Homework 5 Solutions

Jacob S Townson

March 26, 2018

1)

Before we begin, we must read in our data, and create our variables with the new data frame. Using these, we can answer the following parts of the question.

```
HeightData=read.table('D:/obewa/Documents/My real documents/University of Louisville/public_work/Classes
g=as.numeric(HeightData$gender=="female")
m=HeightData$mother
f=HeightData$father
y=HeightData$childHeight
X=cbind(g,g*m,g*f,1-g,(1-g)*m,(1-g)*f);dimnames(X)=list(NULL,c('x1','x2','x3','x4','x5','x6'))
X=data.frame(X)
```

The first 5 entries of the data frame X is given below. These values will be used to do the following problems.

a)

The MLE of β can be found using the following bit of code:

```
X=mutate(X, x7=x2+x5)
X=mutate(X, x8=x3+x6)
height.mod = lm(HeightData$childHeight~X$x1+X$x4+X$x7+X$x8+0)
```

Note, the reason we add $x_2 + x_5 = x_7$ and $x_3 + x_6 = x_8$ is so that we can make sure that the corresponding β for x_2 and x_5 are the same, and the corresponding β for x_3 and x_6 is the same. Now, from the above code, we find that the β's are the following:

```
coefs = summary(height.mod)$coefficients
kable(coefs)
```

	Estimate	Std. Error	t value	Pr(>|t|)
X$x1	16.5212399	2.7272041	6.057940	0
X$x4	21.7362293	2.7222301	7.984714	0
X$x7	0.3176101	0.0310004	10.245360	0
X$x8	0.3928433	0.0286768	13.698987	0

Thus the MLE is $\hat{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^{\mathrm{T}}$ where $\beta_1 = 16.5212399$, $\beta_2 = \beta_5 = 0.3176101$, $\beta_3 = \beta_6 = 0.3928433$, and $\beta_4 = 21.7362293$.

b)

Let's rename $\hat{\beta}$ as $\hat{\beta}_c$. If we use the following code, we can find the F-statistic as desired in this problem:

```
X=data.matrix(X)
X.reduced=cbind(X[,1],X[,2]+X[,5],X[,3]+X[,6],X[,4])
```

```
lm.reduced=lm(y~X.reduced+0)
lm.full=lm(y~X+0)
anova(lm.reduced,lm.full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ X.reduced + 0
## Model 2: y ~ X + 0
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    930 4357.9
## 2    928 4354.1  2    3.8228 0.4074 0.6655
```

Thus $F = .407$ as in the example in the notes.

c)

Here we want to find A, B, C, D, and E such that

$$A(\beta_2 - \beta_5)^2 + B(\beta_3 - \beta_6)^2 + C(\beta_2 - \beta_5)(\beta_3 - \beta_6) + D(\beta_2 - \beta_5) + E(\beta_3 - \beta_6) \leq 1$$

is a 95% confidence ellipse for $(\beta_2 - \beta_5, \beta_3 - \beta_6)^T$.

Using the following code below will help us get started here, giving us numerical values we will need to solve the problem:

```
C=rbind(c(0,1,0,0,-1,0),c(0,0,1,0,0,-1))
beta.hat=solve(t(X)%*%X)%*%t(X)%*%y
CB = C%*%beta.hat
M=solve(C%*%solve(t(X)%*%X)%*%t(C))
SSE=sum(y^2)-sum((X%*%beta.hat)^2)
n=nrow(X)
SSn=SSE/(n-6)
C
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    1    0    0   -1    0
## [2,]    0    0    1    0    0   -1
```

```
beta.hat
```

```
##          [,1]
## x1 18.8335828
## x2  0.3034821
## x3  0.3725423
## x4 19.3128130
## x5  0.3287734
## x6  0.4175562
```

```
CB
```

```
##             [,1]
## [1,] -0.02529128
## [2,] -0.04501389
```

```
M
```

```
##            [,1]       [,2]
## [1,] 1218.74004   80.29407
## [2,]   80.29407 1411.68049
```

```
SSE
```

```
## [1] 4354.052
```

```
n
```

```
## [1] 934
```

```
SSn
```

```
## [1] 4.691866
```

Now we must use the hint given to us in this problem to get a pivot. Note that $||\mathbf{y} - \mathbf{X}\hat{\beta}||^2/(n-6) = SSE/(n-6)$.

