Inferential Linear Regression

Rosie Martinez

Data Science Immersive

Learning Objectives

By the end of this lecture, you will be able to:

- Understand the differences between predictive and inferential linear regression
- Identify the assumptions of an inferential linear regression model
- Define and detect collinearity between features using VIF
- Understand what confounding is and how it can impact a model's results
- Encode categorical features to use in a regression model
- Interpret the model output for a generalized audience

Predictive Linear Regression

Goal: Accurately predict a target

We picked a model based on trying different features, feature engineering, evaluation **using cross validation**

We care that it predicts well on unseen data

We don't really care that:

- Some features may be partially collinear
- Because of that, we can't rely on our parameter estimates to tell us something about their effect on the signal
- We may be violating some fundamental assumptions of inferential linear regression

Inferential Linear Regression

Goal: Learn something accurate about the process that made the data. Infer (estimate) coefficients.

We picked a model based on trying different features, feature engineering, and checking residuals to see **if we are violating some of the assumptions of linear regression**

We care that the parameter estimates are accurate and valid

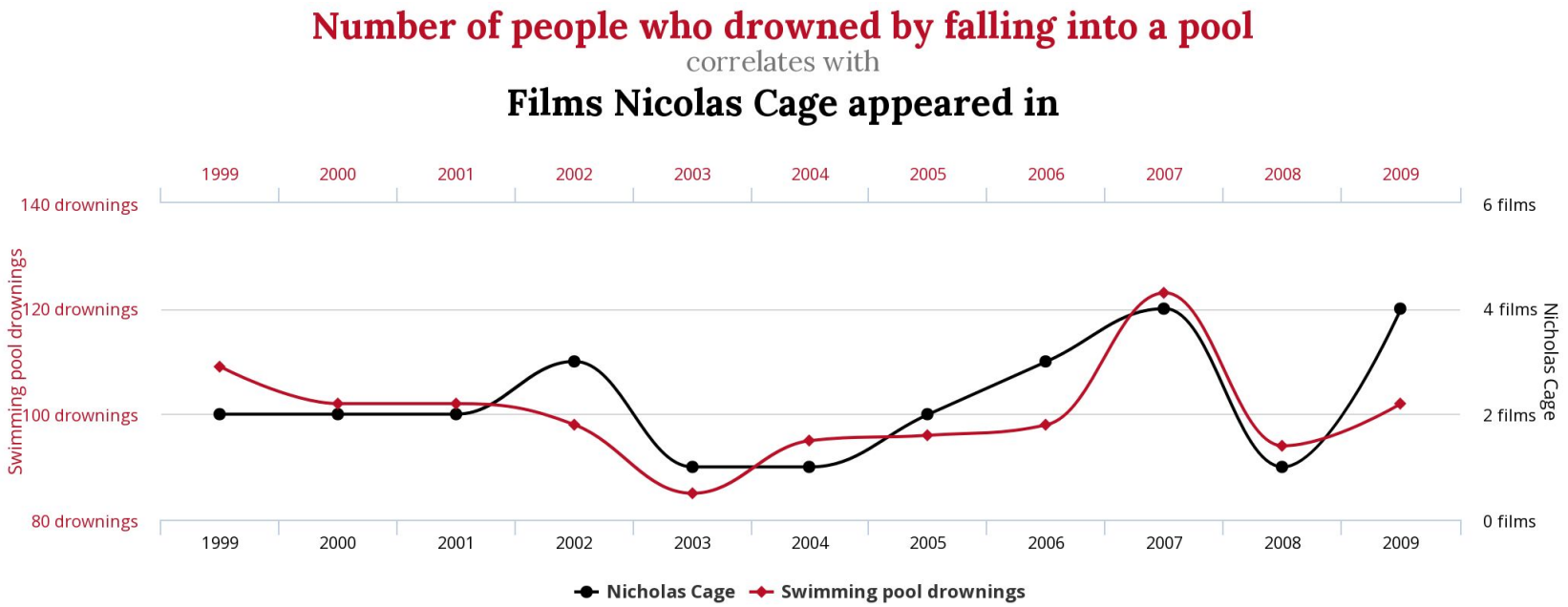
We don't really care that:

- It predicts well (but in theory, it should!)

Inferential Linear Regression

Inferential linear regression is on the way to understanding causality, however, it doesn't mean that we can deduce causality, only correlation or association.

Correlation does not imply causation



Source

So what should you do?

It depends! (The typical data scientist answer)

Most often, we're asked for predictive models, **however**, one major benefit to linear regression is interpretability of the coefficients. If assumptions of inferential linear regression are being violated, then you can't rely on your parameter estimates to provide interpretability.

It is wise to take a hybrid approach: predictive (best model through cross validation) but attempt to fulfill the assumptions of inferential linear regression so that you have confidence in your parameter estimates.

Assumptions of Inferential Linear Regression

1. **Linearity**: the relationship between the X and the y can be modeled linearly
2. **Independence:** the residuals should be independent from each other
3. **Normality:** the residuals are normally distributed
4. **Homoscedasticity:** the variance of the residuals is constant

Note: you might see other assumptions mentioned if you do a quick Google search and we will briefly touch on one of the main ones:

5. **No multicollinearity**: the independent variables are not highly correlated with each other

Assumptions: Linearity

1. **Linearity**: the relationship between the X and the y can be modeled linearly

