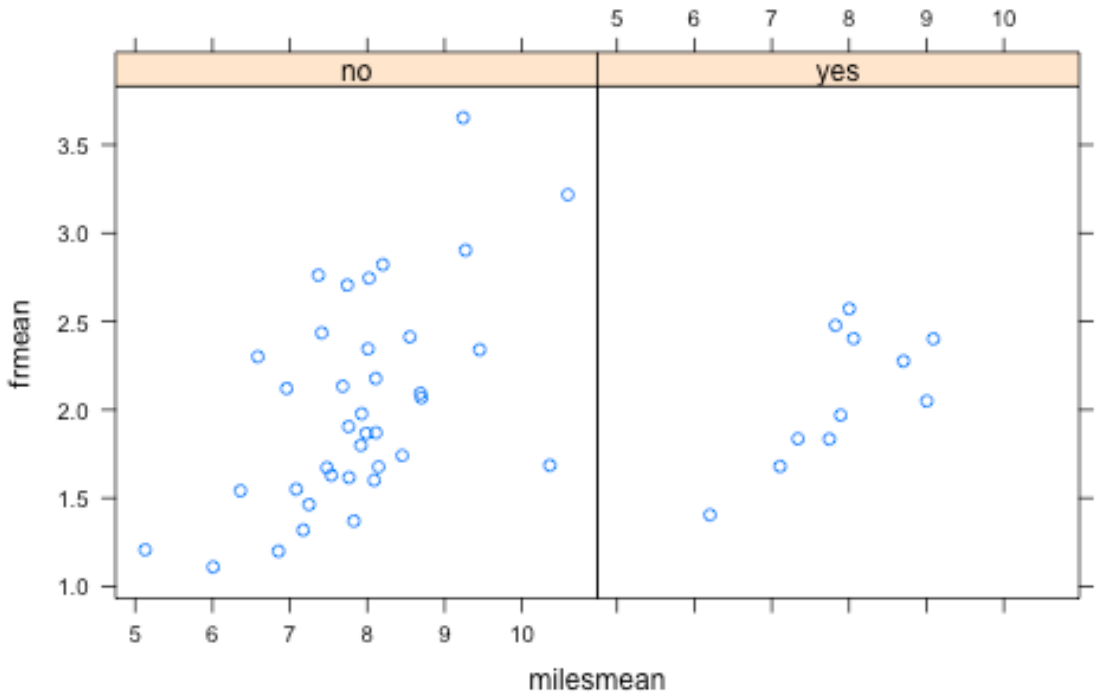
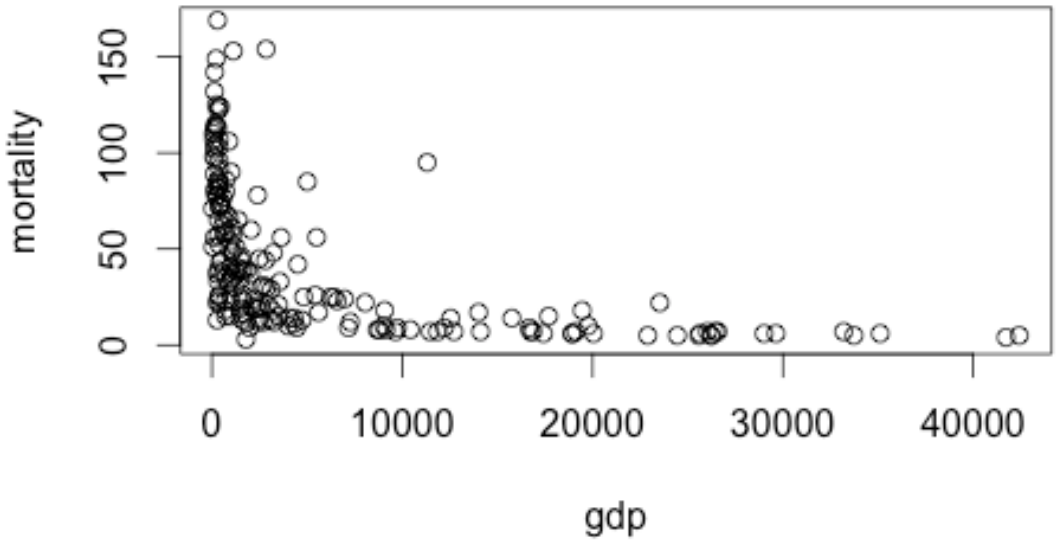
Sabeeha Islam
9:30 AM-11:00 AM

