# STA302 Assignment #2

Rui Qiu
#999292509

Last edited on November 25th

6.9

6.9.1 Solution:

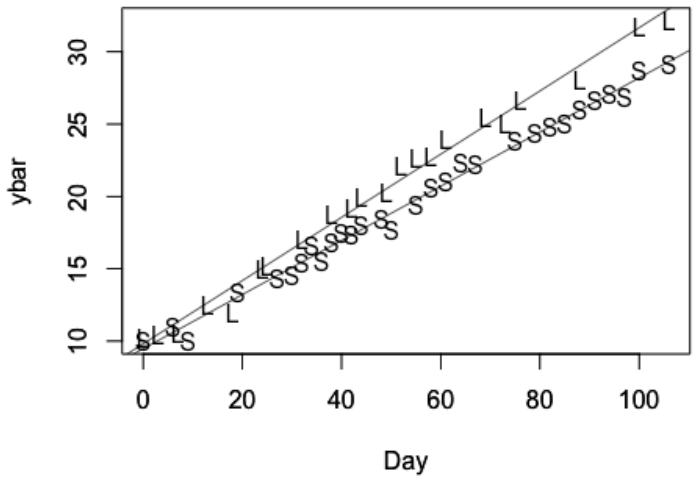
```
        df       SS        pe
long   167 255.1215 1.5276737
short  292 246.7392 0.8449973
pooled 459 374.3000 0.8154683
```

The mean square for pure error of long shoots is 1.5277, the mean square for pure error of short shoots is 0.8450. So the pure error estimate of variance for longs shoots is about the twice of e estimate for short shoots.

Let F = 1.5276737 / 0.8449973 = 1.807904, then 2*(1-pf(F, 167, 292) = 1.029662e-05, the significance level is close to zero. Hence we reject the null hypothesis. As a result, the variances are not equal.

Under the assumption that variances are sigma^2 and 2sigma^2 respectively,
so
the sum of square of long shoots is SSpeLong = 2sigma^2 * 167 = 255.1215
the sum of square of short shoots is SSpeShort = sigma^2 * 292 = 246.7392
the sum of square of pooled estimate is SSpePooled = sigma^2 * (167 + 292) = (1 / 2) * SSpeLong + SSpeShort = 374.3
Then the pooled pure-error estimate of sigma^2 is
sigma^2pePooled = SSpePooled / (167 + 292) = 0.8154684

6.9.2 Solution:



According to the plot, those two linear models are plausible (with L representing long shoots and S representing short shoots), since the two lines don't overlap, the two types of shoots are different.

6.9.3 Solution:

According to problem 6.9.1, the variance of ybar for long shoots is 2(sigma^2)/n, while for short shoots is (sigma^2)n. Then the weight for long is n/2, the weight for short is n. Here, Model 1(most general model) is model 1 in 6.2.2, Model 2 (common intercept) is model 3 in 6.2.2, Model 3 (Coincident regression lines) is model 4 in 6.2.2.

```
Analysis of Variance Table

Model 1: ybar ~ Day + Type + Type:Day
Model 2: ybar ~ Day + Type:Day
Model 3: ybar ~ Day
  Res.Df    RSS Df Sum of Sq        F Pr(>F)
1     48 165.07
2     49 171.24 -1     -6.17   1.7929 0.1869
3     50 929.06 -1   -757.82 220.3570 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8.5 See the hand-writing part in the Appendix. (on next page)

①

Let H be $n \times n$ simple regression

By what we learnt in simple regression $(X'X)^{-1} = \begin{pmatrix} \frac{\sum x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \frac{1}{SXX}$

Then by definition of $h_{ij}$, $h_{ij} = x_i'(X'X)^{-1}x_j$

$$= (1 \ x_i) \frac{1}{SXX} \begin{pmatrix} \frac{\sum x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$= \left( \frac{\sum x_i^2}{nSXX} - \frac{\bar{x}x_i}{SXX} \quad \frac{-\bar{x}}{SXX} + \frac{x_i}{SXX} \right) \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$= \frac{\sum x_i^2}{nSXX} - \frac{\bar{x}x_i}{SXX} + \frac{-\bar{x}x_j}{SXX} + \frac{x_i x_j}{SXX}$$

Let $j=i$ then, $h_{ii} = \frac{\sum x_i^2}{nSXX} - \frac{\bar{x}x_i + \bar{x}x_i - x_i^2}{SXX}$

$$= \frac{\sum x_i^2}{nSXX} - \frac{\bar{x}^2}{SXX} - \frac{\bar{x}x_i + \bar{x}x_i - x_i^2 - \bar{x}^2}{SXX}$$

$$= \frac{\sum x_i^2 - n\bar{x}^2}{nSXX} + \frac{(x_i - \bar{x})^2}{SXX}$$

$$= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}$$

② High $h_{ii}$ $\longrightarrow$ high $(x_i - \bar{x})^2$ $\longrightarrow$ some points in the scatterplot are very far from other points (outliers).

③ Suppose for X, $x_1 = 1, x_2 = x_3 = \cdots = x_n = 0$.

so $\bar{x} = \frac{1}{n}$. $SXX = \sum (x_i - \bar{x})^2 = (\frac{1}{n})^2 (n-1) + (1 - \frac{1}{n})^2$

$$= \frac{n-1}{n^2} + \frac{(n-1)^2}{n^2}$$

$$= \frac{n - 1 + n^2 - 2n + 1}{n^2}$$

$$= \frac{n^2 - n}{n^2}$$

$$= \frac{n-1}{n}$$

Hence $h = \frac{1}{n} + \frac{(1 - \frac{1}{n})^2}{1 - \frac{1}{n}} = 1$ Done

9.2 Solution:

note that p = 5 (we use p here instead of p' in the book, so that for coding convenience), n - p = 46