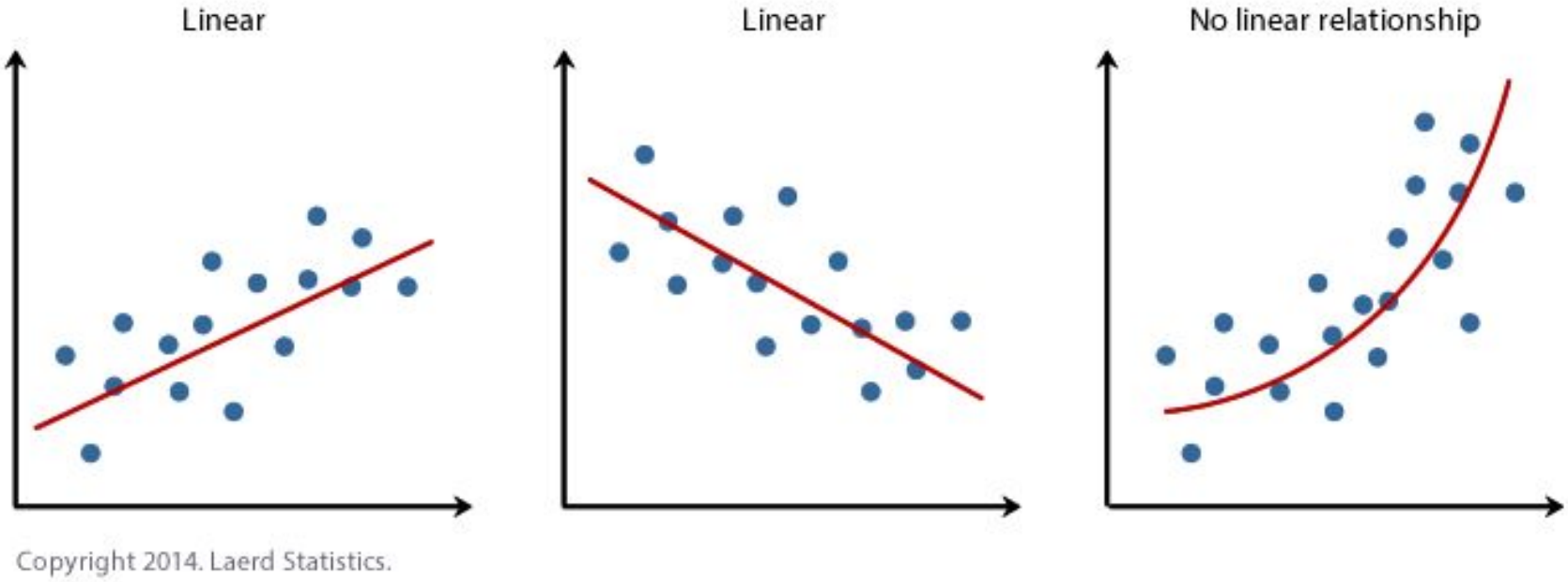
A linear relationship should exist between the independent variable and the dependent variable. Use pair-wise scatterplots to visualize this!

What if the data isn't linear?

- **Transform your data**, typically, log() or and exponentiation of either the dependent or independent variable
- Maybe a linear model isn't the correct model to fit the data, **try a nonlinear regression** (polynomial)

Assumptions: Linearity

1. **Linearity**: the relationship between the X and the y can be modeled linearly

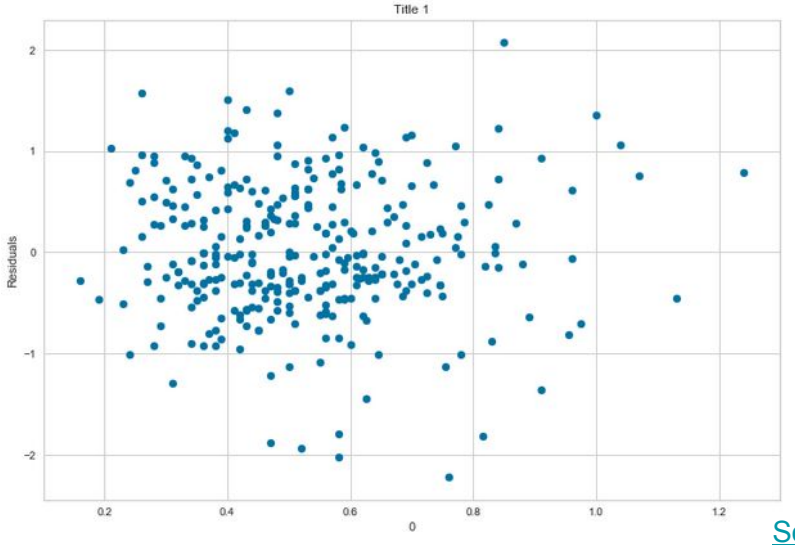


Source

Assumptions: Independence

2. **Independence:** the residuals should be independent from each other

To test for independence, we can use a formal test for this (**Durbin-Watson Test**), but a scatter plot of the residuals with each independent variable will also do the trick. This scatter plot shows how the data is distributed without any specific pattern, thus verifying our assumption



Source

What if there isn't independence?

- This might mean the model isn't linear or that variables have been omitted
- For time-series, one could add a lag variable. Other models, features should be fine-tuned and added to the model

Assumptions: Normality

3. **Normality:** the residuals are normally distributed

To test for normality, we can visualize using a QQ plot, which measures the divergence of the residuals from a normal distribution. There are formal statistical tests as well, Shapiro-Wilk, Kolmogorov-Smironov, Jarque-Bera, or D'Agostino-Pearson, but these are sensitive to large sample sizes - they often conclude that residuals are **not** normal when the sample size is large.