STA 371H Scribing Notes for 1/27/14

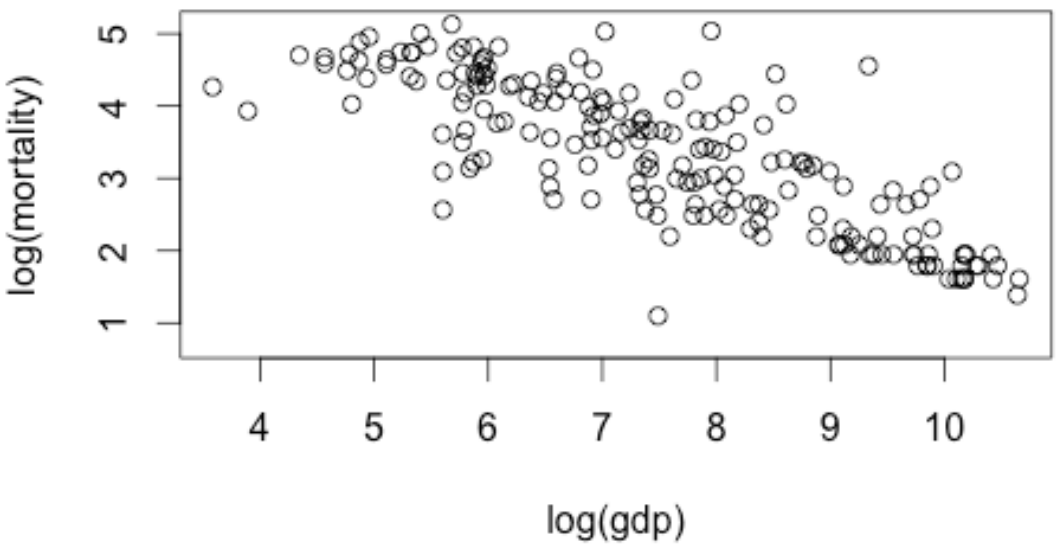
- We began class by reviewing homework due today
  - 1A) Plot Price~Food
  - 1B) Use the dotplot command. Compare to dotplots: the first modeling Price~FoodScore and the second Price~FeelScore
  - 1C) For this question, you can graph the residuals and find the mean of the residuals. The points closest to the line were the "most-valuable" and the ones farthest from the lines were the "least-valuable"
  - 2A) Plot Stock~S&P500 and find a linear model and extract the coefficients. For the residuals, you can plot the residuals and then use "favstats(resid(model))", assuming you have already saved the linear model.
  - 2B) This question is asking to interpret the y-intercept. Since we know the y-intercept is when x=0, we know the y-intercept tells us where the stock will be when the market (S&P500) is at 0. When the y-intercept is positive, it means that on average, the stocks perform better than the market.
  - 2C) Yahoo!'s beta score is their estimation for the slope. We assume the scores are different because of possibly different dates or years taken into calculation. Note that our data is only based on information from 2011-2012.
  - 2D) For this problem, there are two possible ways of approaching the solution. You can calculate the coefficient of determination (r-squared) or by comparing the averages of the residuals found in 2A. If the residuals of TGT and WMT are similar, you can guess that WMT may be predicted by TGT.
- Trafficdeaths Script and Data set
  - Open both the trafficdeaths script and csv files: fips & trafficdeaths
  - Learned to merge files (fips.csv into trafficdeaths.csv)
    - traffic2 = merge(trafficdeaths, fips, by.x = "state", by.y="fipsnum")Df
  - After merging the files, you can define new aggregated variables
    - frmean = mean(mrall~fipsalpha, data=traffic2)
    - For more, see the script
    - You can then take these aggregated new variables and create linear models with them
  - Learned how to add text labels on graphs
    - For r code, see trafficdeaths.R line 28-32
  - Learned how to stratify the aggregated variables by another 3rd variable



