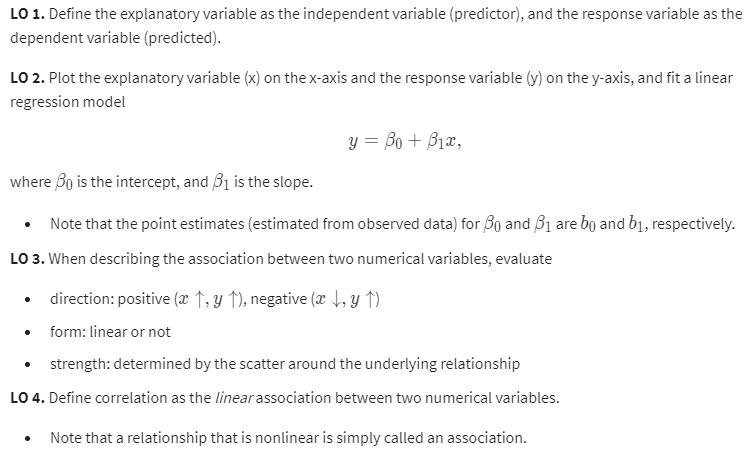
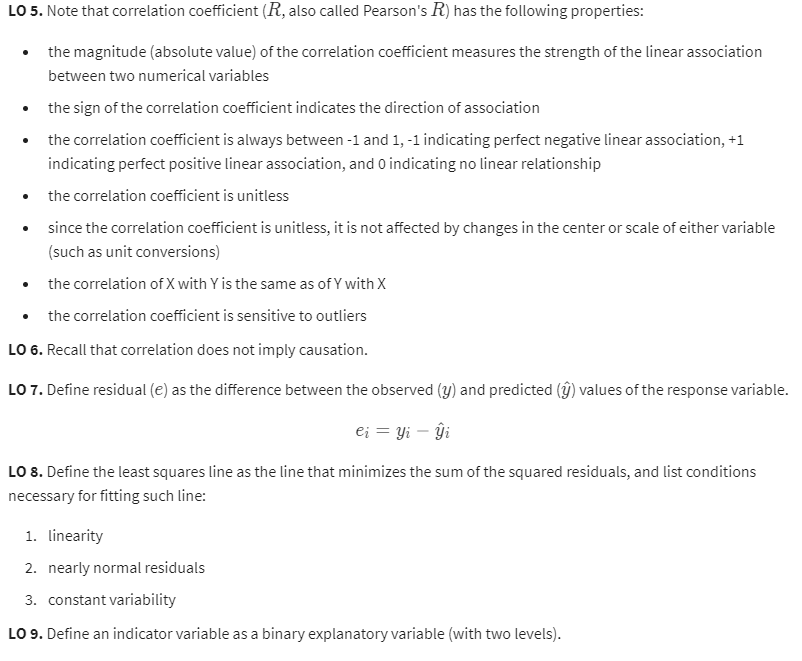
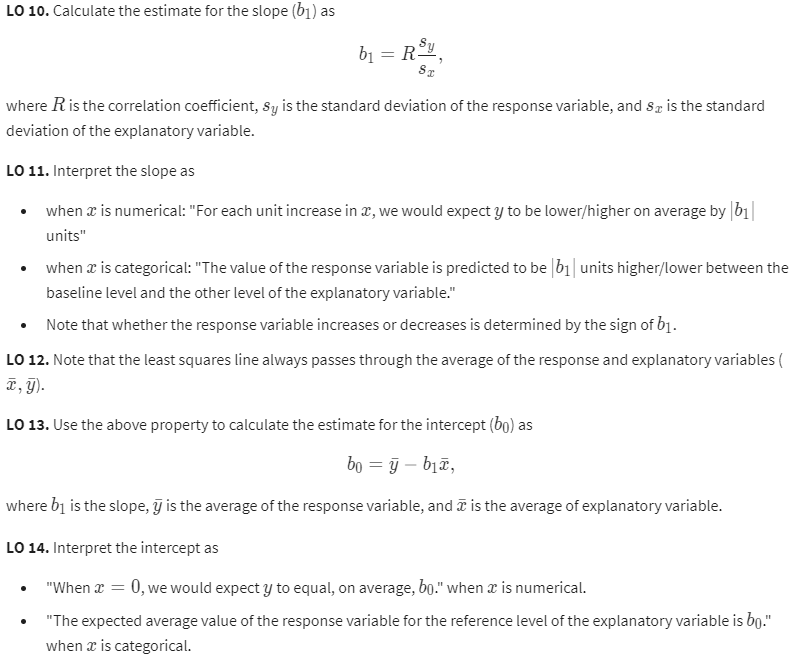
**Course 3 Week 2 Notes**

