Peter Feng

2/2/2015 Notes

Class Updates:

-Be comfortable with basic R skills.

Traffic Case Study code

/ start up library **/**

```
library(mosaic)
```

/ load up the files **/**

```
trafficdeaths <- read.csv("YourDirectory/trafficdeaths.csv")

View(trafficdeaths)

fips <- read.csv("YourDirectory/fips.csv")

View(fips)
```

/ Merge the data. Join by columns state from trafficdeaths.csv and fipsnum from fips**/**

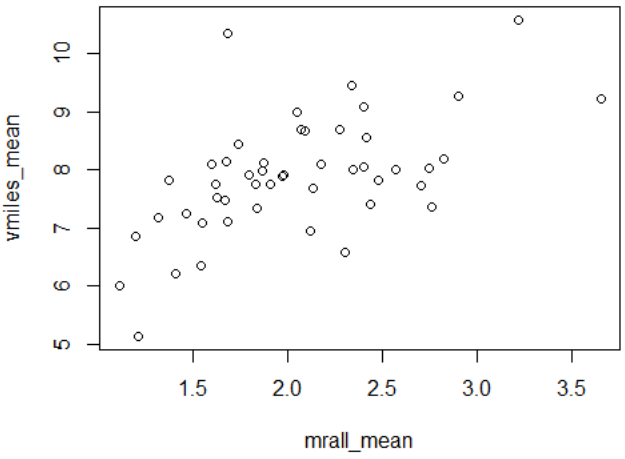
```
traffic2 = merge(trafficdeaths, fips, by.x = "state", by.y="fipsnum")
```

/ Find the means stratified by state for the deaths per state**/**

```
mrall_mean <- mean(mrall~fipsalpha,data=traffic2)
vmiles_mean <- mean(vmiles~fipsalpha,data=traffic2)
perinc_mean <-mean(perinc ~fipsalpha,data=traffic2)
```

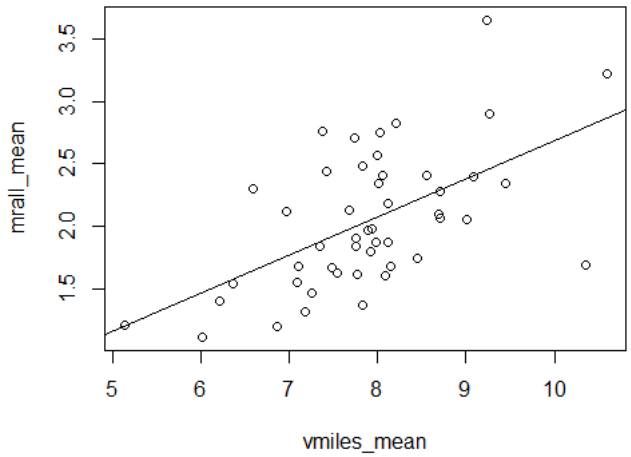
/ Make a scatter plot and fit a linear models. Fatality rate vs Miles per Driver **/**

```
plot(mrall_mean~vmiles_mean)
```



/ Plot a linear fit**/**

abline(lm(mrall_mean~vmiles_mean))

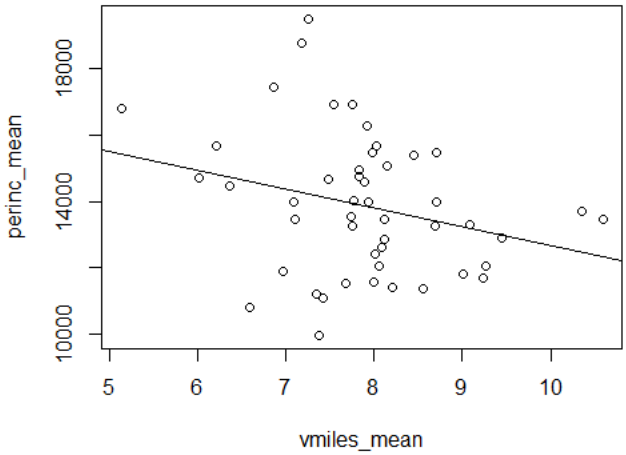


/ Scatter plot Per capita income vs miles per driver**/**

plot(perinc_mean~vmiles_mean)

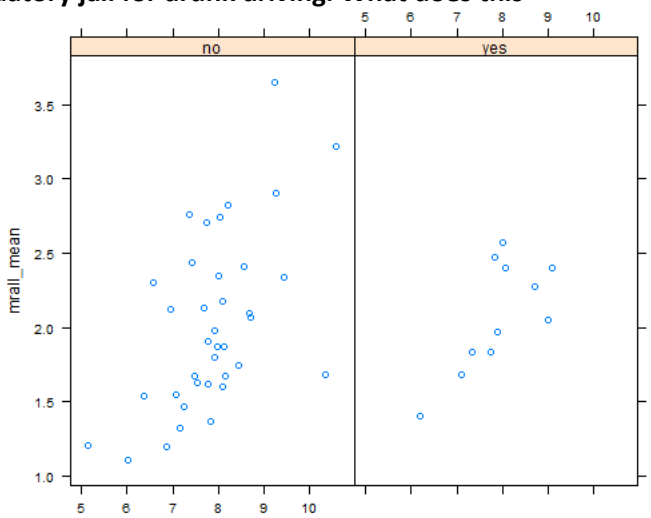
/ Plot a linear fit**/**

abline(lm(perinc_mean~vmiles_mean))