After some manipulation, we find that

$$\frac{3.801 + 1218.740(\beta_2 - \beta_5)^2 + 1411.680(\beta_3 - \beta_6)^2 + 160.588(\beta_2 - \beta_5)(\beta_3 - \beta_6) + 68.164(\beta_2 - \beta_5) + 131.066(\beta_3 - \beta_6)}{9.384} F_{2,928}$$

The critical point can be found in R by the following code:

```
qf(.5,2,928)
```

```
## [1] 0.6936652
```

Now we can use the pivot and the above critical value to find the 95% confidence ellipse:

$$P(\frac{3.801 + 1218.740(\beta_2 - \beta_5)^2 + 1411.680(\beta_3 - \beta_6)^2 + 160.588(\beta_2 - \beta_5)(\beta_3 - \beta_6) + 68.164(\beta_2 - \beta_5) + 131.066(\beta_3 - \beta_6)}{9.384}$$

$$\leq 3.005424) = .95$$

$$P(49.944(\beta_2 - \beta_5)^2 + 57.851(\beta_3 - \beta_6)^2 + 6.581(\beta_2 - \beta_5)(\beta_3 - \beta_6) + 2.793(\beta_2 - \beta_5) + 5.371(\beta_3 - \beta_6) \leq 1) = .95$$

This gives us that $A = 49.944, B = 57.851, C = 6.581, D = 2.793$, and $E = 5.371$.

2)

Answer on separate page.

3)

First we must set up the data. It could have also been read in through a file, however this makes it easier to work from 2 different pc's on the same project.

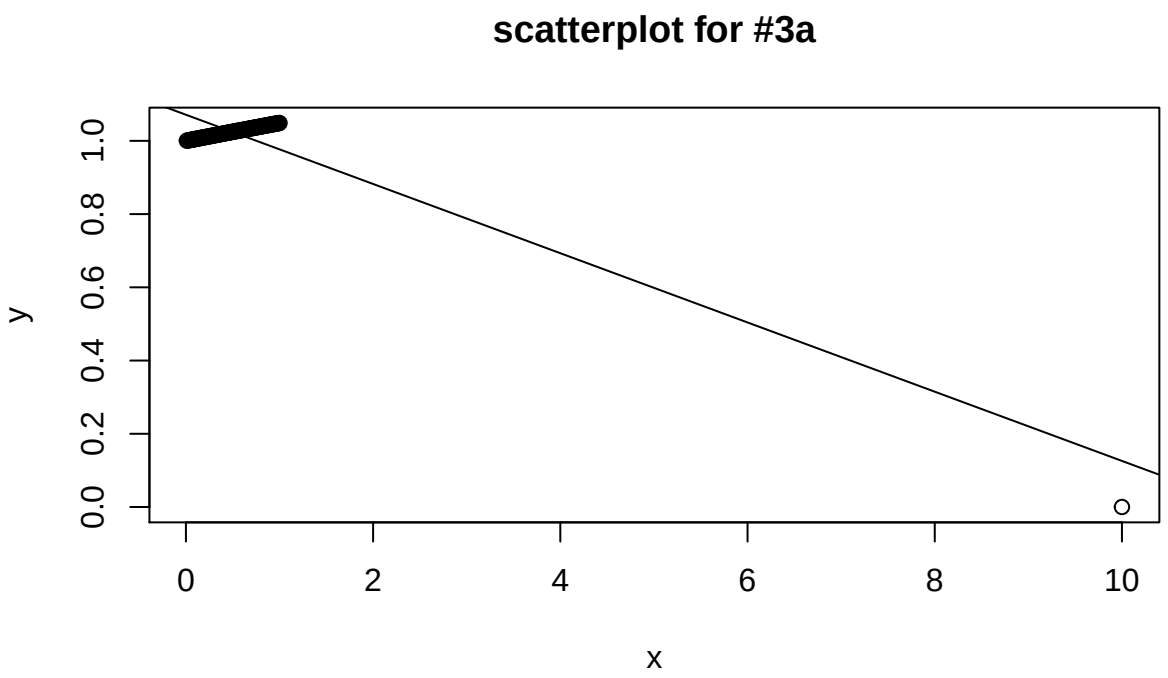
```
set.seed(123)
x0=1:100/100
x=c(x0,10)
y0=sqrt(1+.1*x0+rnorm(100,sd=.0001))
y=c(y0,0)
hw5_3 = cbind(x,y)
outrm_hw5_3 = hw5_3[1:100,]
```

Note the last two lines of code contain the creation of two data frames. The first of which contains the entire data set, the second of which removes the outlier.

a)

Below is a scatterplot of the data, with the line representing our model going through it. As you can see, the outlier at $x = 10$ makes the model seemingly strange for the rest of the data. We will address this in later parts of the problem.

```
model3 = lm(y~x)
plot(x,y, main = "scatterplot for #3a")
abline(model3$coef)
```



b)

