**Discussion 8: Regression**

1. What is correlation?

Correlation is the metric which measures how two different variables change in relation to each other. It indicates the strength and direction of their linear relationship.

1. Order the charts below (a,b,c) in order of highest to lowest correlation. Additionally, interpret the charts’ correlation coefficient.

Chart, scatter chart

Description automatically generated

*(source:* [*https://www.sv-europe.com/wp-content/uploads/correlation-1.png*](https://www.sv-europe.com/wp-content/uploads/correlation-1.png)*)*

Highest to lowest correlation: a, c, b

A’s correlation indicates that Horsepower and Vehicle Weight have a very strong, positive relationship.

B’s correlation indicates that Time to Accelerate and Miles Per Gallon have a moderate, positive correlation.

C’s correlation indicates that Time to Accelerate and Horsepower have a strong, negative correlation.

However, we cannot necessarily say that there is a causal effect between any pairing of variables (ie: we can’t say that Horsepower causes Vehicle Weight to go up, only that the data shows that higher Vehicle Weight is often accompanied by higher Horsepower.)

1. What is multicollinearity?

Multicollinearity is when 2 or more independent variables are highly related (correlated).

1. Explain Regression in a paragraph as if you were talking to someone without a background in coding or statistics.

Regression is a statistical method that analyzes whether or not an occurrence of one or more things is related to the outcome of something else by making a best guess at each possible combination of occurrences.

1. What is the difference between classification and regression models?

The dependent variable in a classification model is limited to a finite amount of whole numbers (classes, ie: red (1), blue (2), orange (3) or Yes (1) vs. No (0)) whereas regression models allow for an independent variable that can take on an infinite amount of values (- inf, inf).

1. What is the error (hint: what value difference?) that we are trying to minimize in a regression model?

The error we are trying to minimize in a regression model is the difference between our model’s predicted value for the dependent variable vs the actual value of the dependent variable in our dataset for their corresponding independent variables’ values.

1. Why do we split our data into training and test sets? What is a common split ratio?

We split our data into training and testing sets in order to see how our model does when being tested on data that it doesn’t use to train itself. It’s a way to make sure that the model works on data that it’s unfamiliar with. A common training/test split of data is 80% training data, 20% testing data.