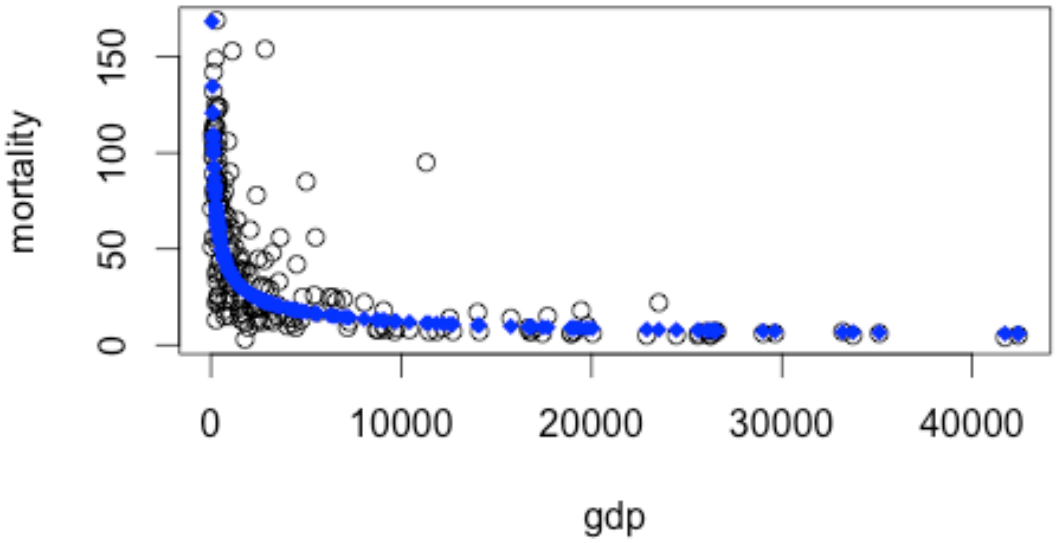
- Transformations.R (learning how to use linear models for non-linear data)
  - Need mosaic package and faraway package (see R script)
  - Import the infmort.csv
  - When we plot infant mortality~gdp per capita, we get a non-linear plot



  - We notice that the data is "squished" towards the bottom and towards the left, but there are long tails on the x & y axes
  - Therefore, we take the log of both the x variable & y variable (because data has tails on both)