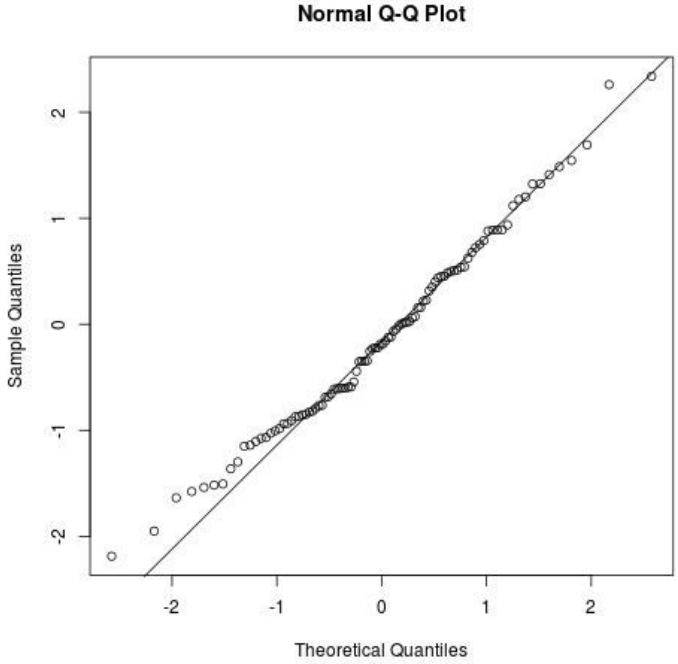
What if there isn't normality?

- This might mean there is a large outlier problem or other assumptions are being violated
- First verify the other assumptions, then check for outliers and see if data could be subset

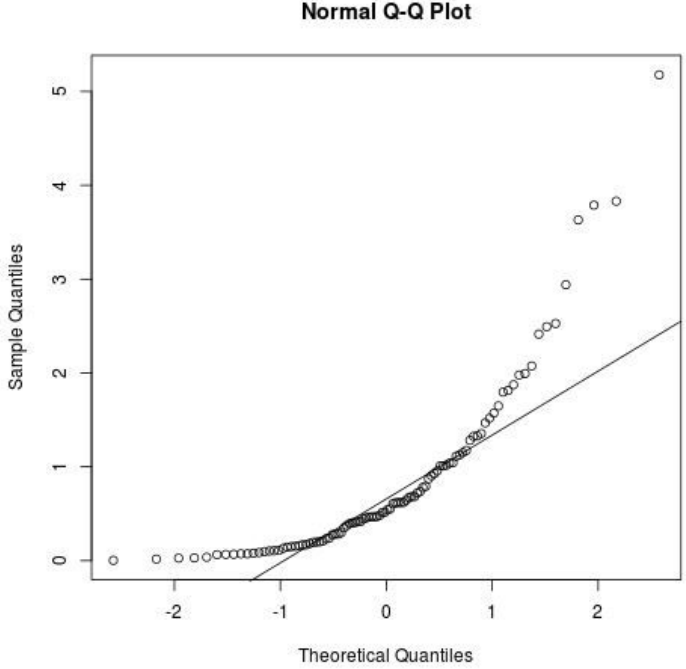
Assumptions: Normality

3. **Normality:** the residuals are normally distributed

Good QQ-Plot



Bad QQ-Plot

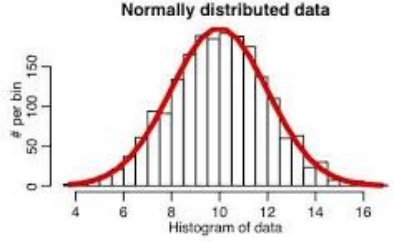


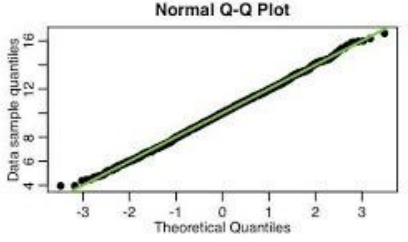
Source

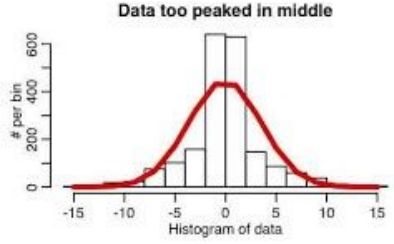
Assumptions: Normality

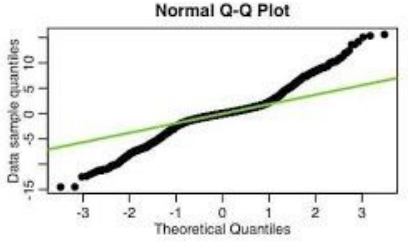
3. **Normality:** the residuals are normally distributed

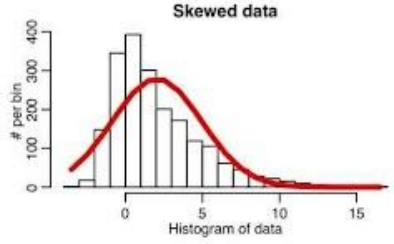
What is my QQ-plot telling me?

