ROBUSTNESS

Our model for simple linear regression has four assumptions:

1. Linear mean function: \( E(Y|x) = \eta_0 + \eta_1 x \)

2. Constant variance of conditional distributions: \( \text{Var}(Y|x) = \sigma^2 \) (constant variance)
   
   (Equivalently: Constant variance of conditional errors: \( \text{Var}(e|x) = \sigma^2 \))

3. Independence of observations: \( y_1, \ldots, y_n \) are chosen independently from \( Y|\mathbf{x}_1, Y|\mathbf{x}_2, \ldots, Y|\mathbf{x}_n \), respectively.

4. \( Y|x \) is normal for each \( x \) (or at least for each \( x_i \) and for each \( x \) where we wish to do inference.)

Robustness is the question of how valid our procedures are if the model doesn't exactly fit.

Robustness to departures from linearity:

- Not all relationships are linear, but sometimes a linear model can be useful even if the relationship is known not to be linear. (e.g., to check for an increasing or decreasing trend, or as a good-enough approximation.) However, results need to be interpreted appropriately.
- Remember that a high \( R^2 \) does not mean that the relationship is linear.
- Often we can transform to linearity to get a better model fit. [More later]
- Outliers (observations that don't fit the general pattern of the data) can have a strong influence on the least squares fit.

Wise practice: If there is just one predictor, always look at a scatter plot before calculating a simple linear regression -- and make decisions about transforming variables and whether or not to include outliers in the analysis.

Robustness to departures from constant variance:

- \( \hat{\eta}_0 \) and \( \hat{\eta}_1 \) are still unbiased estimators of \( \eta_0 \) and \( \eta_1 \).
- Since the constant variance assumption was important in inference, the inference procedures are not reliable in the presence of non-constant variance ("heteroskedasticity"). Another good reason to plot data.
- Possible remedies for nonconstant variance:
  1. Transform to constant variance
  2. Weighted least squares (Chapter 9)
**Robustness to departures from independence of observations:**
- \( \hat{\eta}_0 \) and \( \hat{\eta}_1 \) are still unbiased estimators of \( \eta_0 \) and \( \eta_1 \).
- Since independence of observations was used in developing inference procedures, the inference procedures are not reliable.
- However, if observations are "almost independent," it's probably OK to use inference procedures.
  
  *Important example:* We often sample with replacement, which does not give independent observations -- but with large populations, the covariances are negligible.

**Robustness to departures from normality**
- \( \hat{\eta}_0 \) and \( \hat{\eta}_1 \) are still unbiased estimators of \( \eta_0 \) and \( \eta_1 \).
- Since normality of conditional distributions was used in developing inference procedures, the inference procedures might be questioned.
- However, if \( n \) is large, the Central Limit Theorem implies that the sampling distributions of the estimates are approximately normal.

*Empirical Rule of Thumb:* Inference for \( \hat{\eta}_0 \), \( \hat{\eta}_1 \), and \( \hat{E}(Y|x) \) is approximately valid unless \( n \) is small and the distributions of the \( Y \backslash x \)'s are strongly skewed or bimodal.

**However:**
- The inference procedures are not as powerful -- i.e., they are not as good at distinguishing between close values -- so they are less likely to show evidence against NH when NH is false.

*Thus:* Transforming to (or close to) normality is still desirable. [more later]

b. Prediction is less robust -- since \( y \) may dominate in prediction.