**8 Linear Modeling- Introduction**

* What is a model?  
  An object in R that represents the data and can be used for predictions.
* The basis of linear models is the idea that it is possible to fit a line to a set of data points that represents the connection between an independent variable and a dependent variable.
* The independent variable is usually represented on the X axis, which is the input, and the dependent variable is represented in the Y axis as the output.
* The dependent variable is not necessarily a cause of the independent variable i.e. correlation does not mean causation.
* In the real world, all data points will not fit on a single line and the line which minimizes the distance of the line from the points is regarded as the best fit.
* The equation of a line is:  
  Y = MX + B  
  where,  
  Y: dependent variable,  
  X: independent variable,  
  M: slope of the line,  
  B: intercept of the line.
* With the existing data i.e. the X and Y values, the slope and the intercepts can be found where the difference between the actual values of Y and the real values of Y is minimized. Then with the values of slope and intercept, the value of the dependent variable can be found out for any independent variable value.
* R-squared value or the coefficient of determination represents the proportion of variation that is accounted for in the dependent variable by the whole set of independent variables.
* R-squared ranges from 0 to 1, both inclusive.
* There is no range for a good value of R-squared. The context of the problem is important.
* p-value corresponds to the probability of observing the result this extreme by randomness.
* p-value below 0.05 means that the condition was true rather than giving the R-squared value by chance.
* Residuals are the differences between the actual and the predicted values.
* There should not be an independent variable in the analysis if it has a formulaic connection to the dependent variable. That inflates the R-squared value.

**8.2 Example of Linear Modeling**

We use the example where we try to predict the number of repairs a car will need.

We are given attributes like number of oil changes and miles completed which are the independent variables.

We first plot the dependent variables versus the independent variable to gauge whether there seems to be any correlation,

* plot(oil$oilChanges, oil$repairs)
* plot(oil$miles, oil$repairs)  
  The visualization suggests that miles does not seem to be correlated with the number of repairs.
* model1 <- lm(formula = repairs ~ oilChanges, data = oil)  
  We call the lm function which is short for linear models. We give the dependent variable first and the dependent variable with a tilde in between and specify the data frame.
* summary(model1)  
  Gives the summary of the model we built. It gives the R-squared value and the p-value.
* abline(model1)  
  Gives the best fitted line for the linear model.

**Another example of linear model**

* oil$oilChangeCost <- oil$oilChanges \* 350  
  Creates a new column which takes into account the oil change cost.
* oil$totalCost <- oil$repairs + oil$oilChangeCost  
  Takes into consideration total cost.
* m <- lm(formula = totalCost ~ oilChanges, data =oil)  
  Creates a new linear model with total cost as the dependent variable.
* abline(m)  
  Gives the best fitted line for the linear model m.
* The inference is that it is better to have minimum amount of oil changes to minimize the cost.
* test = data.frame(oilChanges=0)  
  predict(m, test, type = “response”)  
  This predicts the value of the dependent variable which corresponds to a particular independent variable value.

**Questions from the videos**

* What might be the cost to change the oil?
* The cost to change the oil will be the number of repairs times the cost per oil change.
* What might be the ranges of the cost?
* The ranges of cost of changing the oil will be zero to (9 \* oil change cost).
* How accurate is the model? Did we have all the facts? Did we have all the data?
* There are other possible attributes that could add more context to the information. The weather conditions and the cost of maintenance could be helpful to better predict the number of repairs.