FTAP Stats HW 6

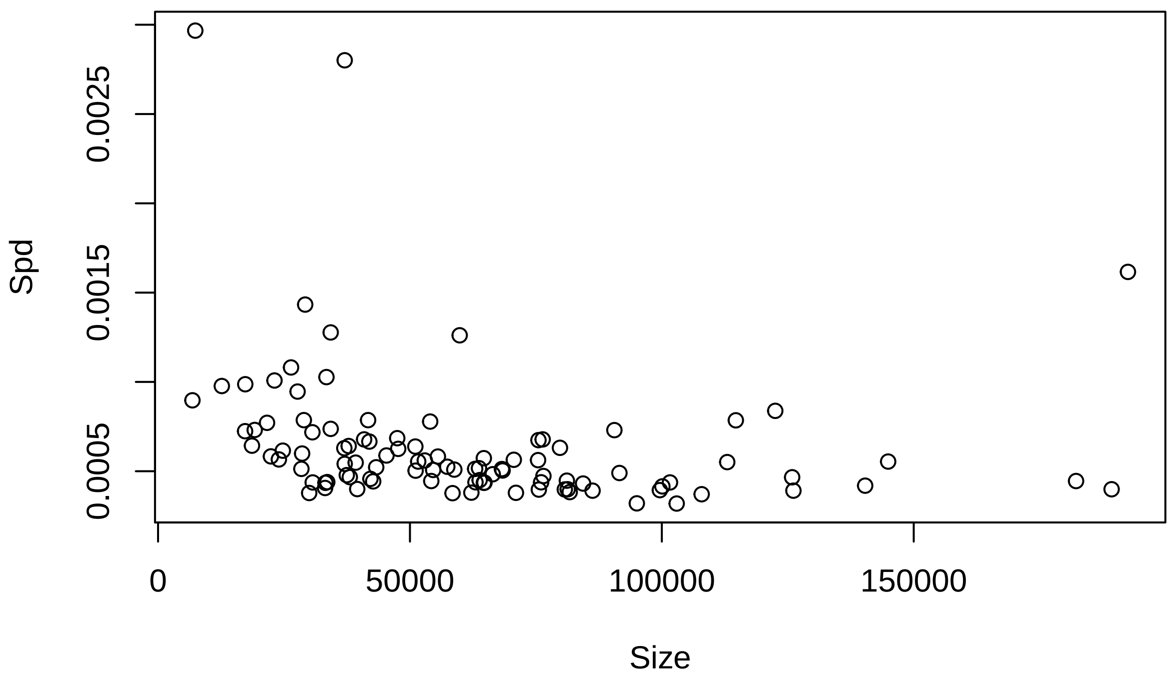
Zachary Fogelson

July 20, 2015

Problem 1

a

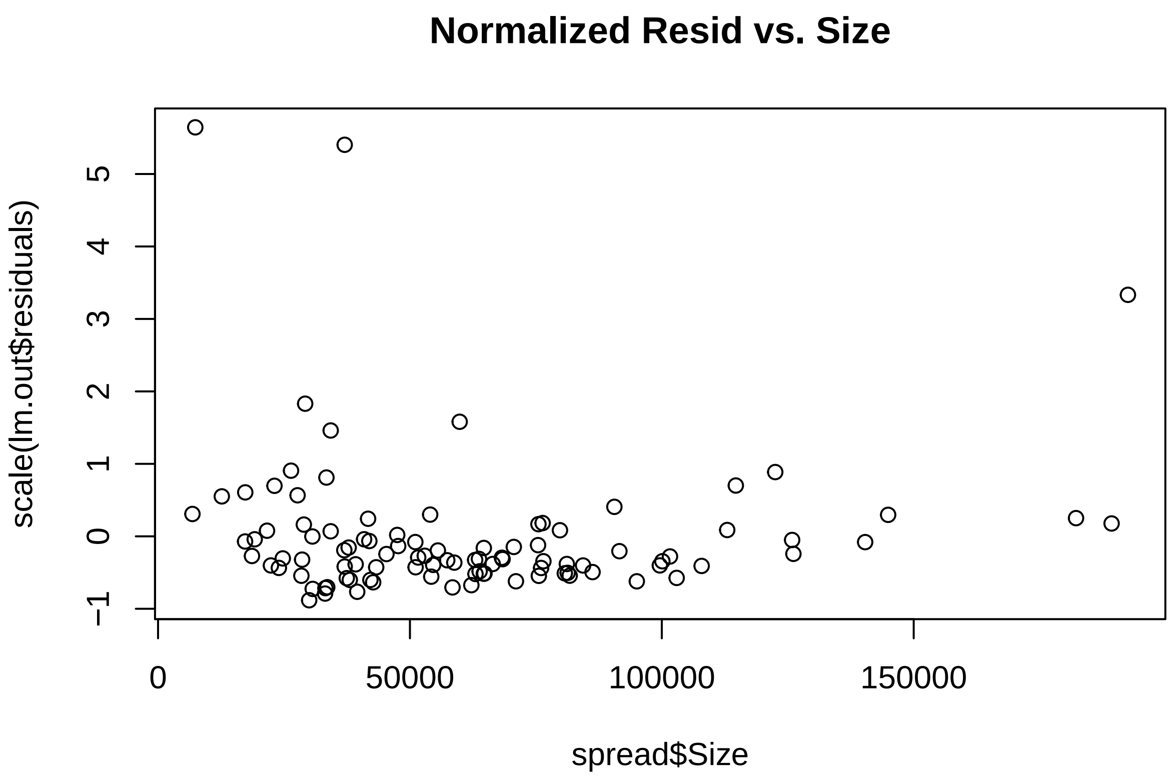
```
spread <- read.xls("spread.xls")
plot(Spd ~ Size, spread)
```