```
  r          D          ti
1 -2.9147602 0.5846591 -3.1927822
2 -2.3163746 0.2074525 -2.4376317
3 -1.7711013 0.1627659 -1.8147106
4  2.9546191 0.1601094  3.2465847
5 -0.9963719 0.1408527 -0.9962917
```

We note that the forth data, Wyoming has the largest residual and largest ti, so it is a suspected outlier. We are going to test this. Then since the Bonferroni-adjusted p-value is the p-value multiply by the sample size (which is 46 + 5 = 51 in this problem). We calculate the p-value by the following codes:

```
> 2*pt(-abs(3.2465847),46)
[1] 0.00218225
```

then the Bonferroni-adjusted p-value is:

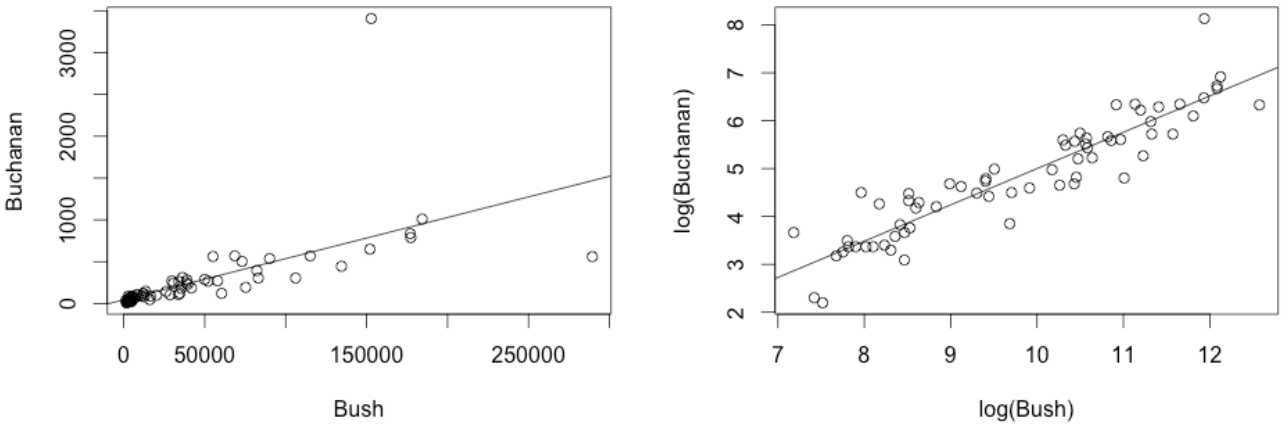
```
> 2*pt(-abs(3.2465847),46)*51
[1] 0.1112947
```

It is large, so Wyoming is not an outlier. Moreover, we compare the Cook Distance, and Alaska has the largest D, therefore, (deleting) it has the largest influence on the whole data.

9.8 Solution:

First of all, we are going to give two scatterplots of Buchanan versus Bush, both with original data and with log data. (codes can be referred in appendix). Note that the first model is required, and the second model is considered as the "better satisfy the assumptions of the simple linear regression model).
We find that in the first plot, a point obviously behaves as a 'suspicious outlier', which is Palm Beach County (above the line) with a very high number of votes for Buchanan; and similarly, in the right corner, the point is Dade County (below the line) with a very low number of votes for Buchanan. (we can do this through identify() function in R)

In details, we would like to do the outlier tests for both of the points, in both of cases, namely the original data and the log data.




Then according to the outlier tests, we find out that the data from Palm Beach County is an outlier while the data from Dade County is not.

Why? The following codes explain the reason:

For Palm Beach County and Dade County, the outlier in the first model, namely the linear model with original data, is:

```
> outlierTest(lm(Buchanan~Bush))
   rstudent unadjusted p-value Bonferonni p
50 24.08014         8.6246e-34   5.7785e-32
```

The outlier in the second model, namely the log linear model is:

```
> outlierTest(lm(log(Buchanan)~log(Bush)))
   rstudent unadjusted p-value Bonferonni p
50 4.066282         0.00013325    0.0089278
```

Therefore, in both cases, the results of outlierTest function are the same, that the data with index 50 is an outlier. That is to say, Palm Beach County is an outlier, while Dade County is not.

# Appendix

```
6.9

6.9.1
> data(allshoots)
```

```
> Day <- allshoots[,c(1)]
> n <- allshoots[,c(2)]
> ybar <- allshoots[,c(3)]
> SD <- allshoots[,c(4)]
> Type <- allshoots[,c(5)]
> long <- Type == 1
> short <- Type == 0
> pure.error <-
data.frame(df=c(sum(n[long]-1),sum(n[short]-1),sum(n-1)),SS=c(sum((n[l
ong]-1)*SD[long]^2),sum((n[short]-1)*SD[short]^2),sum((n[long]-1)*SD[l
ong]^2)/2 + sum((n[short]-1)*SD[short]^2)))
> pure.error
   df       SS
1 167 255.1215
2 292 246.7392
3 459 374.3000
> pure.error$pe <- pure.error$SS/pure.error$df
> row.names(pure.error) <- c('Long', 'Short', 'Pooled')
> pure.error
        df       SS        pe
Long   167 255.1215 1.5276737
Short  292 246.7392 0.8449973
Pooled 459 374.3000 0.8154683
> 1.5276737/0.8449973
[1] 1.807904
> 2*(1-pf(1.807904, 167, 292))
[1] 1.029662e-05
> 1/2*255.1215 + 246.7392
[1] 374.3
> 374.3/(167+292)
[1] 0.8154684
```

6.9.2

```
>
plot(Day[long],ybar[long],pch=c('L'),ylim=range(ybar[short],ybar[long]
),xlim=range(Day[short],Day[long]),xlab='Day',ylab='ybar')
> par(new=T)
>
plot(Day[short],ybar[short],pch=c('S'),ylim=range(ybar[short],ybar[lon
g]),xlim=range(Day[short],Day[long]),xlab='Day',ylab='ybar')
> par(new=F)
> abline(lm(ybar[long]~Day[long]))
> abline(lm(ybar[short]~Day[short]))
```

6.9.3

```
> pooledweights <- allshoots$n / 0.8154684
> m1 <- lm(ybar ~ Day + Type + Type:Day, weights=pooledweights)
> m3 <- lm(ybar ~ Day + Type:Day, weights=pooledweights)
> m4 <- lm(ybar ~ Day, weights=pooledweights)
> anova(m1,m3,m4)

9.2
> ehat <- c(-163.145, -137.599, -102.409, 183.499, -49.452)
> lev <- c(0.256, 0.162, 0.206, 0.084, 0.415)
> sig <- 64.891
> r <- ehat/(sig*sqrt(1-lev))
> D <- (1/5)*r^2*(lev/(1-lev))
> ti <- r*sqrt((46-1)/(46-r^2))
> data.frame(r, D, ti)

9.8
> names(florida)
[1] "County"   "Gore"     "Bush"     “Buchanan"
> County <- florida[,c(1)]
> Bush <- florida[,c(3)]
> Buchanan <- florida[,c(4)]
> plot(Bush, Buchanan)
> abline(lm(Buchanan~Bush))
> plot(log(Bush),log(Buchanan))
> abline(lm(log(Buchanan)~log(Bush)))
> identify(Bush, Buchanan, County)
[1] 13 50
```