Regression - Summary

1. Regression

   We are interested in exploring the relationship between two variables y (the response) and x (the predictor). If the model is linear we can express it as:

   $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

   where y_i is the response for observation i, x_i is the predictor for observation i, and $\epsilon_i \sim iidN(0, \sigma^2)$ is the random error for observation i. The goal is to estimate the values of β_0 and β_1. We do this by minimizing:

   $$SSE = \sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_i])^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

   with respect to β_0 and β_1, which gives:

   $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

   If we wish to test:

   $H_0 : \beta_1 = 0$
   vs.
   $H_A : \beta_1 \neq 0.$

   and we are willing to assume:

   - The model is correct. (A straight line makes sense for the data.)
   - The observations are independent.
   - The variance around the fitted line is constant for all values of x.
   - The random error around the fitted line is normal for each x.

   then we can do this with a T-test.

   $$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

   and we estimate σ^2 using:

   $$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2}.$$

   Then:

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE(\hat{\beta}_1)}} \sim t_{n-2}.$$

Assumptions can be checked by using a scatterplot to check linearity, residuals vs fitted values plot for checking constant variance, and a QQ plot of residuals to check normality.

If we want to predict the y value for a given $x = x^*$, we use:

$$(\text{Fitted value for } x = x^*) = \hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

A measure of accuracy of the prediction, if we are interested in the position of the fitted line is:

$$SE(E(\hat{y}|x^*)) = \sigma\sqrt{1/n + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

If we define $SSTot = \sum_{i=1}^{n}(y_i - \bar{y})^2$, we can create a quantity called R^2:

$$R^2 = \frac{SSTot - SSE}{SSTot}.$$

This can be interpreted as the fraction of the total sum of squares that is explained by the regression line.