The original data does not appear linear in X because the vast majority of the data is in the lower left quadrent.

b

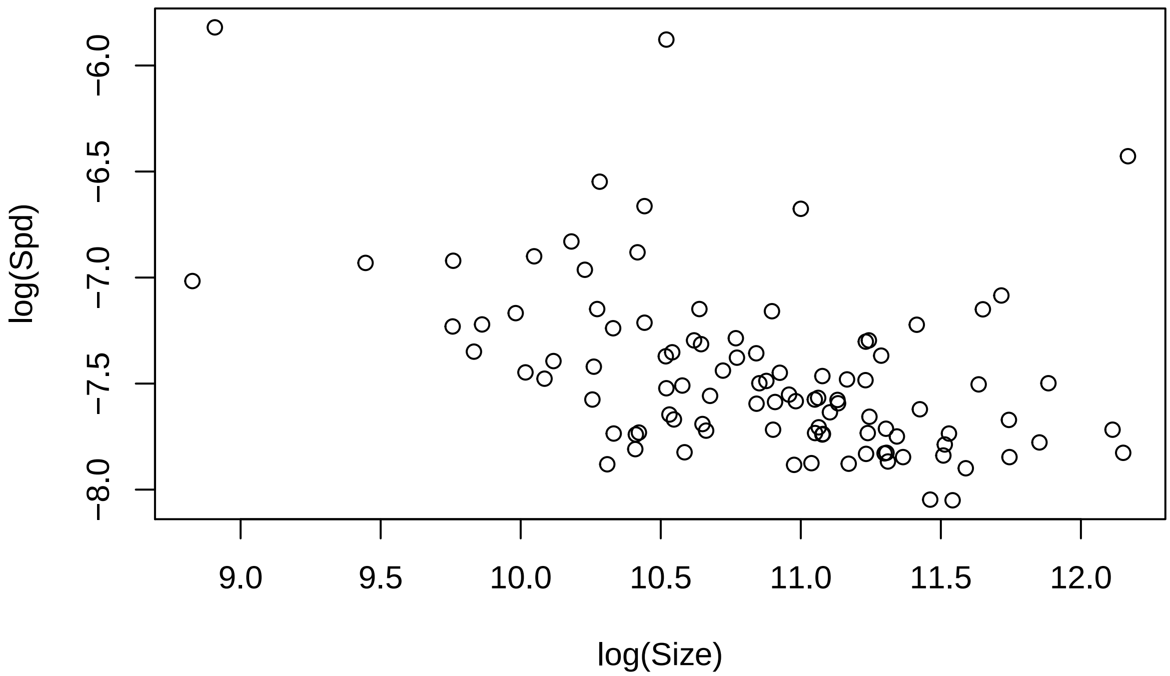
```
lm.out <- lm(Spd ~ Size, spread)
plot(spread$Size, scale(lm.out$residuals), main = "Normalized Resid vs. Size")
```