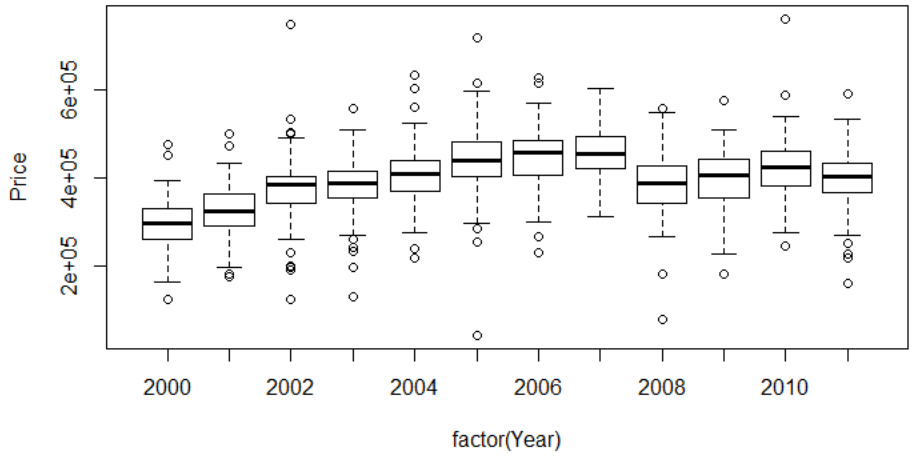
/ Make a lattice plot that stratifies the scatter plot of fatality rate versus miles driven by the categorical variable indicating whether the state has mandatory jail for drunk driving. What does this plot suggest? **/**

xyplot(mrall_mean ~vmiles_mean | jaild, data = traffic2)



Homework stuff:

1. Describe the within and between group variability. Overall the trend shows in increase despite occasional dips



2. How was the most expensive/cheapest food/price-weighted restaurants found?

Method 1

plot a linear model through the data points of mean food score per neighborhood vs mean food price per neighborhood. Find the residuals of the data points. The two greatest residuals represent the 2 most overpriced restaurants. The two lowest residuals represents the two cheapest restaurants according to the linear model.

Method 2

Divide mean price by neighborhood by mean food score per neighborhood to calculate Dollars paid per Food Score metric and create a new column. Sort the data to find restaurants with highest dollars paid per food score and lowest dollars paid per food score.

3. What do the intercepts mean?

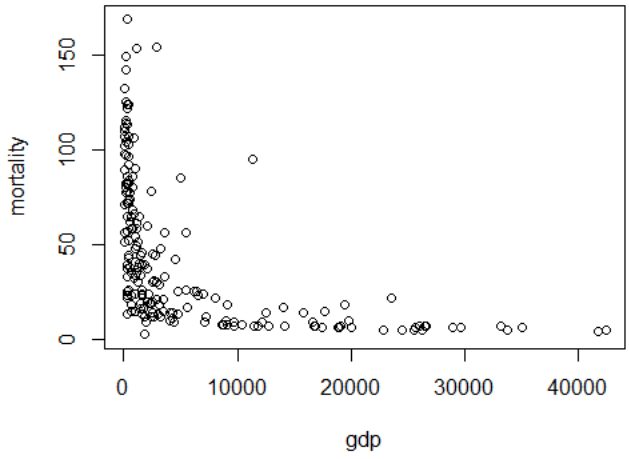
On a day of 0 returns, how much does the corresponding stock return according to the model.

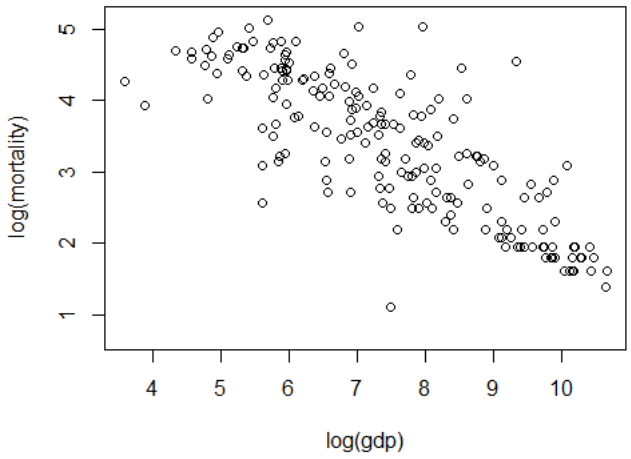
Infant Case:

Key is to scale the data using logs in order to create a cleaner linear fit.

/** Import the data, view summary, plot mortality vs gdp, adjust using logarithm**/

```
infmort <- read.csv("C:/Users/Jenny Feng/Downloads/infmort.csv")
  View(infmort)
summary(infmort)
          country      mortality           gdp
Afghanistan   :  1   Min.   :  2.00   Min.   :   36
Albania       :  1   1st Qu.: 12.00   1st Qu.:  442
Algeria       :  1   Median : 30.00   Median : 1779
American.Samoa:  1   Mean   : 43.48   Mean   : 6262
Andorra       :  1   3rd Qu.: 66.00   3rd Qu.: 7272
Angola        :  1   Max.   :169.00   Max.   :42416
(Other)       :201   NA's   :6        NA's   :10
plot(mortality ~ gdp,data=infmort)
plot(mortality ~ log(gdp),data=infmort)
plot(log(mortality) ~ log(gdp),data=infmort)
```





Remember to convert back to normal values when using the linear model with the log scale.

```
> lm1 = lm(log(mortality) ~ log(gdp),data=infmort)
> abline(lm1)
> new_gdp = 1100
> coef(lm1)
(Intercept)    log(gdp)
  7.0452008  -0.4932026
> log_infmort = 7.0452008 +(-.4932026)*log(new_gdp)
> log_infmort
[1] 3.591271
> y_hat = exp(log_infmort)
> y_hat
[1] 36.28015
```

Equation conversion