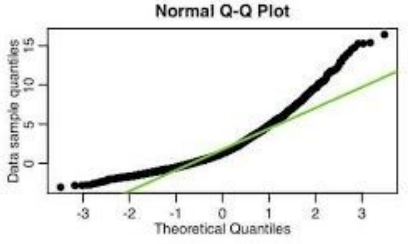












Source

Assumptions: Homoscedasticity

4. **Homoscedasticity:** the variance of the residuals is constant

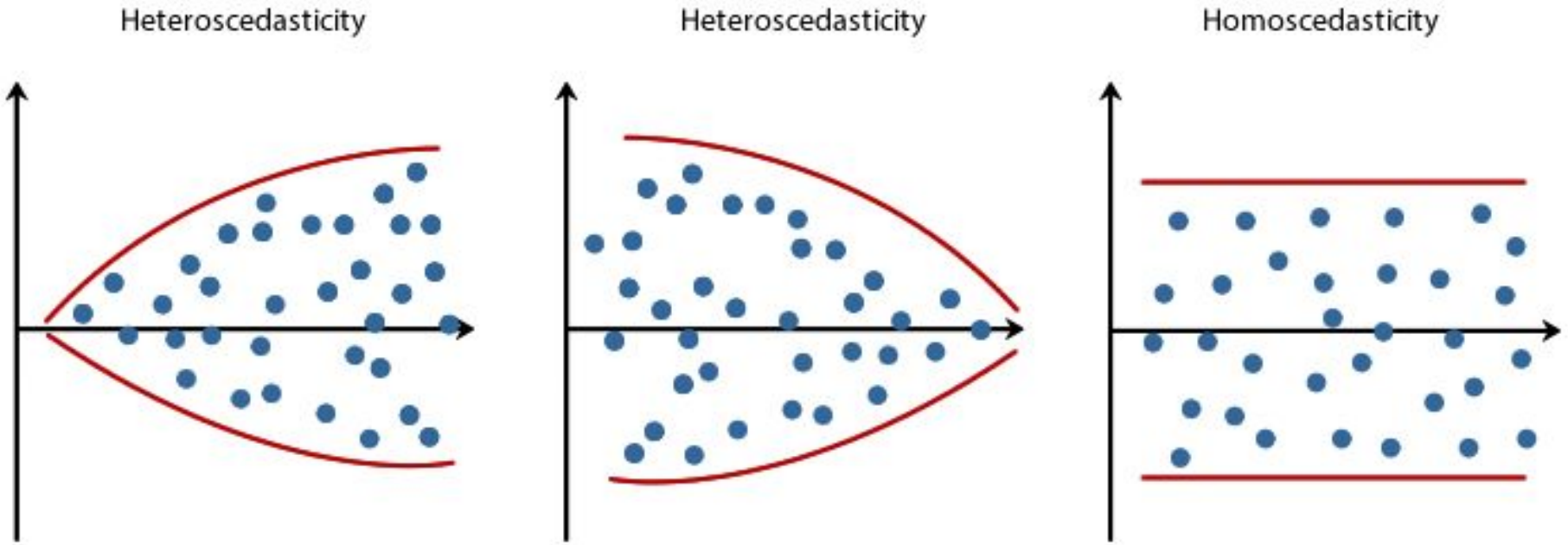
To test for homoscedasticity, we can visualize using a residual plot, which will verify the variance of the error term is constant across all the values of the dependent variable. If this scatter plot forms any sort of pattern in the data, then the data is **NOT** homoscedastic, instead it is heteroscedastic.

What if the variance is heteroscedastic?

- This means that the model will **not** fit all parts of the model equally and **will** lead to bias in the predictions. It often means that confounding variables (more on this later!) have been omitted
- Examine the features, are there confounding variables missing? If so, do you have those variables? Add them in the model and check the assumption again

Assumptions: Homoscedasticity

4. **Homoscedasticity:** the variance of the residuals is constant



Source

Assumptions: Multicollinearity

5. No multicollinearity: the independent variables are not highly correlated with each other

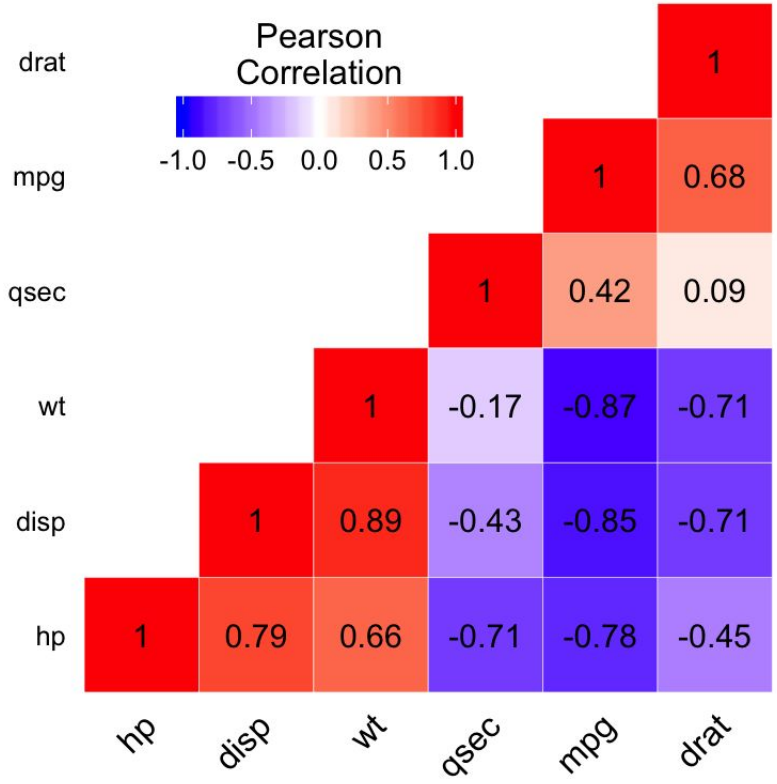
Multicollinearity, or collinearity, refers to the fact that two or more independent variables are highly correlated. While this instance may not be important for non-parametric models, it is necessary for parametric models such as linear regression.

What if there is collinearity in the model?

- Often, the wrong sign (+ or -) of the estimated coefficients is a tell-tale sign of collinearity. Pairwise correlations could be the first step to identify potential relationships
- A more thorough method, however, would be to look at the Variance Inflation Factors (VIF)
- If there is a feature that is collinear with other features, **REMOVE IT!**

Assumptions: Multicollinearity

5. **No multicollinearity**: the independent variables are not highly correlated with each other



A great first step to identifying correlated independent variables, however it can only pick up **pairwise** effects

If you are looking for multicollinearity, USE VIF!

Source

Assumptions: Multicollinearity - VIF

5. No multicollinearity: the independent variables are not highly correlated with each other

Variance Inflation Factor is defined as: $$VIF = \frac{1}{1 - R^2}$$

VIF runs OLS for each independent variable as a function of all the other predictors. K times for k predictors.