Two of the plots will be supplied below. To calculate the residuals, we use the following code:

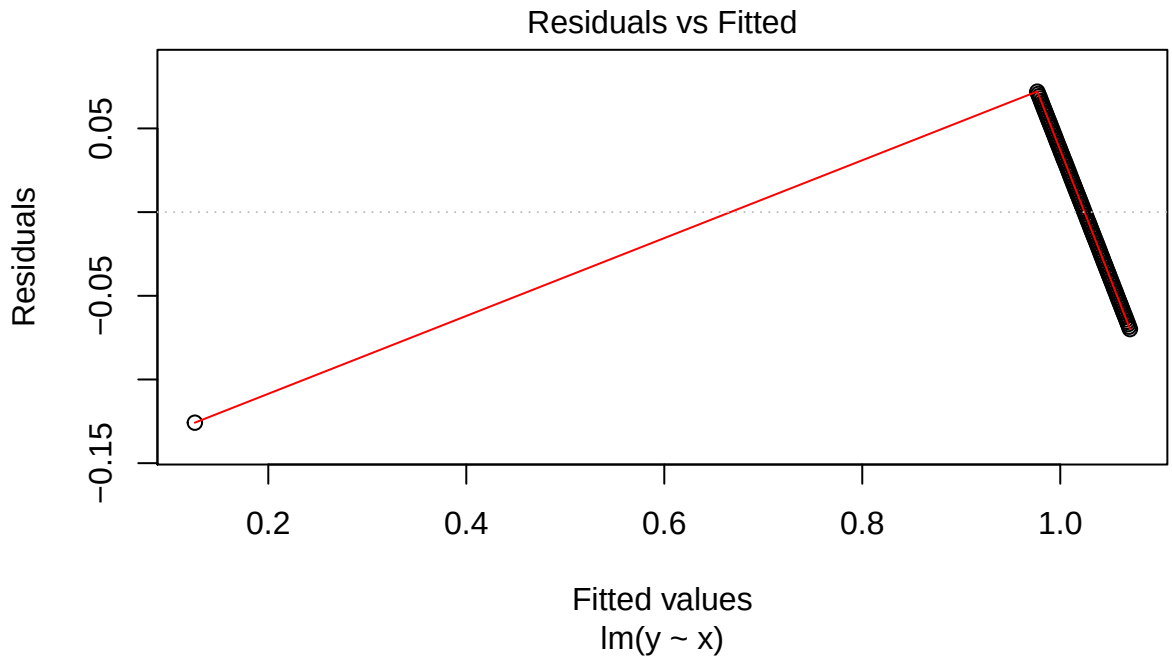
```
epsilon.hat=residuals(model3)
r=rstandard(model3)
t=rstudent(model3)
resids=cbind(epsilon.hat,r,t)
kable(resids[101,])
```

epsilon.hat	-0.1258052
r	-9.9498514
t	-4602.2968772

This is a table containing the values for the residuals of the outlier in the data. As we can see, these values lead us to believe that the outlier should most likely not be contained, as it throws off the entire model.

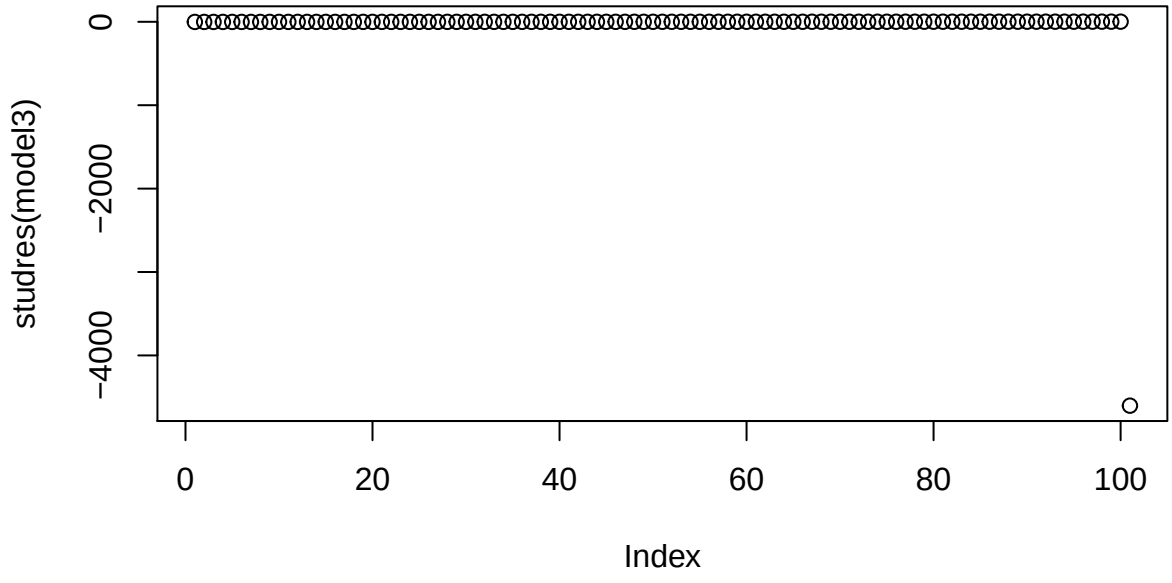
Below is the residual plot.

```
plot(model3, which = 1, labels.id = '')
```



