1 Multiple Regression

From Last Time:
- Which variables should you include? Does it matter if a variable that should be in the model is omitted?
- Omitted variable bias. True model:
  \[ y_i = b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i \]
  where \( \varepsilon \) is the random error term and \( x = (X_i - \bar{X}) \).
- Estimated model:
  \[ y_i = b_1 x_{1i} + \epsilon_i \]
  where the estimator for \( \hat{b}_1 \) is:
  \[ \hat{b}_1 = \frac{\sum_i x_{1i}y_i}{\sum_i x_{1i}^2} \]
- Substitute the true model into the least squares formula. (Note, think about what you would get if the bi-variate was the true model):
  \[ E(\hat{b}_1) = \frac{\sum_i x_{1i}(b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i)}{\sum_i x_{1i}^2} \]
- Multiply out the bracket, and note that the \( E(x_1, \varepsilon) = 0 \), the bias is:
  \[ E(\hat{b}_1) = b_1 + b_2 \frac{\sum_i x_{1i}x_{2i}}{\sum_i x_{1i}^2} \]

For this week:
- Inference in the multivariate regression line:
  \[ Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \]
- Testing the significance of individual regression coefficients: (dropping \( i \) subscripts)
- Intercept term (a):
  \[ \hat{\sigma}_a^2 = \left[ \frac{1}{n} + \frac{\sum x_1^2}{\sum x_1^2} + \frac{\sum x_2^2}{\sum x_2^2} - \frac{2 \sum x_1 x_2}{\sum x_1^2 (\sum x_1 x_2)^2} \right] \hat{\sigma}_{YX}^2 \]
  and the standard error
  \[ \hat{\sigma}_a = \sqrt{\hat{\sigma}_a^2} \]
• Slope coefficients \((b_1 + b_2)\):

\[
\hat{\sigma}_{b_1}^2 = \left[ \frac{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \right] \hat{\sigma}_{YX}^2
\]

\[
\hat{\sigma}_{b_1} = \sqrt{\hat{\sigma}_{b_1}^2}
\]

\[
\hat{\sigma}_{b_2}^2 = \left[ \frac{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \right] \hat{\sigma}_{YX}^2
\]

\[
\hat{\sigma}_{b_2} = \sqrt{\hat{\sigma}_{b_2}^2}
\]

where \(\hat{\sigma}_{YX}^2 = \frac{\sum \varepsilon^2}{n-k}\), \(k = \) number of parameters in the model; \(n = \) is the sample size.

• Sum of squares identity is now:

\[
\sum y^2 = \left( \hat{b}_1 \sum x_1 y + \hat{b}_2 \sum x_2 y \right) + \sum \varepsilon^2
\]

• Coefficient of determination:

\[
R^2 = \frac{\hat{b}_1 \sum x_1 y + \hat{b}_2 \sum x_2 y}{\sum y^2}
\]

• F-test for the joint test \(H_0: b_1 = b_2 = 0:\)

\[
F^* = \frac{\left( \hat{b}_1 \sum x_1 y + \hat{b}_2 \sum x_2 y \right) / (k - 1)}{\sum \varepsilon^2 / (n - k)}
\]

\(F^* F_{(k-1),(n-k)}:\)

• Relationship between \(R^2\) and the F-test:

\[
F^* = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}
\]

• Leads to an adjustment on \(R^2:\)

\[
\overline{R^2} = 1 - \left(1 - R^2\right) \left( \frac{n - 1}{n - k} \right)
\]