- VIF starts at 1 and has no upper limit
- VIF = 1, no correlation between the independent variable and the other variables
- VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others
- VIF is an iterative process - what does this mean? Let's take a look at an example

Assumptions: Multicollinearity - VIF Example

5. No multicollinearity: the independent variables are not highly correlated with each other

We are trying to predict salary (dependent variable) based on the independent variables.This is what our data looks like:

	Gender	Age	Years of service	Education level	Salary
0	0.0	27.0	1.7	0.0	39343.0
1	1.0	26.0	1.1	1.0	43205.0
2	1.0	26.0	1.2	0.0	47731.0
3	0.0	27.0	1.6	1.0	46525.0
4	0.0	26.0	1.5	1.0	40891.0

From our data, let's calculate VIF using Stats Models: VIF

$$VIF = \frac{1}{1 - R^2}$$

Source

Assumptions: Multicollinearity - VIF Example

5. No multicollinearity: the independent variables are not highly correlated with each other

Here we have calculated the VIF scores for each of the independent variables and can see both Age and Years of Service have a VIF greater than 10. Let's remove the variable that has the highest VIF, Age (13.71) and re-calculate the VIF for the remaining variables to see if multicollinearity still exists.

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

Source

Assumptions: Multicollinearity - VIF Example

5. No multicollinearity: the independent variables are not highly correlated with each other

We can see here that after removing Age and re-calculating VIF, the independent variable, Years of Service, is no longer correlated with the other variables.

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

	variables	VIF
0	Gender	1.863482
1	Years of service	2.478640
2	Education level	2.196539

Source

Assumptions: Multicollinearity - VIF

5. **No multicollinearity**: the independent variables are not highly correlated with each other

Remember that identifying multicollinearity is an iterative process, you might have several independent variables that have a high VIF score. Use your best judgement which ones should be removed before re-calculating VIF.

Assumptions: Multicollinearity

5. **No multicollinearity**: the independent variables are not highly correlated with each other

A quick aside:

Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set, it only affects calculations regarding individual independent variables, which is why it isn't necessarily always deemed as an assumption of inferential linear regression

Assumptions of Inferential Linear Regression

In general, these are the main assumptions of inferential linear regression, however there are many other factors that can impact this model that haven't been mentioned. One of the other main secondary assumptions is **that there are no influential outliers**.

Breakout: Assumptions

Do a quick Google search on the other secondary assumption: **that there are no influential outliers**.

1. What happens to the model if there are influential outliers?

2. What can we do to combat these outliers?

3. Are there any formal statistical tests for these outliers?

Assumptions of Inferential Linear Regression

In general, these are the main assumptions of inferential linear regression, however there are many other factors that can impact this model that haven't been mentioned. One of the other main secondary assumptions is **that there are no influential outliers**.

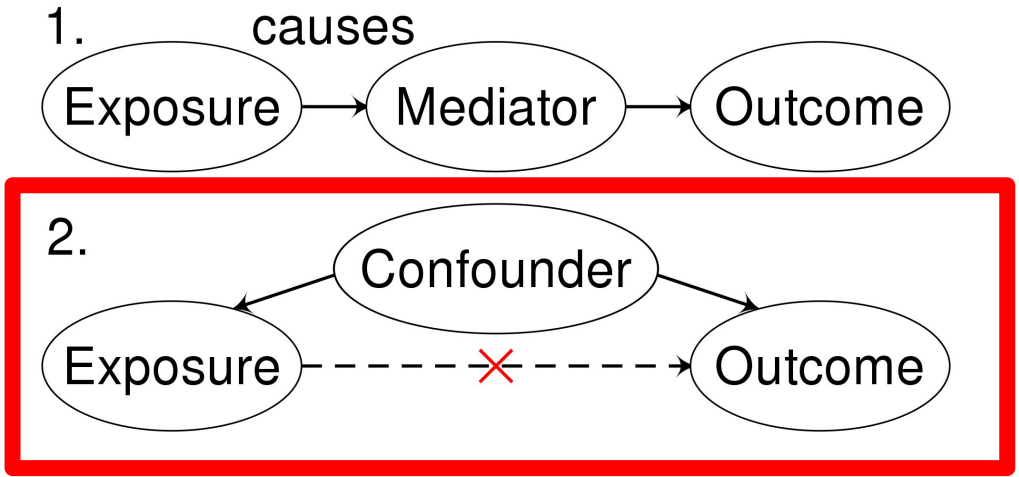
If the assumptions aren't met, it reduces the reliability of the model itself and the results may not be valid or meaningful.

In an ideal world, you will create a predictive linear regression model that holds at least linearity, normality, and homoscedasticity to create a powerful predictive model!

Confounding in Regression

A confounding variable is formally defined as:

1. The confounding variable is correlated with the dependent variable
2. The confounding variable is correlated with the independent variable
3. The confounding variable does not lie on the path from the independent variable to the dependent variable



Source

Confounding in Regression

A confounding variable is formally defined as:

1. The confounding variable is correlated with the dependent variable
2. The confounding variable is correlated with the independent variable
3. The confounding variable does not lie on the path from the independent variable to the dependent variable

Informally, a confounding variable may be the reason for the association between the independent and dependent variable, or conversely, may be the reason you don’t see an association between the independent and dependent variable