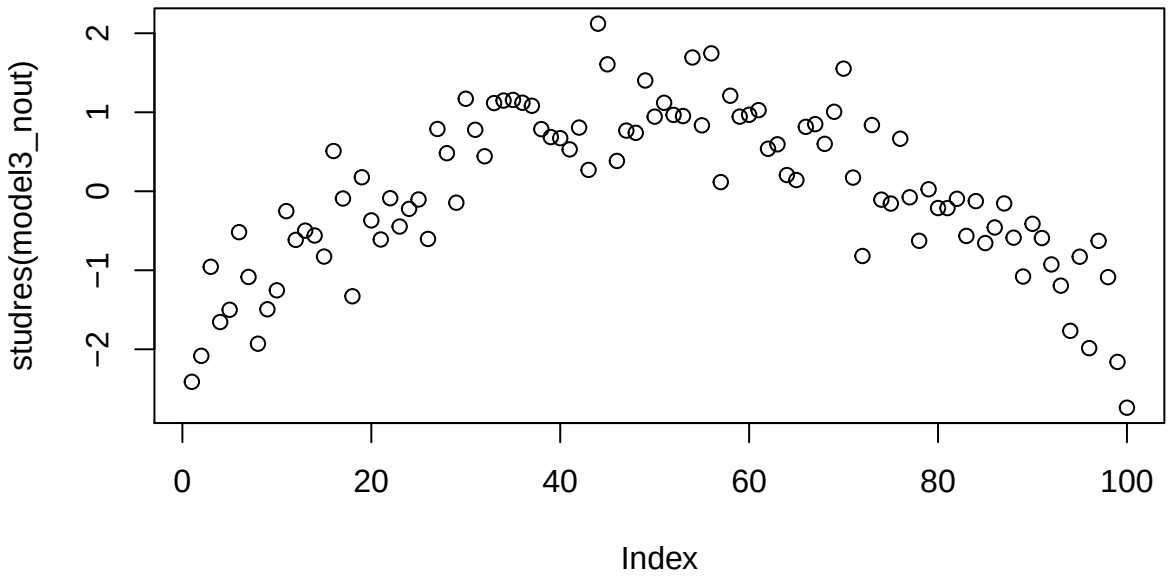
Below is the studentized residual plot with the outlier.

```
plot(studres(model3))
```



Below is the studentized residual plot without the outlier.

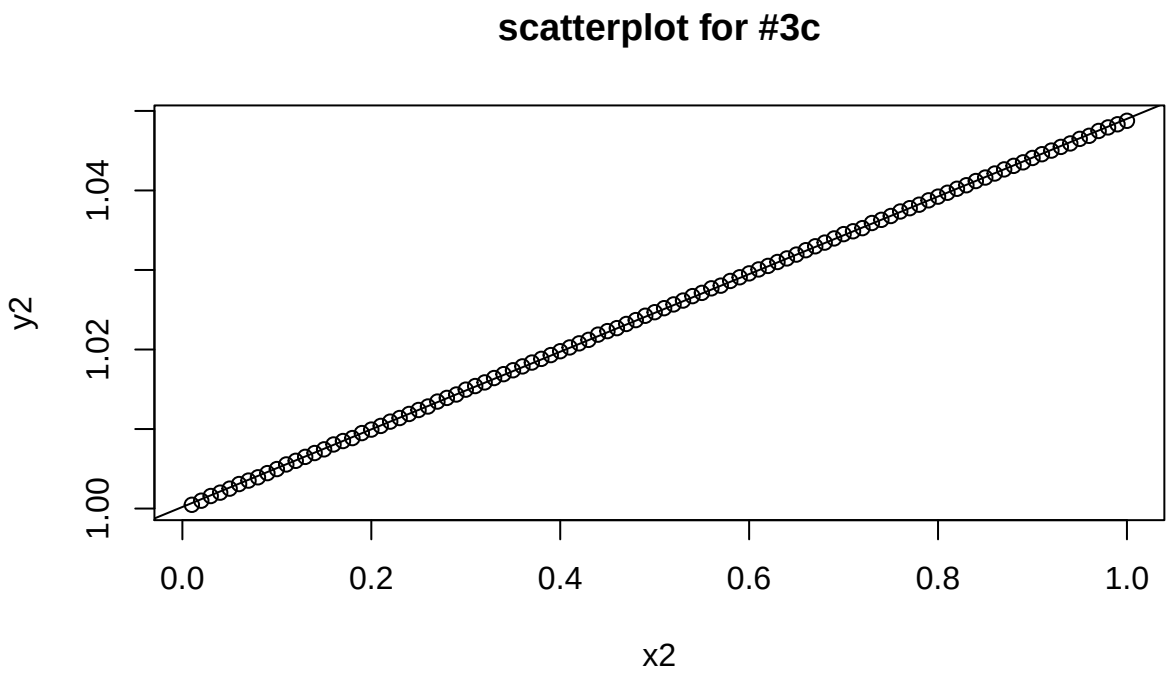
```
x2=outrm_hw5_3[,1]
y2=outrm_hw5_3[,2]
model3_nout = lm(y2~x2)
plot(studres(model3_nout))
```



