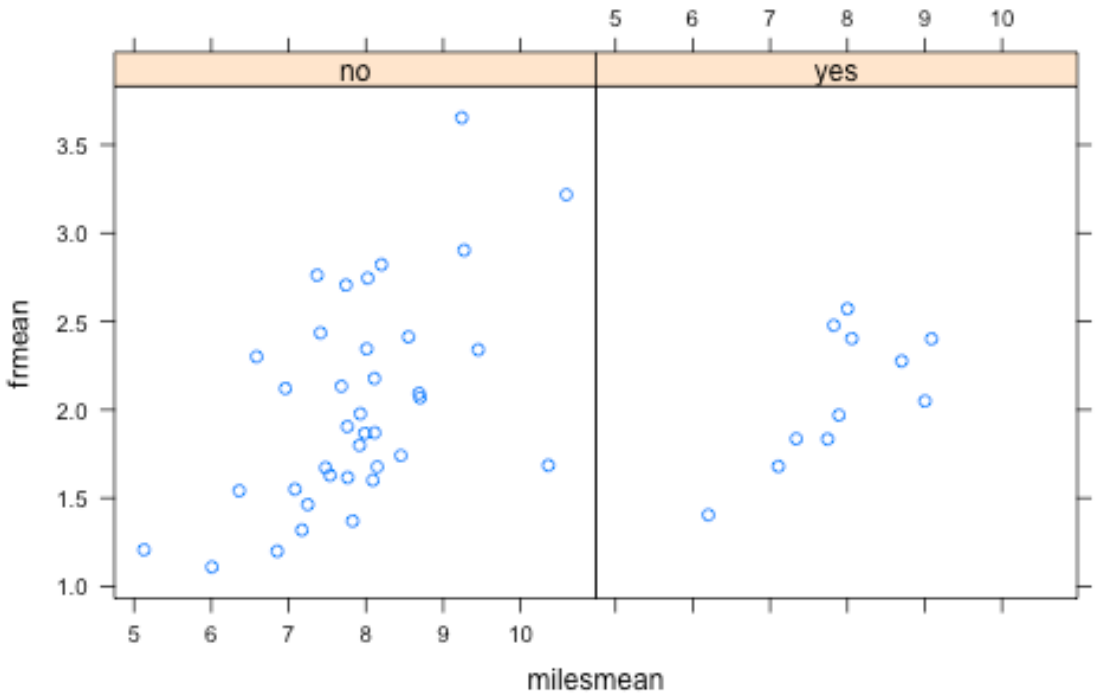
Sabeeha Islam
Cynthia Ramirez
9:30 AM-11:00 AM

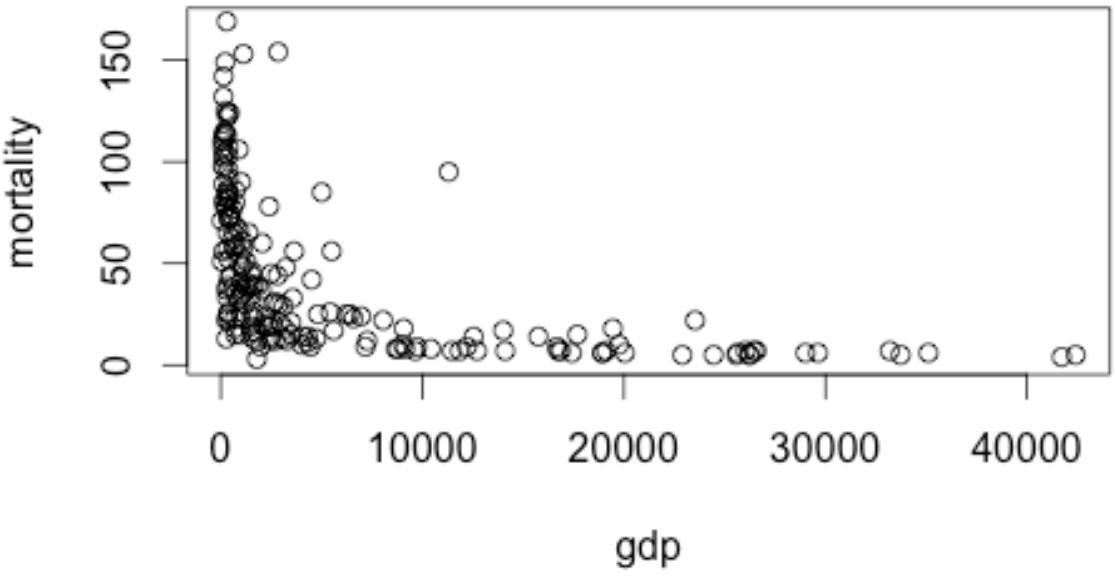
STA 371H Scribe Notes 1/27/14

A) **Homework Review** for afc.R and afc.csv Files
- o1A) Plot Price~Food
- o1B) Use the dotplot command. Compare to dotplots: the first modeling Price~FoodScore and the second Price~FeelScore
- o1C) 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 furthest away from the lines in either direction were the "least-valuable"
- o2A) Plot Stock~S&P500 and find a linear model and extract the coefficients. You can plot the residuals and then use "favstats(resid(model))", assuming you have already saved the linear model.
- o2B) Here we were asked to interpret the y-intercept. Since we know what the y-intercept is when x=0, we know the y-intercept will tell us where the stock market will be when the market is at 0. When the y-intercept is positive, it means that on average, the stocks perform better than the market.
- o2C) Yahoo!'s beta score is their estimation for the slope. We can assume that the scores are different due to a variety of factors, including the possibility that different years may have been used or different data taken into account. Our data in the file was only based on information from 2011-2012.
- o2D) There are two possible ways of approaching this problem. You can calculate the coefficient of determination (r-squared) or you can compare 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.

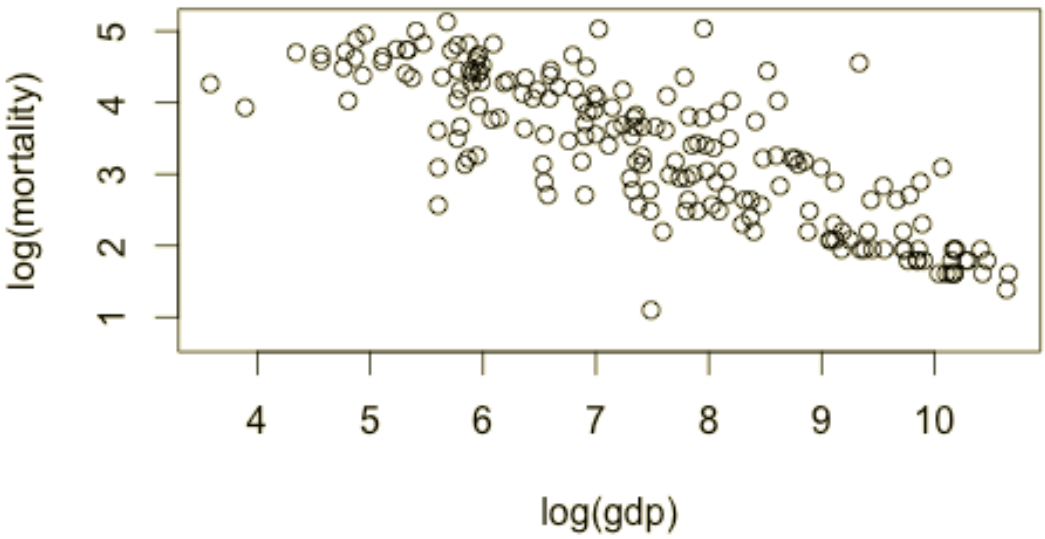
B) **Combining and Transforming Variables**: Trafficdeaths Script Data set.
How to generate new variables from old ones.
- Open both the trafficdeaths script and csv files: fips & trafficdeaths
- We learned to merge files (fips.csv into trafficdeaths.csv)
  - traffic2 = merge(trafficdeaths, fips, by.x = "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
- How to Add Text Labels on Graphs
  - For R code, see trafficdeaths.R lines 28-32
- Learned how to stratify the aggregated variables by another 3rd variable
  - Xyplot-stratify(lattice plot)
  - In a **lattice graph**, in each panel you have the same two variables stratified by a third variable. Here it is stratified by jail sentence. This allows for multivariable thinking.



- 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 so it cannot be explained with typical linear function.

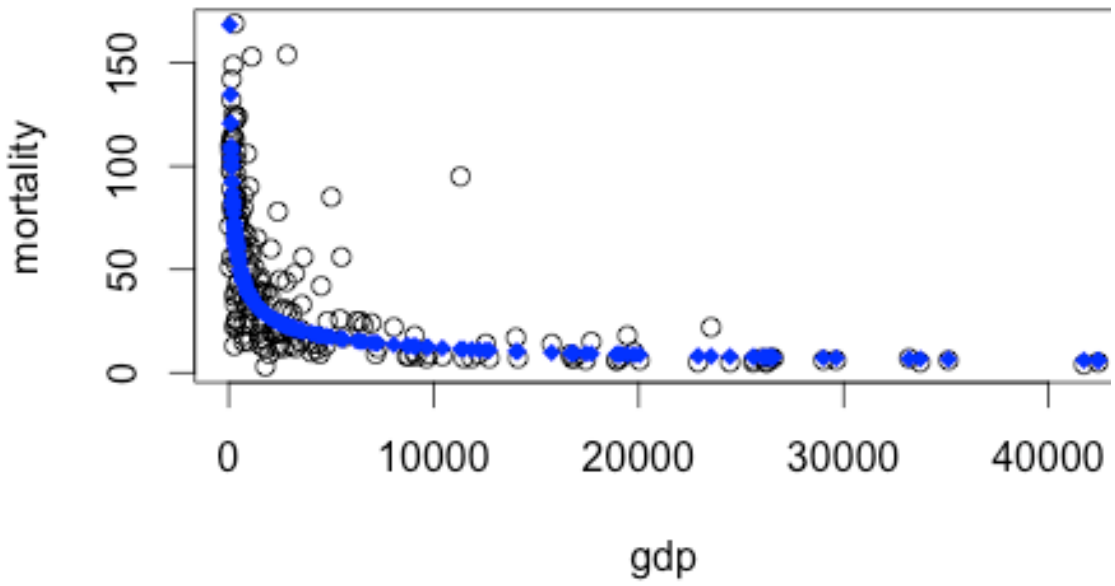


- **Log Function**- "Unsquishing" Consider using the Log Function when you notice that data is "squished" at low levels of GDP, but there are long tails on the x and y axes (both directions)
- Here, we would take the log of both the x and y variables because there are long "tails" on both ends.



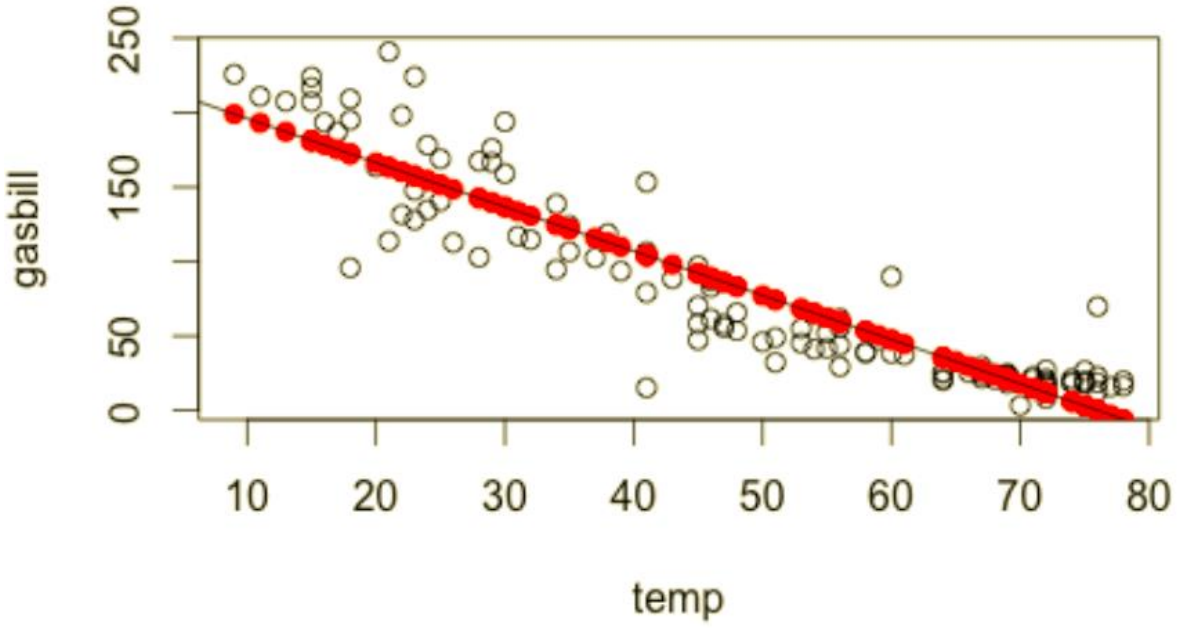
Once you have plotted the logs of the variables, you can add a linear model. From there, you can predict on this logical scale and then undo the transformation.