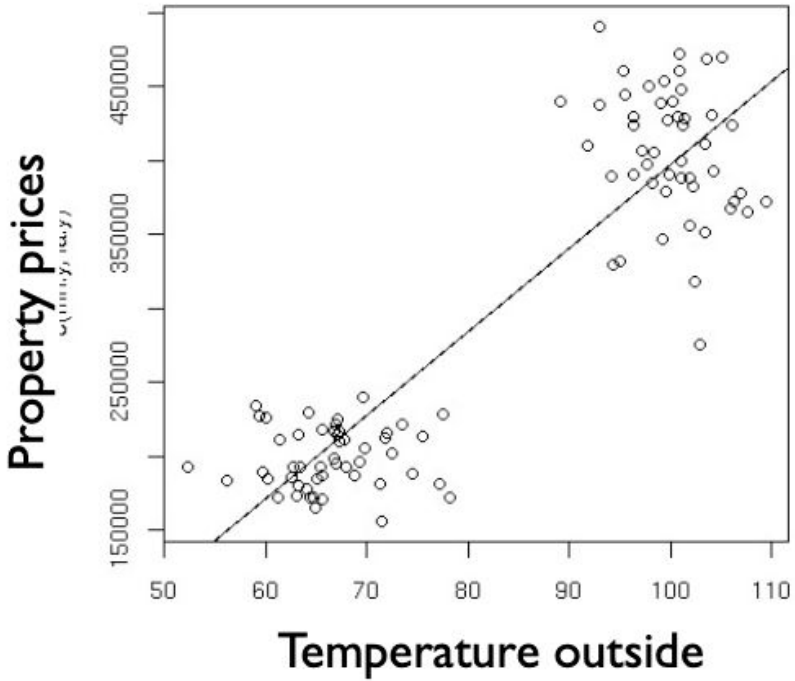
Confounding in Regression: Example 1

A realtor made an interesting observation. She discovered that the correlation between the price of the property and temperature outside while closing a sale was 0.902. From that point on, she planned to close sales only on warm days.

- The data the realtor was using were collected in two places, LA and a town in northern Minnesota
- The realtor forgot that **correlation does not mean causation** but only association and made a mistake by completely disregarding the effect of the location (a confounding factor)!

Confounding in Regression: Example 1

The data the realtor was using were collected in two places, LA and a town in northern Minnesota



Confounding in Regression: Example 1

By not taking into account confounding factors (adjusting for them, putting them in the model), we might obtain a positive relationship, while in fact there is no relationship between closing price and temperature in each location.

- We would label this as confounding away from the null (there is no association when adjusted for location)

Confounding in Regression: Example 2

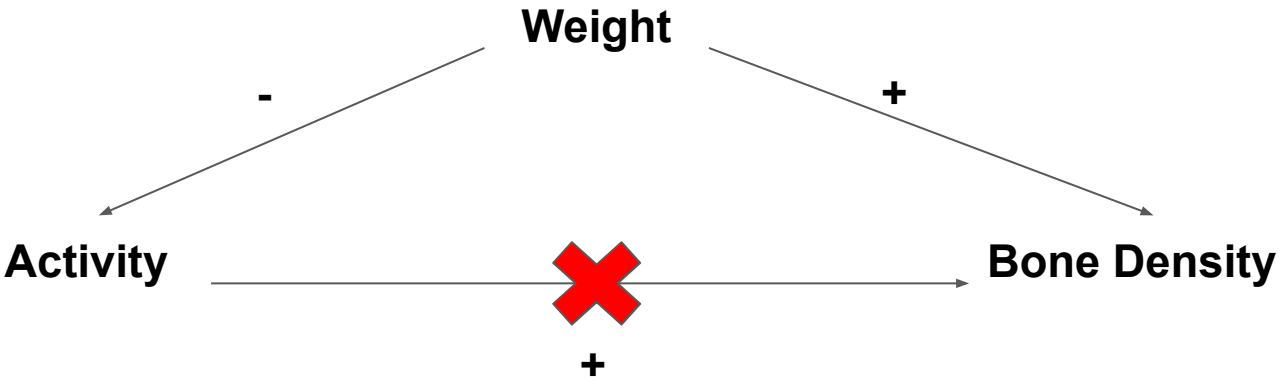
In a study assessing the effects of physical activity on bone density, various characteristics were measured in each subject including, activity level, weight, and bone density, to name a few. The leading theory is that higher activity level produces greater bone density.

In a study, a simple regression analysis was performed to determine whether there was a relationship between activity and bone density. Assuming all the data were valid, the hypothesis is that there would be a positive relationship, however, the simple regression showed **no relationship** at all.

What is happening here?

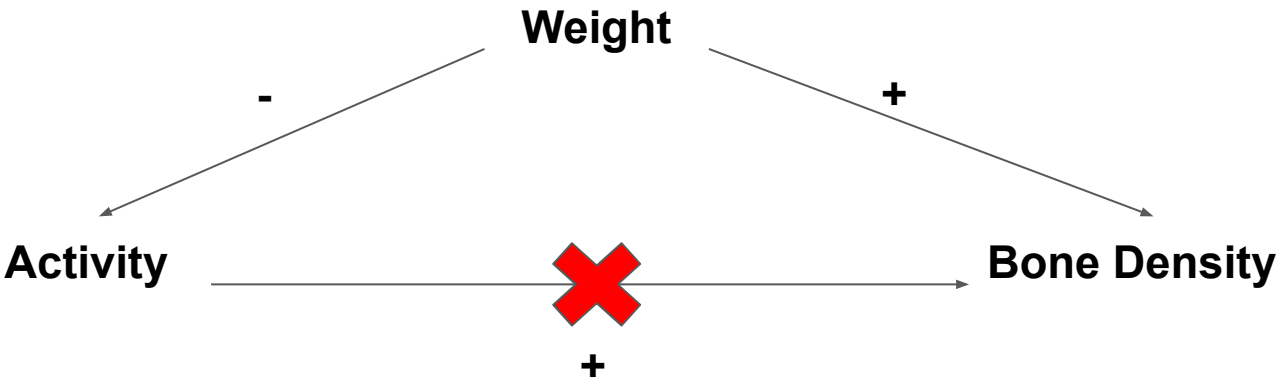
Confounding in Regression: Example 2

This is a case of the model exhibiting **omitted variable bias**, meaning that there is a confounding variable that isn't being accounted for. Another variable that was measure was subjects' weight, let's think through the properties of a confounding variable in regards to weight.



Confounding in Regression: Example 2

So adding in weight to the regression model, along with activity, resulted in a statistically significant positive association between both weight → bone density and activity → bone density. So why did we see a no association when we didn't account for weight?



