1/28/2015

Last time:

- Encompassing idea of stat models partition variation
- Yi=y^I +ei
- Yi= fitted value + residual
- =systematic +unpredictable
- Std deviation of y^I compared to ei to decide within or between group variation is the dominate model
- Basline/ offset form or dummy variable
    - Data in two groups ex) architecture and business
    - Xi (group 1) is a dummy variable
        - Xi= 1 if case I is in group 1
        - Xi=0 if case I is in group 0
    - Yi=Bo+B1Xi+ei
        - B1= difference between group means
    - If Xi=0; Yi= Bo+ei
    - If Xi=1; Yi= Bo+B1+ei
        - Xi is the dummy variable
        - Bo is the baseline/ intercept
        - B1 is the slope or the coefficient on the dummy variable
    - More than 2 groups? More dummy variables
        - Groups: 0,1,2
        - Xi=1 if case I is in group 1
        - Xi= 2 if case I is in group 2
        - Xi=0 if case I is in group 0
        - Group 0 Yi=Bo+ei
        - Group 1 yi=Bo+B1+ei
        - Group 2 yi=Bo+B2+ei

SAT R Script

- UT200.csv load into Rstudio

-
  ```
  mean(SAT.Q ~ School, data=ut2000)
  ```
    - change mean to LM for get it on baseline offset form

    -
      ```
      lm1 = lm(SAT.Q ~ School, data=ut2000)
      ```

    -
      ```
      coef(lm1)
      ```

OLS: ordinary least squares

- Yi=response
    - Ex asking price of a truck on craigslist
    - Xi is the predictor, feature, covariant, independent variable
    - Straight line: Yi=Bo+B1X1

- But this doesn’t hit all the points
- Error is the “misses” of the equation
- Legendre says minimize the sum of ei^2
    - SUM(ei^2,1,n)
    - SUM((Yi-Y^i)^2,1,n)
    - SUM(Yi-(Bo+B1x1))^2
- Why square?
    - No negatives
    - Square the errors punishes the outliers more (proportionate)
    - Real reason: Legendre did this by hand and to minimize the function you take the derivative of the square-> can’t take the derivative of the absolute value
    - Deep conection between sum of sq errors and normal distributions *later in the course*
    - Variance deomposision: special property that can be invoked by pathagreom *we will look at this on Monday*

Pickup walk through

- Upload dataset
    - Picture of coverage interval
    - (lm(price~miles, data=pickup)
        - Calculus in a micro second
- Save the output
    - Store the command as a model
        - Model1=lm(price~miles, data=pickup)
        - Find coef(model1)
        - Find fitted(model1)
        - Resid(model1)
        - Fitted(model1) + resid(model1)=original
    - Line of best fir
        - Plot(price~miles, data=pickup)
        - Abline(model1)
- 4 stories of linear regression

1. Plug in prediction: plug in your value and the prediction comes out
    a. Sell truck on craigslist
    b. Newx = c(25000,50000,100000)
    c. Yhat= 14419.3762 + (-0.0643)*newx
2. Summarize the trend: why we collect data
    a. What will the change in Y be if I wait to sell my truck online
    b. Y^=Bo + B1X
        i. Take derivate with respect to X
        ii. Dy^/dx=B1
    c. Coef(model1) -6.42E-02 * 10000 =-642.9944
3. Taking the X-ness out of Y (statistical adjustment)

a. Residual can be thought of a Y adjusted for X

4. Qualifying the Reduction in Uncertainty
    a. Sd(pickup$price)
    b. Guess sample mean of price, use std dev of prices to see how good that guess will be, graph and pick the point on the graph you are interested in
        i. Quantify with residuals
    c. Sd(resid(model1))
    d. Truecar.com makes money on data analysis
        i. Graph of price comparisions (stories 1-4)
            1. Normal distribution describing residuals on their stat model

Most important points: the uncertainty that remains after those 4 things are done is a huge factor to decision making

Case Study:

1) Define new variables that compute a mean value for each state's statistics (vmiles, mrall, and perinc) across all years. Remember how to compute groupwise means in R.

```
mean(mrall~statename, data=traffic2)
```

way more than 50 dots: we need to store means

```
mrall_mean = mean(mrall ~ statename, data traffic2)
vmiles_mean = mean(vmiles ~ statename, data=traffic2)

plot(mrall_mean~vmiles_mean, data=traffic2)
coef(fred)
abline(fred)
```

2) Make a scatter plot and fit a linear models that shows the relationship between fatality rate (response variable) and miles per driver (predictor).
3) Do the same thing, except with per capita income as the predictor.
4) Make a lattice plot that stratifies the scatter plot of fatality rate versus miles driven by the categorical variable indicating whether the state has mandatory jail for drunk driving. What does this plot suggest?