prob(JB):**	0.349
Kurtosis:	2.882	**Cond. No.**	357.

According to the simple linear regression model we built, $\hat{\beta}_1$ is 141.1904. So for every one-millimeter increase in the bill length of a penguin, we would expect a penguin to have about 141.1904 more grams in body mass. The estimate has a p-value of 0.000, which is less than 0.05, meaning that the coefficient is "statistically significant." Additionally our estimate has a 95% confidence interval of 131.788 and 150.592. Let's review these short sentences a bit more.

Previously you may have learned about p-values and confidence intervals within the context of hypothesis testing. Even though it may seem unintuitive, even in regression analysis we are testing hypotheses.

P-values

When running regression analysis, you want to know if X is really correlated with y or not. So we do a hypothesis test on the regression results. In regression analysis, for each beta coefficient, we are testing the following set of null and alternative hypotheses:

- H_0 (null hypothesis): $\beta_1 = 0$
- H_1 (alternative hypothesis): $\beta_1 \neq 0$

In our example, because the p-value is less than 0.05, we can reject the null hypothesis that β_1 is equal to 0, and state that the coefficient is statistically significant, which means that a difference in bill length of a penguin is truly correlated with a difference in body mass.

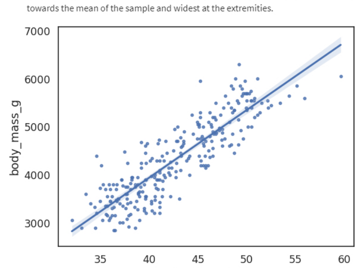
Confidence Intervals

Each beta coefficient also has a confidence interval associated with its estimate. A 95% interval means the interval itself has a 95% chance of containing the true parameter value of the coefficient. So there is 5% chance that our confidence interval [131.788, 150.592] does not contain the true value of β_1. More precisely, this means that if you were to repeat this experiment many times, 95% of the confidence intervals would contain the true value of β_1.

But, since there is uncertainty in both of the estimated beta coefficients, then the estimated y values also have uncertainty. This is where confidence bands become useful.

Example Graph

- **Confidence band:** the area surrounding the line that describes the uncertainty around the predicted outcome. You can think of the confidence band as representing the confidence interval surrounding each point estimate of y. Since there is uncertainty at every point in the line, we use the confidence band to summarize the confidence intervals across the regression model. The confidence band is always narrowest towards the mean of the sample and widest at the extremities.



Key Takeaways

- Regression analysis utilizes **estimation** techniques, so there is always uncertainty around the predictions.
- We can measure uncertainty using confidence intervals, p-values, and confidence bands.
- For every coefficient estimate, we are testing the hypothesis that the coefficient equals 0.

Go to next item ✓ Completed