Independent Variables in Regression

There are two types of independent variables in the feature matrix:

1. Continuous Variables
2. Categorical Variables

Independent Variables: Continuous

Interpreting continuous variables:

As briefly described this previously:
Holding all else constant, on average, a 1 unit increase in *independent variable* increases/decreases the *dependent variable* by *estimated coefficient of independent variable*.

Grumpiness = -7.5sleep + 85

How would you interpret the sleep independent variable?

Source

Independent Variables: Continuous

There are two ways to deal with continuous variables

Leave as is (the standard)	Standardize/Transform (if necessary)
Great for interpretation, we all know what a one year increase in age is!	Harder for lay-people to understand what a one unit increase in a standardization unit
Might not hold true with some of the assumptions	Transforming the data can help with upholding the assumptions
	Standardization of independent variables allows for comparisons of estimated coefficients on the dependent variable, **not** importance of the variable

Independent Variables: Categorical - Binary

Binary encoding (yes/no, 1/0, True/False, exists/not)

Example: Biological Sex (Male/Female)

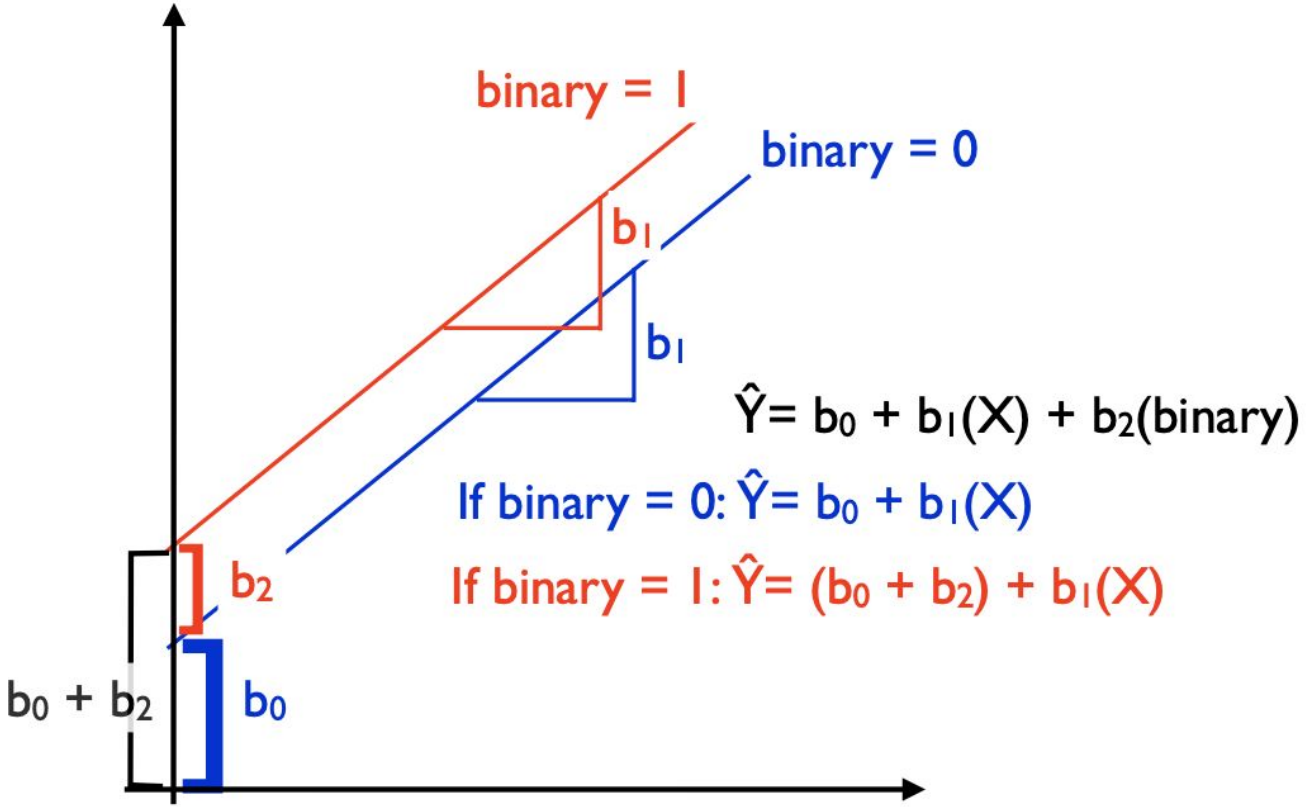
$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

B_1 quantifies how much being female changes the response compared to being male, in general you are comparing the group encoded as 1 to the group encoded as 0.

We call the group encoded as 0 the “reference group”

Independent Variables: Categorical - Binary



Independent Variables: Categorical - Binary

How do we interpret this type of variable?

We want to predict weight (in lbs) using biological sex (F = 0, M = 1) and height (inches).

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

$$\text{weight} = b_0 + b_1\text{sex} + b_2\text{height}$$

b_0 = $b_0 + b_1*0 + b_2*0$ = the population average weight of women who are 0 inches tall

b_1 = holding height constant, the average weight among men is **b_1 lbs** more/less compared to women

b_2 = holding a constant sex, on average, for every one inch increase in height, weight will increase/decrease **b_2 lbs**

Independent Variables: Categorical - Multilevel

Multilevel encoding

Example: Ethnicity (Asian, Caucasian, African American)

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

Independent Variables: Categorical - Multilevel

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

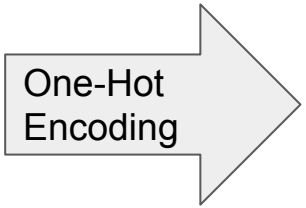
B_1 quantifies how much being Caucasian changes the response relative to being African American.
B_2 quantifies how much being Asian changes the response relative to being African American

Our reference group here is African American, but how did this get coded?

