Class Notes: 1/28/15

Review of last class:

- Encompassing idea of stat models:
  - Takes observation y_i into 2 parts: y-hat and e_i
    - y-hat is fitted value, e_i is the residual
    - Also written as y_i = systematic + unpredictable part
- Between group v. Within group distribution: std dev of y-hat v. std dev of e_i
- Some variation is predictable, some isn't
- Baseline: offset form or dummy variable

Today:

- Observe data in 2 groups: for example, two groups of clinical trial or 2 colleges within a university
  - tracking mean response
  - x_i is a dummy/indicator variable
    - takes value of 1 if case i is in group 1
    - takes value of 0 if case i is in group 0
  - $y_i = B_0 + B_1X_1 + e_i$
    - we're almost always interested in the differences—B_1 is this difference between group means
    - if $x_i = 0$, $y_i = B_0 + e_i$
      - group mean is B_0
      - if $x_i = 1$, $y_i = B_0 + B_1 + e_i$
        - B_1 therefore the difference we care about
    - x_i: dummy variable
    - B_0: baseline/intercept
    - B_1: slope/coefficient on dummy variable
- What if we have more than 2 groups?
  - Groups 0, 1 or 2
    - introduce x_{i1} and x_{i2}
    - if in group 0
      - $y_i = B_0 + e_i$
    - if in group 1
      - $y_i = B_0 + B_1 + e_i$
    - if in group 2
      - $y_i = B_0 + B_2 + e_i$
  - Baseline offset form: always choosing one group to be 0

SAT R Script

- mean SAT.Q ~ School), data=ut2000
- lml = lm(SAT.Q ~ School, data = ut2000) → baseline offset form

Ordinary Least Squares

- ex: What's the asking price of a truck on craigslist?
  - $y_i = B_0 + B_1x_i + e_i$
    - e_i are the "misses"
    - Legendre's idea
    - choose B0 as intercept and B1 as slope to minimize sum of e_i^2
- Why sum of squares?
  - don't want negative values
  - how much does it hurt to miss things?
  - greater punishment for outliers than just taking absolute values
  - real reason: Legendre did calculation by hand
    - minimizing: taking derivative and make it equal to 0
    - can't take derivative of absolute value
  - 2 deeper reasons discovered subsequently:
    - connection between sum of squared errors and normal distribution
      - talked about later in the course
    - variance decomposition
      - sum of squared errors is a very special property
      - sum of squared errors is equivalent to Pythagorean theorem

Pickup Data Set

- lm function helps fit a straight line to scatterplot
  - gives intercept and slope
  - create model called 'model1'
    - nothing happens in console, but keeps model available to do various things
  - add trend line: plot and use abline function: plot(price ~ miles, data = pickup), abline(model1)
- 4 stories to emphasize
  - Story 1: Plug-In Prediction
    - ex: want to sell my truck
      - need to decide how much to charge for it

            - decide reasonable market price by looking up expected price given by the fitten line
        - ex: have 3 cars you want to sell
            - define new variable newx
            - c is a vector
            - y-hat: can plug whole vector into equation instead of individually plugging in
            - y-hat will now give you predicted values
    - Story 2: Summarizing the Trend
        - Ex: I drive my truck that has 50,000 miles. What happens if I wait to sell my truck for a year after it accumulates an additional 10,000. What is the corresponding change in y?
            - y-hat = B0 + B1x
            - take the derivative
            - answer: 642 dollars less
    - Story 3: Taking the "x"-ness out of y: Statistical adjustment
        - Yi = B0 + B1Xi + ei
            - B1Xi is the systematic part that corresponds to X
            - adjustment process: subtract
            - In the residual plot, the highest point is most overpriced
    - Story 4: Quantifying the Reduction in Uncertainty
        - Starting off, we had a certain degree of uncertainty of the price of trucks: How good is the guess? The standard deviation will measure this amount.
        - After we have truck mileage information, we can apply story 1
            - When you don't know mileage, guess sample mean. Use the standard deviation of prices to see how good the guess is.
            - Now we have the straight line model. Look at the residuals from the line to quantify the reduction of uncertainty: 4200 v 5500
        - True Cars makes its money on data analysis
            - data analysis graph tells you: what you'd expect to pay for the car, market average, spread of expected prices.
            - normal distribution that describes residuals from their statistical models

Case Study

- mrall_mean = mean(mrall ~ state, data=traffic2)
- vmiles_mean = mean(vmiles ~ state, data=traffic2)
- plot(mrall_mean ~ vmiles_mean, data=traffic2)
- fred=lm(mrall_mean ~vmiles_mean, data=traffic2)
- coef(fred)
- abline(fred)