The residuals do not look iid normal in x. The residuals appear to be distrobuted in the exact same way as the original data!

c

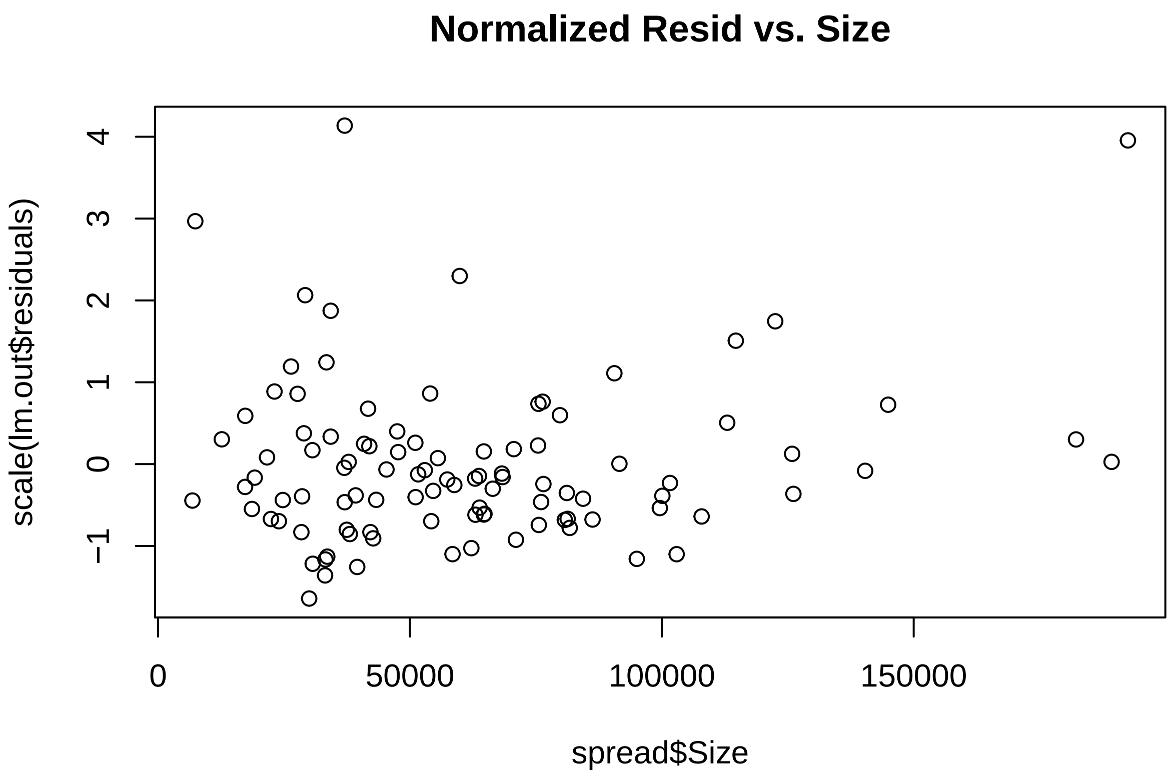
```
plot(log(Spd) ~ log(Size), spread)
```



This data appears more more linear and possibly normally distributed.

d

```r
lm.out <- lm(log(Spd) ~ log(Size), spread)
plot(spread$Size, scale(lm.out$residuals), main = "Normalized Resid vs. Size")
```



This data looks more linear but the values still do not look normally distributed because of the concentration of residuals in the lower left along with the number of outliers which are >4 standard deviations away.

