DATA 106 - Linear Regression Cheat Sheet

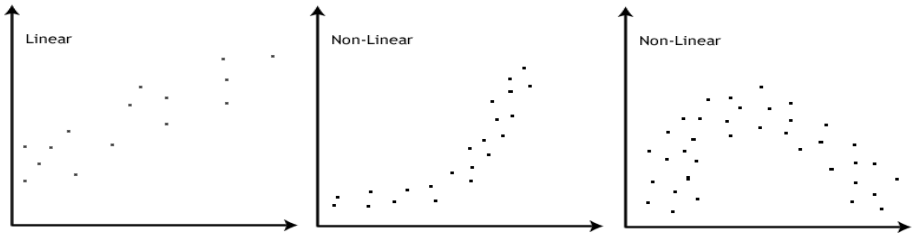
Predictor variable – the thing you are using to predict the response (this is the x variable)

Response variable – the thing you are trying to predict using the predictor (this is the y variable)

1. **Is there a relationship?**

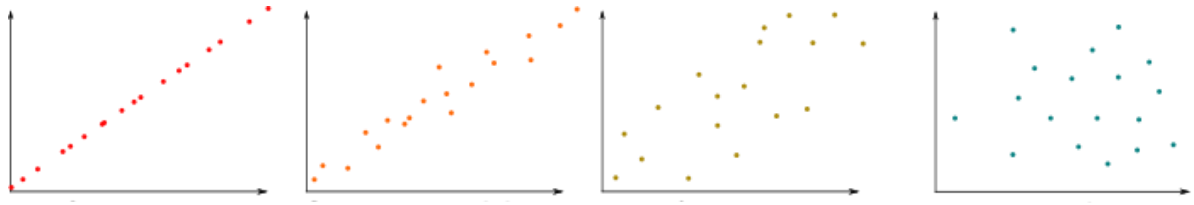
   Look the graph to determine this. If you don't' think there is one, there probably isn't one!

2. **Is the relationship linear?**



   Again, look at the graph to determine this

3. **How strong is the linear relationship/association?**



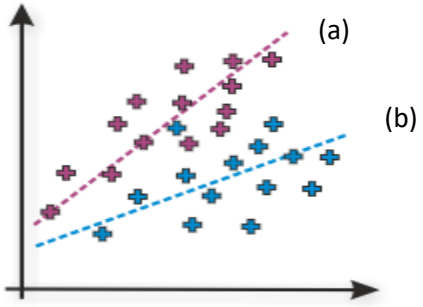
Perfect Strong Weak None

Use Coefficient of Determination (R^2) to determine this

- R^2 closer to 1 has stronger association
- R^2 closer to 0 has weaker association

4. **Which predictor variable has a bigger effect on the response variable?**

For effect, we talk about the magnitude of change in the response variable when there is a change in the predictor variable. In this case, we look at the slope. The bigger the slope is, it means that the effect of the predictor variable on the response variable is larger. Consider the following plot:



In (a) there is a bigger slope compared to (b), so it means that for a unit change in the predictor variable, there will be a greater change in the response variable in (a) than in (b)

DATA 106 - Linear Regression Cheat Sheet

5. **Is the predictor variable a good predictor of the response variable? Or does the predictor contribute to predicting the response?**

Here we care about the size of the slope. Is there evidence to say that the slope is large enough to be different from zero (slope bigger than zero means that the change is meaningful)? For example, consider a slope of 0.005. This is so close to 0, which would imply that changing x doesn't change y (you basically have a horizontal line). But it is close enough to 0 to make this conclusion?

You check this by using a hypothesis test:

Research Hypothesis: the slope is not zero

Null Hypothesis: the slope is zero

To decide which hypothesis to go with:

1) if Pr(>|t|) is less than 0.05 for the variable that you estimated the slope, then you have evidence to suggest that the slope is not zero. THIS suggests THAT THE VARIABLE IS A GOOD PREDICTOR OF THE RESPONSE.
2) if Pr(>|t|) is greater than 0.05 for the variable that you estimated the slope, then you do not have evidence to suggest that the slope is not zero. THIS suggests THAT THE VARIABLE IS **NOT** A GOOD PREDICTOR OF THE RESPONSE.

In statistics, we call if Pr(>|t|) a 'p-value' and scientists and statisticians have agreed that 0.05 is a good threshold to use to determine what decision to make. Calculating a p-value and determining the threshold is beyond the scope of this course since it requires understanding gained in higher level statistics courses.

6. **What effect does the predictor have on the response?**

We look at the slope of the line, but this time interpreting the slope. So, for example, a slope of 0.556 would mean that for a unit change in the predictor, the response would increase by 0.556. Recall the meaning of the equation of a line $y = mx + b$, where $m = \beta_1$= slope and $b = \beta_0 =$ y intercept.

7. **How accurately can we predict the response using the predictor variable? Which variable more accurately predicts the response?**

When we talk of accuracy of predictions, we talk about Root Squared Error or Root Mean Squared Error. This tells you how far away the predictions are from the actual responses. The smaller the RSE or (RMSE), it means that the predictions are closer to the actual response, and hence more accurately predicts.

Example: