Niket Parikh; np6896

January 27 Scribing

# Went over HW 2

1. **Problem Number 1**
    a. **Question A: Find the most expensive neighborhoods to eat at on average**
        i. Used the mean function to find group means of neighborhoods and prices
            1. Mean(Price~Neighborhood, data = afc)
        ii. The cheapest was the Drag and Bouldin Creek
    b. **Question B**
        i. Which predicts price of meal better, food score or Feel Score
            1. Found the coefficient of the scatter plots of price vs FoodScore and price vs FeelScore
            2. FoodScore predicts better
    c. **Question C**
        i. Find the best value neighborhood
        ii. Graph the residuals of Food Score vs Price to its line of best fit
        iii. Find the average residual for each neighborhood.
        iv. Code: model1= lm(Price~FoodScore, data=afc)
        mean(resid(model1)~Neighborhood, data=afc
        v. Two best value neighborhood's were Hyde Park and the Drag (Lowest residual), the worst value neighborhoods were Congress and the Convention Center (highest residual).
2. **Problem 2**
    a. Question A: Do a regression of each stock vs the SP500, find the intercept, slope and standard Deviation of the residuals

    Code: plot(AAPL~SP500, data=marketmodel)

    model1 = lm(AAPL~SP500, data=marketmodel)

    sd(resid(model1))

    coef(model1)

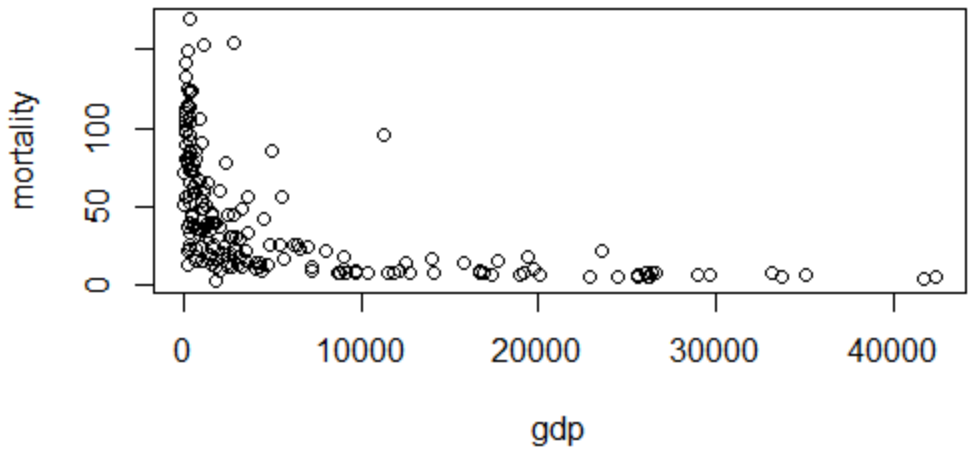
    You can make 2 arguments for which stock is most tightly coupled to the movements of the wider market. First, JNJ because its residual standard deviation is the smallest or second GOOG because its slope is closest to 1

    ***In future we look at R^2 because it combines the two numbers; standard deviation and Slope***

b. **Question B**: The intercepts are very close to 0 and mostly small since the stocks tend to move similarly to the S&P 500 with the exception of the Apple stock. According to the intercept, Apple is outperforming the market by over 10%.
c. **Question C**: Yahoo states that Wal-Mart's beta is 0.33, and a beta below 1 means that the stock is not as volatile as the market. This means that the stock will follow the trend of the market but not as severely. This makes sense because Wal-Mart has the lowest slope of 0.47; therefore, the stock is moving less than the movement of the stock market. In conclusion, for every 1% that the market increases, Wal-Mart only gains 0.47%.
d. **Question D:** This question can go either way, the question asks if WMT is most closely related to TGT. You could make the argument that it is more closely to JNJ by looking at their slope, intercept and Residual Standard Deviation
   OR
   This statement is correct because when we plot TGT with WMT, we find that the slope is 0.85, which means that for every 1% that Wal-Mart fluctuates, Target moves in the same direction by 0.85%. These two companies are both in the large inexpensive retail industry; therefore, it would make sense that their stock prices fluctuate by the same percentage.

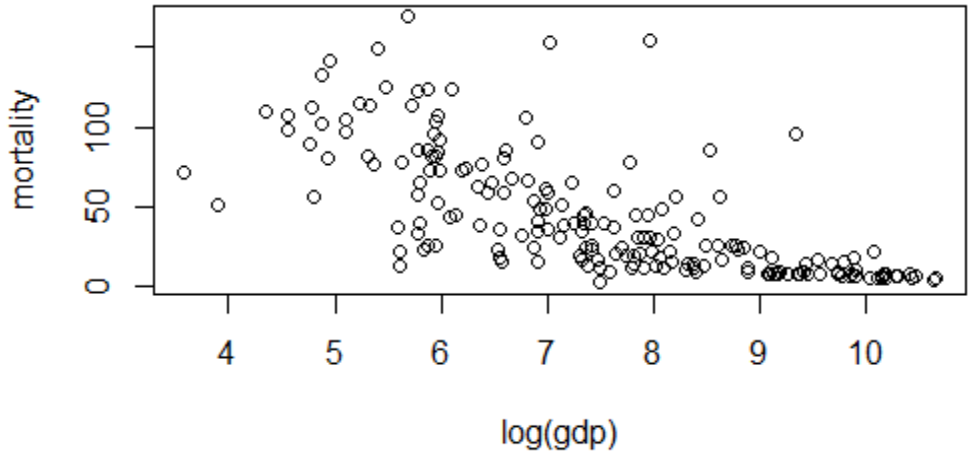
## Class

1. **Trafficdeaths.R (Learning to generate new variables)**
   a. Merge 2 data sets, Traffic Deaths and Fips to cross reference and create new variables using both data
   b. Code: traffic2 = merge(trafficdeaths, fips, by.x = "state", by.y="fipsnum")
      i. *Fipsnum* and *State* both have the same correlating data in the that's the connector
   c. Next, you define new variables that aggregate a state's statistics across years
      i. frmean = mean(mrall~fipsalpha, data=traffic2)
         1. new variable frmean is the average mortality rate by state (mrall = mortality rate, fipsalpha= state)
2. **Transformations.R (Transforming variables to fit logarithmic, exponential, and power-law relationships and Non Linear Curve Fitting)**
   a. Infmort Data File
   b. To see what type of line to fit, try to make scatter plot linear, plot log of X variable vs Y, if that doesn't look linear, plot log of X variably by Log of Y Variable.
      i. In example we did log of both examples
      ii.



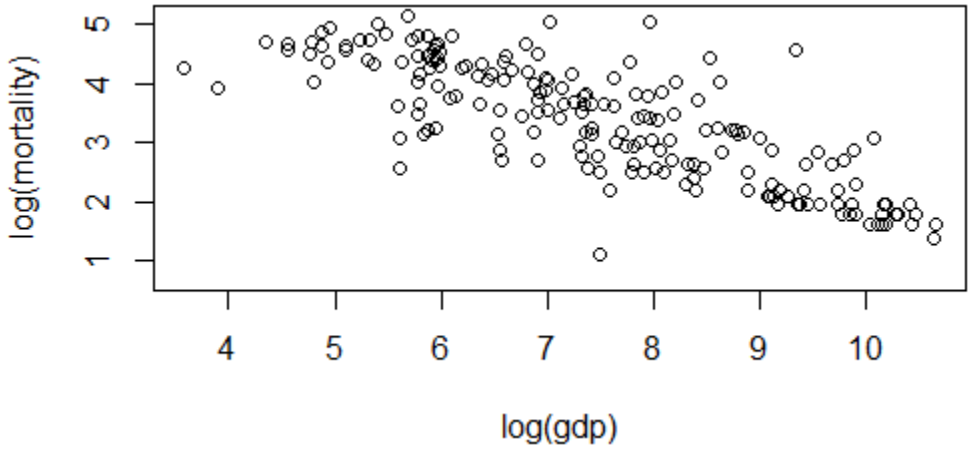

iii.

iv. Normal Graph^^




v.

vi. Log of X Var ^^




vii.

viii. Log of X and Y Var ^^^-most linear

c. Why take Logs?

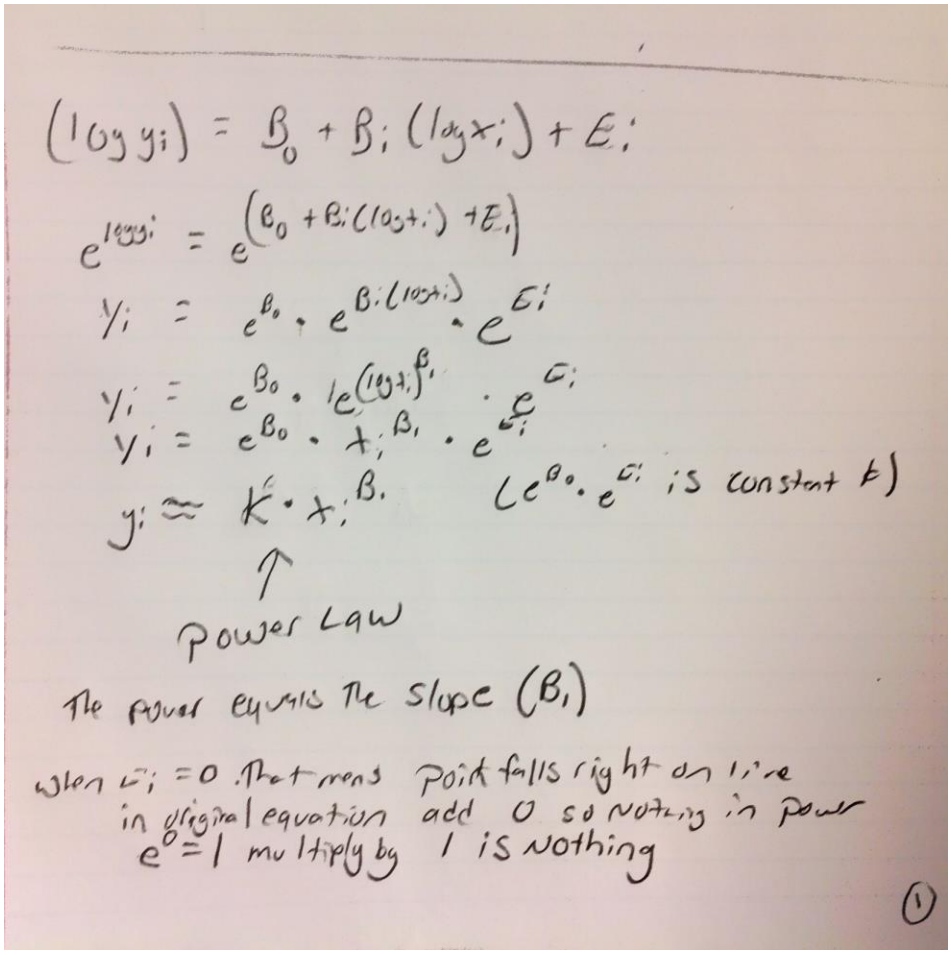
i.

| .1 | 1 | 10 | 100 | 1000 |
|---|---|---|---|---|
| $10^{-1}$ | $10^{0}$ | $10^{1}$ | $10^{2}$ | $10^{3}$ |
| -1 | 0 | 1 | 2 | 3 |

ii. "Unsquishes" the data, you use log when a lot of data is squished up against the left axis and then a long right tail, puts the on a similar scale

d. How to interpret these log graphs
   i. $Log_{yi} = B_0 + B_1(Log_{Xi}) + e_i$

e. Now you have it linear on a log scale so you have to undo the log scale by exponentiating the whole thing.
   i.




f. Code to do this

```
# Plot the curve on the original scale
plot(mortality ~ gdp, data= infmort)
beta = coef(lm1)

# Predict on the log scale, and then undo the transformation
logmort.pred = beta[1] + beta[2]*log(infmort$gdp)
mort.pred = exp(logmort.pred)
```

g. Now you can plot the original plot with a log line of best fit

```
plot(mortality ~ gdp, data= infmort)
points(mort.pred ~ gdp, data=infmort, col='blue', pch=18)
```

**3. Utilities R Script and Utilities Data file (Quadratic fit)**

a. When you plot the residuals of a linear model it looks systematic which means back to the drawing board to fit a new model
b. Try a quadratic model: Code:
   lm2=lm(gasbill ~ temp + I(temp^2), data=utilities)
   plot(gasbill ~ temp, data=utilities)
   points(fitted(lm2)~temp, data=utilities, col='blue', pch=19)

Last we Opened the Milk Excel file and tried to determine how much to charge for milk, we did this on our own and we will discuss it in more depth next class.