Course 3 - Linear Regression and Modeling

Week 1

LO 1. Define the explanatory variable as the independent variable (predictor), and the response variable as the dependent variable (predicted).

LO 2. Plot the explanatory variable (x) on the x-axis and the response variable (y) on the y-axis, and fit a linear regression model.

LO 3. When describing the association between two numerical variables, evaluate direction, form, and strength.

LO 4. Define correlation as the linear association between two numerical variables.

LO 5. Note that correlation coefficient (R, also called Pearson's R) has the following properties:

- the magnitude (absolute value) of the correlation coefficient measures the strength of the linear association between two numerical variables
- the sign of the correlation coefficient indicates the direction of association
- the correlation coefficient is always between -1 and 1, -1 indicating perfect negative linear association, +1 indicating perfect positive linear association, and 0 indicating no linear relationship
- the correlation coefficient is unitless
- since the correlation coefficient is unitless, it is not affected by changes in the center or scale of either variable (such as unit conversions)
- the correlation of X with Y is the same as of Y with X

- the correlation coefficient is sensitive to outliers

LO 6. Recall that correlation does not imply causation.

LO 7. Define residual (e) as the difference between the observed (y) and predicted ($\hat{y}$) values of the response variable.

LO 8. Define the least squares line as the line that minimizes the sum of the squared residuals, and list conditions necessary for fitting such line:

- Linearity
- nearly normal residuals
- constant variability

LO 9. Define an indicator variable as a binary explanatory variable (with two levels).

LO 10. Calculate the estimate for the slope (b_1) in terms of the correlation coefficient, the standard deviation of the response variable, and the standard deviation of the explanatory variable.

LO 11. Interpret the slope of a regression coefficient correctly.

LO 12. Note that the least squares line always passes through the average of the response and explanatory variables $(\bar{x}, \bar{y})$.

LO 13. Use the above property to calculate the estimate for the intercept b_0.

LO 14. Interpret the intercept as

- "When $x = 0$, we would expect y to equal, on average, b_0." when x is numerical.
- "The expected average value of the response variable for the reference level of the explanatory variable is b_0." when x is categorical.

LO 15. Predict the value of the response variable for a given value of the explanatory variable, $x^\star$, by plugging in $x^\star$ in the linear model.

LO 16. Define R^2 as the percentage of the variability in the response variable explained by the the explanatory variable.

Week 2

LO 1. Define a leverage point as a point that lies away from the center of the data in the horizontal direction.

LO 2. Define an influential point as a point that influences (changes) the slope of the regression line.

LO 3. Do not remove outliers from an analysis without good reason.

LO 4. Be cautious about using a categorical explanatory variable when one of the levels has very few observations, as these may act as influential points.

LO 5. Determine whether an explanatory variable is a significant predictor for the response variable using the t-test and the associated p-value in the regression output.

LO 6. Set the null hypothesis testing for the significance of the predictor as $H_0 : \beta_1 = 0$, and recognize that the standard software output yields the p-value for the two-sided alternative hypothesis.

LO 7. Calculate the T score for the hypothesis test.

LO 8. Note that a hypothesis test for the intercept is often irrelevant since it's usually out of the range of the data, and hence it is usually an extrapolation.

LO 9. Calculate a confidence interval for the slope.

Week 3

LO 1. Define the multiple linear regression model.

LO 2. Interpret the estimate for the intercept (b_0) as the expected value of y when all predictors are equal to 0, on average.

LO 3. Interpret the estimate for a slope (say b_1) as "All else held constant, for each unit increase in x_1, we would expect y to be higher/lower on average by b_1."

LO 4. Define collinearity as a high correlation between two independent variables such that the two variables contribute redundant information to the model – which is something we want to avoid in multiple linear regression.

LO 5. Note that R^2 will increase with each explanatory variable added to the model, regardless of whether or not the added variable is a meaningful predictor of the response variable. Therefore we use adjusted R^2, which applies a penalty for the number of predictors included in the model, to better assess the strength of a multiple linear regression model.

LO 6. Define model selection as identifying the best model for predicting a given response variable.

LO 7. Note that we usually prefer simpler (parsimonious) models over more complicated ones.

LO 8. Define the full model as the model with all explanatory variables included as predictors.

LO 9. The significance of the model as a whole is assessed using an F-test.

LO 10. Note that the p-values associated with each predictor are conditional on other variables being included in the model, so they can be used to assess if a given predictor is significant, given that all others are in the model.

LO 11. Stepwise model selection (backward or forward) can be done based on p-values (drop variables that are not significant) or based on adjusted R^2 (choose the model with higher adjusted R^2).

LO 12. The general idea behind backward-selection is to start with the full model and eliminate one variable at a time until the ideal model is reached.

LO 13. The general idea behind forward-selection is to start with only one variable and adding one variable at a time until the ideal model is reached.

LO 14. Adjusted R^2 method is more computationally intensive, but it is more reliable, since it doesn't depend on an arbitrary significance level.

LO 15. List the conditions for multiple linear regression.

LO 16. Note that no model is perfect, but even imperfect models can be useful.