e

```r
lm.out <- lm(log(Spd) ~ log(Size) + log(Trd) + log(Num) + log(Turn) + log(Vol), spread)

summary(lm.out)
```

```
##
## Call:
## lm(formula = log(Spd) ~ log(Size) + log(Trd) + log(Num) + log(Turn) +
##     log(Vol), data = spread)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28878 -0.10493 -0.01636  0.07060  0.39509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.79697    0.45227  -1.762  0.08130 .
## log(Size)   -0.13953    0.02399  -5.816 8.25e-08 ***
## log(Trd)    -0.16832    0.03551  -4.740 7.57e-06 ***
## log(Num)    -0.01587    0.04814  -0.330  0.74232
```

```
## log(Turn)   -0.10158    0.03234  -3.141  0.00225 **
## log(Vol)     1.02517    0.05322  19.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1419 on 94 degrees of freedom
## Multiple R-squared:  0.8826, Adjusted R-squared:  0.8763
## F-statistic: 141.3 on 5 and 94 DF,  p-value: < 2.2e-16
```

(assumption: The problem set asks for log(sd) does this mean the square root of volatility? I decided to assume it was the same as volatility.)

Based on the p-values of the coefficients, log(Size), log(Trd), and log(Vol), we are confident at the 99.9% level that they are significant in predicting the log(spd). For log(Turn) we reject the null hypothesis that it is uncorrelated with log(Std) at the 99% level. However, we fail to reject the null hypothesis that log(Num) is correlated with log(Spd)

f

```r
1 - pf(summary(lm.out)$fstatistic[1], summary(lm.out)$fstatistic[2], summary(lm.out)$fstatistic[3])
```

```
## value
##     0
```

Because the p-value of the F-statistic is well below .001, we reject the null hypothesis that none of the variables are correlated with the log(Spd) at a 99.9% confidence level.

g

```r
lm.outWo <- lm(log(Spd) ~ log(Size) + log(Trd) + log(Turn) + log(Vol), spread)
cat("With Num\n", "R^2: ", summary(lm.out)$r.squared, ", Adj R^2: ", summary(lm.out)$adj.r.squared, "\n
```

```
## With Num
##  R^2:  0.8825876 , Adj R^2:  0.8763423
```

```r
cat("Without Num\n",  "R^2: ", summary(lm.outWo)$r.squared, ", Adj R^2: ", summary(lm.outWo)$adj.r.squa
```

```
## Without Num
##  R^2:  0.8824518 , Adj R^2:  0.8775024
```

Based, on the two tests the R^2 and adjusted R^2's are remarkably similar.

h

```r
tester <- data.frame(Size=exp(10.5), Turn=exp(-1.1), Trd=exp(7.6), Vol=exp(-3.5))

pr <- predict(lm.outWo,tester, interval = "prediction")

cat("Expected = ", exp(pr[1]))
```

```
## Expected =  0.0008582373
```

i

```r
cat("PI = (", exp(pr[2]), ", ", exp(pr[3]), ")")
```

```
## PI = ( 0.0006462183 ,  0.001139818 )
```

j

$$f = \frac{\Delta R^2/q}{(1 - R^2_{full}/(n-k-1))}$$

```r
lm.rest <- lm(log(Spd) ~ log(Size) + log(Trd) + log(Vol), spread)
lm.unrest <-  lm(log(Spd) ~ log(Size) + log(Trd) + log(Vol) + log(Turn) + log(Num), spread)

fstat <- ((summary(lm.unrest)$r.squared-summary(lm.rest)$r.squared) / 2) / ((1 - summary(lm.unrest)$r.s

1 - pf(fstat, 2, length(spread$Spd) - 6 - 1)
```

```
## [1] 0.008727642
```

Because the pvalue of the f-statistic is less than .01 we reject the null hypothsis that the value for the coefficient of the log of turnover and the log of number of analysts are both 0 with 99% confidence.

k

```r
sumCol <- (log(spread$Size) + log(spread$Trd))
lm.rest <- lm(log(Spd) ~  sumCol + log(Vol) + log(Turn) + log(Num), spread)
lm.unrest <-  lm(log(Spd) ~ log(Size) + log(Trd) + log(Vol) + log(Turn) + log(Num), spread)

fstat <- ((summary(lm.unrest)$r.squared-summary(lm.rest)$r.squared) / 1) / ((1 - summary(lm.unrest)$r.s

1 - pf(fstat, 1, length(spread$Spd) - 6 - 1)
```

```
## [1] 0.5084317
```

Because the pvalue of the f-statistic is greater than .05 we fail to reject the null hypothsis that the value for the coefficients for log of size and trd are the same.