**Week 3, Day 1**

* **Conditions for the Linear Model**
  + Linear between X & Y
  + Error terms (residuals) should be:
    - Normally distributed 🡨 only necessary for inference
    - Constant variance
    - Zero mean 🡨 automatic for least squares regression
  + Independence of the error terms
  + “Random” = Data are representative of the population of interest 🡨 only necessary for inference
* **How do we check these conditions? 🡪 residual plots**
  + plot(model) 🡪 1st two plots only
* **Transformations**
  + Problem: non-linear   
    Solutions:
    - See Figure 3.13 (below)
    - Whenever there’s “bunchiness” on one side, try log.
  + Problem: non-constant variance   
    Solutions:
    - See Figure 3.15 (below)
    - May have to make a simultaneous transformation on X to keep the relation linear. For count data, square root transforms are often helpful.
  + Problem: non-normality of residuals  
    Solutions:
    - Often goes along with non-constant variance, so improving one will often improve both.
  + US States Example: student spending vs. insured
* **Outliers & Influential points**
  + What is an outlier?
    - Standardized residuals : rstandard()
    - Studentized residuals: rstudent()
    - Any standardized/studentized resids outside ±2 (±3) are suspect (*very* suspect)
  + Influential points
    - Does inclusion (or exclusion) of this point change the regression line substantially?
    - Leverage: points with *x* values far away from the mean (); *x*-outlier.
    - Cook's distance (Sec 4.3): measure of influence (takes into account how much it fits the overall pattern and leverage).
      * if Di > 0.5 the point is moderately influential
      * if Di > 1 the point is very influential.
* **Inference in Regression: 3 tests**
  + Questions?

**For Thursday**:

* Read Sections 2.3-2.5 in *Stat2*
* Watch the “Chapter 2: Rsq” and “Chapter 2: Intervals” videos (Lecture Videos link under General tab on Moodle) 🡪 ~ 20 min.