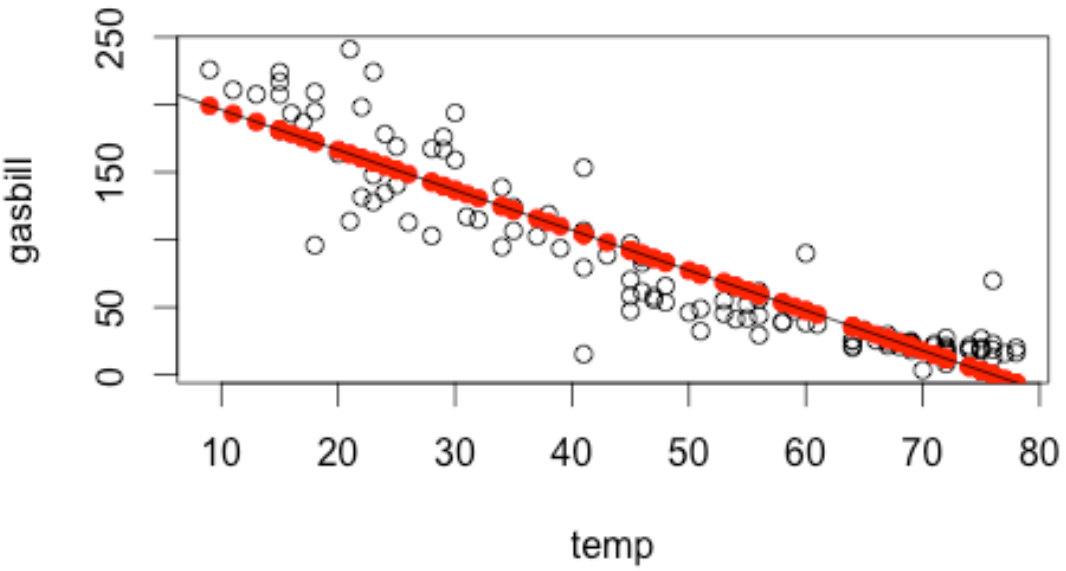


- Once you have plotted the logs of the variables, you can add a linear model. From there, you can predict on this log scale and then undo the transformation
  - After undoing the transformation, you add the predicted plots back to the original plot and end up with…..(for code, see R script for transformations)

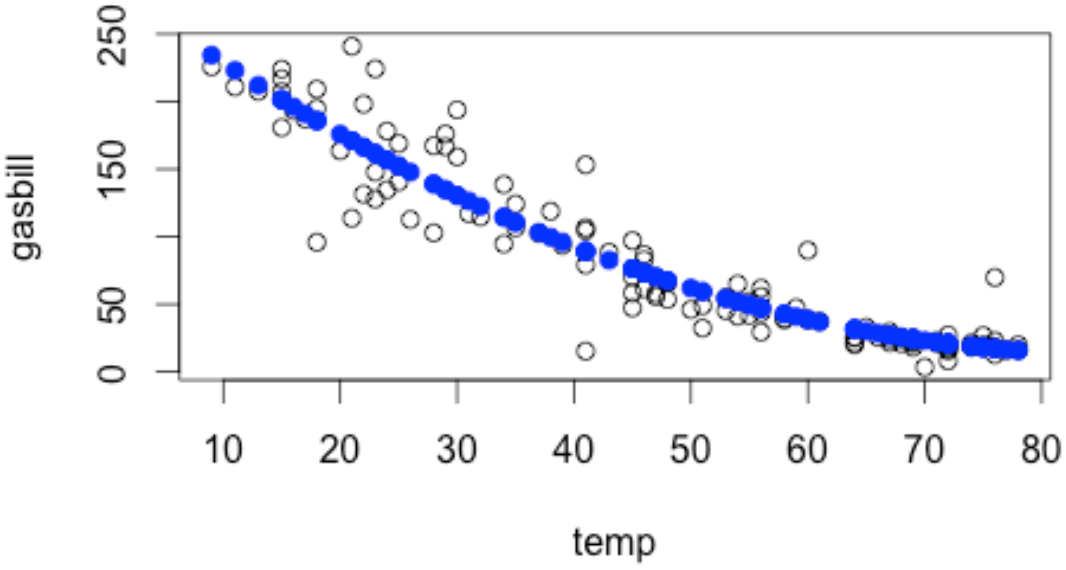


- **Note: just because we took the log of both variables this time does NOT mean you must always take the log of both variables. Sometimes, you will just have to take the log of one of the variables.

- Utilities.R
  - Open utilities.csv
  - Plot gasbill~temperature
  - Again a linear model does not look good



- Add in a quadratic model to raise the predictor variable to a power (in this case to 2). Then replot the data and add the fitted values
  - For specific code, see the utilities.R script



- In this example, we can guess to use a quadratic model because the data plot has a <u>natural "smile" shape</u>

- Interpretation of Power Law (we did some algebra today)
  - Our original equation is $y_i=B_o+B_1x_i+e_i$
  - We take the log of both sides: $\log y_i= B_o+B_1 [\log x_i]+e_i$
  - Raise it all to the e power: $e^{\log y_i}= e^{(B_o+B_1 [\log x_i]+e_i)}$

- Simplify: $y_i=(e^{\wedge} B_o)(x_i^{\wedge} B_1)(e^{\wedge} e_i)$
- This proof proves that x is proportional to y to a power. Therefore, we CAN fit a linear model to the log of x & y
- **Note: The same proof can be done if you only take the log of x or if you only take the log of y