Independent Variables: Categorical - Multilevel

For multi-level variables we have to create **dummy variables** (in Python) by a process called **one-hot encoding**. A dummy variable is an indicator (1/0) variable which indicates one level out of k-categories.

id	color
1	red
2	blue
3	green
4	blue



id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Independent Variables: Categorical - Multilevel

For multi-level variables we have to create **dummy variables** (in Python) by a process called **one-hot encoding**. A dummy variable is an indicator (1/0) variable which indicates one level out of k-categories.

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Now, to pick a reference group, aka the column you drop from your dataset and the column that you want all the other groups to be compared to. Typically the majority group is dropped or if there is some sort of ordinality to your category levels (first, second, third), chose the group you want to compare to. This is subjective and up to you. **Note: Best practice is that you do not compare non-reference groups to each other, though with a little algebra, it is possible**

Independent Variables: Categorical - Multilevel

For multi-level variables we have to create **dummy variables** (in Python) by a process called **one-hot encoding**. A dummy variable is an indicator (1/0) variable which indicates one level out of k-categories.

id	color_red	color_green
1	1	0
2	0	0
3	0	1
4	0	0

We picked blue, as in this case it was the majority. Now the regression model will include color_red and color_green, each compared to the reference group, color_blue.

For k-level categorical independent variables, we need to create k-1 dummy variables!

Independent Variables: Categorical - Multilevel

How do we interpret this type of variable?

We want to predict weight (in lbs) using ethnicity (Asian, AA, Caucasian-reference) and activity level (min).

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$$\text{weight} = b_0 + b_1\text{activity} + b_2\text{Asian} + b_3\text{AA}$$

$\mathbf{b_0} = b_0 + b_1*0 + b_2*0 + b_30$ = the population average weight of Caucasians who are 0 inches tall

$\mathbf{b_1}$ = holding a constant ethnicity, on average, for every one minute increase in activity, weight will increase/decrease $\mathbf{b_1}$ **lbs**

$\mathbf{b_2}$ = holding a constant activity, on average, weight will be $\mathbf{b_2}$ **lbs** more/less among Asians compared to Caucasians

$\mathbf{b_3}$ = holding a constant activity, on average, weight will be $\mathbf{b_3}$ **lbs** more/less among AA compared to Caucasians

Quick Aside on Interpretations

When looking at the estimated coefficients, do a sanity check first.

- Do the coefficients point in the correct direction
    - Does the coefficient positive sign make sense for that specific independent and dependent variable?
- Check the magnitude (only if you have a rough idea of what it should be)
- Are the p-values significant?
    - This tells us whether or not that estimated coefficient differs from 0 or not, if so, it might be worth interpreting

Quick Aside on Interpretation

In inferential regression, we are interested in which independent variables are associated with the dependent variable, but how do we identify this?

OLS Regression Results

Dep. Variable:	Qty	R-squared:	0.995
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	539.9
Date:	Tue, 29 Dec 2015	Prob (F-statistic):	1.52e-17
Time:	19:57:24	Log-Likelihood:	-162.20
No. Observations:	23	AIC:	338.4
Df Residuals:	16	BIC:	346.3
Df Model:	6		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Icpt	2467.8182	6953.489	0.355	0.727	-1.23e+04	1.72e+04
P	-2.047e+04	1987.262	-10.302	0.000	-2.47e+04	-1.63e+04
P1	9210.1391	5371.340	1.715	0.106	-2176.594	2.06e+04
P2	1887.7725	1992.192	0.948	0.357	-2335.486	6111.031
DI	0.0241	0.004	5.389	0.000	0.015	0.034
Pop	63.6892	30.960	2.057	0.056	-1.943	129.321
CPI	-3974.8493	4377.245	-0.908	0.377	-1.33e+04	5304.496

Omnibus:	1.435	Durbin-Watson:	2.458
Prob(Omnibus):	0.488	Jarque-Bera (JB):	0.395
Skew:	0.228	Prob(JB):	0.821
Kurtosis:	3.453	Cond. No.	5.75e+07

Source

Breakout: Interpretation

In terms of y and the respective independent variables, assume everything is on a continuous scale:

OLS Regression Results

Dep. Variable:	Qty	R-squared:	0.995
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	539.9
Date:	Tue, 29 Dec 2015	Prob (F-statistic):	1.52e-17
Time:	19:57:24	Log-Likelihood:	-162.20
No. Observations:	23	AIC:	338.4
Df Residuals:	16	BIC:	346.3
Df Model:	6		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Icpt	2467.8182	6953.489	0.355	0.727	-1.23e+04	1.72e+04
P	-2.047e+04	1987.262	-10.302	0.000	-2.47e+04	-1.63e+04
P1	9210.1391	5371.340	1.715	0.106	-2176.594	2.06e+04
P2	1887.7725	1992.192	0.948	0.357	-2335.486	6111.031
DI	0.0241	0.004	5.389	0.000	0.015	0.034
Pop	63.6892	30.960	2.057	0.056	-1.943	129.321
CPI	-3974.8493	4377.245	-0.908	0.377	-1.33e+04	5304.496

Omnibus:	1.435	Durbin-Watson:	2.458
Prob(Omnibus):	0.488	Jarque-Bera (JB):	0.395
Skew:	0.228	Prob(JB):	0.821
Kurtosis:	3.453	Cond. No.	5.75e+07

Source

Is the model worth looking over?

If so:

- Which of these independent variables would you interpret and why?
- What would those interpretations be?

If not:

- What steps would you take to check the validity of the model?

Learning Objectives

By the end of this lecture, you will be able to:

- Understand the differences between predictive and inferential linear regression
- Identify the assumptions of an inferential linear regression model
- Define and detect collinearity between features using VIF
- Understand what confounding is and how it can impact a model's results
- Encode categorical features to use in a regression model
- Interpret the model output for a generalized audience