- After undoing the transformation, you add the predicted plots back to the original plot and end up with ….(See R code in lines 20-22 of Transformations.R)
- Conceptually, to get back to the regular values you would take the whole thing on each side and use e^ function.
- Power Law- We proved that **logy and logx follow a linear relationship.**

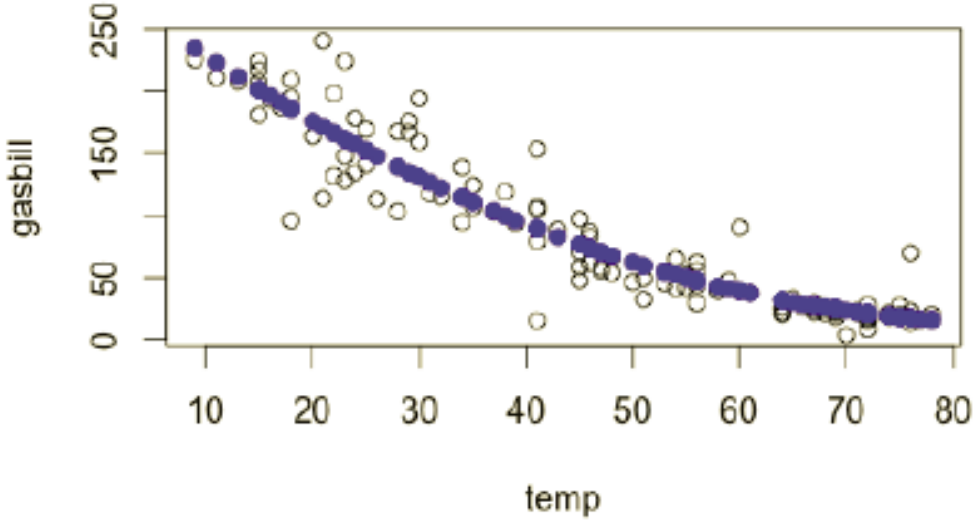


- **Note: just because we took the log of both variables this time does NOT mean you should always take the log of both variables. Sometimes, you will just have to take the log of one of the variables.

- Utilities.R
  - Open utilities.csv
  - Plot gasbill~temperature
  - Linear model does not look good
- You can add the curve directly onto the plot using the curve function. (See Line 41 in R script)



- Add in the quadratic equation 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 that we should use a quadratic function because the data plot has a <u>natural "smile" shape and pattern.</u>

- You DO NOT want there to be a pattern when you plot the residuals, as this would indicate that there is still some "x-ness" left in y.

Interpretation of Power Law (Algebra proof done in class)

- Original equation is $y_i=B_o+B_1x_i+e_i$
- Take the log of both sides: $\log y_i=B_o+B_1[\log x_i]+e_i$ (This approach is less common)
- Raise it all to the e power: $e^{(B_o+B_1[\log x_i]+e_i)}$
- Simplify: $y_i=(e^{B_o})(x_i^{B_1})(e^{e_i})$
- This proves that x is proportional to y to a power. Therefore, we CAN **fit a linear model to the log of x and y.**
- **Note: The same proof can be done if you only take the log of x or y

***You could plot the residuals again and again adding x^3, x^4, etc. You can add as many powers of x as you would like. However, then you can run into the issue of OVERPLOTTING. See Line 22 in R Script.* "There is a trade-off between fit and simplicity." – Dr. Scott

In conclusion, we learned 3 Key Skills:

1. How to combine old variables to generate new variables.
2. When and how to take the log and extract coefficients.
3. How to fit quadratic models.