**Subtopic: Relationship between two numerical variables**

LLOs







Correlation

See LLOs for correlation properties

Residuals

Difference between the observed and predicted y (observed response variable minus predicted response variable)

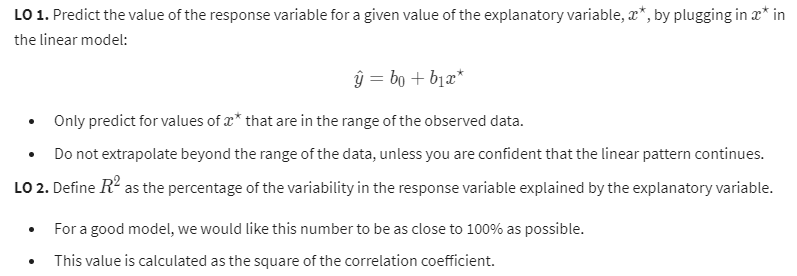
Least Squares Line

Least Squares: minimizing the sum of squared residuals

* Common approach
* A residual is twice as large as another is more than twice as bad
* Calculate the slope and intercept (by estimating)
* Interpret the slope:
  + Example:
    - For each % point increase in HS graduate rate (predictor), we would expect the % living in poverty (response) to be lower on average by 0.62% points
  + Notice that we must avoid causal language
* Calculate the intercept estimate by using the fact that the least squares line always goes through x bar and y bar, the means of x and y
  + The means are calculated from the means of the data
  + Remember x and y can be thought of as 2 columns in the data (i.e. age and price in the same dataset)

**Subtopic: Linear regression with one predictor**

LLOs



Prediction and Extrapolation

* We use the least squares line to predict the response
* Extrapolation
  + Applying a model estimate to values outside of the realm of the original data

Conditions for Linear Regression

1. Linearity

* Relationship between the explanatory and the response variable should be linear
* Check condition using a scatterplot of the data or a residuals plot
  + Residual plots
    - Residuals are the observed value minus the predicted value
    - Ideal residual is zero – the data point falls exactly on the regression line
    - We want residuals to be randomly scattered around zero
    - So we want the residual plot to reflect this
      * Absolutely no pattern in the residual plot is ideal

1. Nearly normal residuals

* Residuals should be nearly normally distributed, centered at 0
* May not be satisfied if there are unusual observations that don’t follow the trend of the rest of the data
* Check using a histogram or a normal probability plot of residuals
  + Want a normal probability plot to be roughly linear

1. Constant variability

* Variability of points around the least squares line should be roughly constant
* Implies that the variability of residuals around the 0 line should be roughly constant as well
* Also called HOMOSCEDASTICITY
* Check using a residuals plot

Notes:

* Generally, we can check the conditions using model diagnostics in R

R Squared

* Square of the correlation coefficient
* Always between 0 and 1
* Example interpretation:
  + Percentage of variability of the response variable explained by the model
  + Incorrect if it mentions the predictor variable

Regression with Categorical Explanatory Variables

* Interpreting the slope and intercept:
  + Example
    - Poverty = 11.17 + 0.38 region:west
    - Intercept: The model predicts an 11.17% average poverty percentage in eastern states
    - Slope: The model predicts that the average poverty percentage in western states is 0.38% higher than in the eastern states

Quick note: In a scatter plot, if there is more scatter around the main trend of the data, it means a weaker strength of association