c)

Below is the scatterplot for the data, excluding the outlier. As we can see, when the outlier is removed, the model follows the data extremely well.

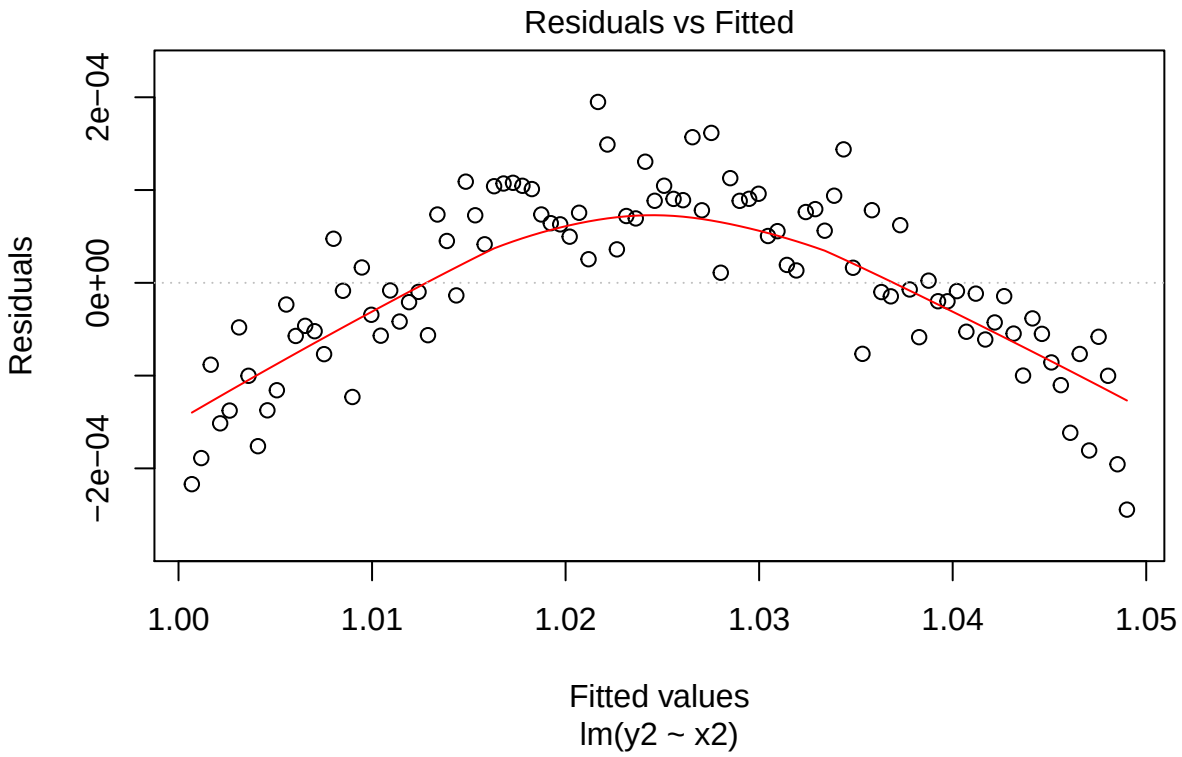
```
plot(x2,y2, main = "scatterplot for #3c")
abline(model3_nout$coef)
```



d)

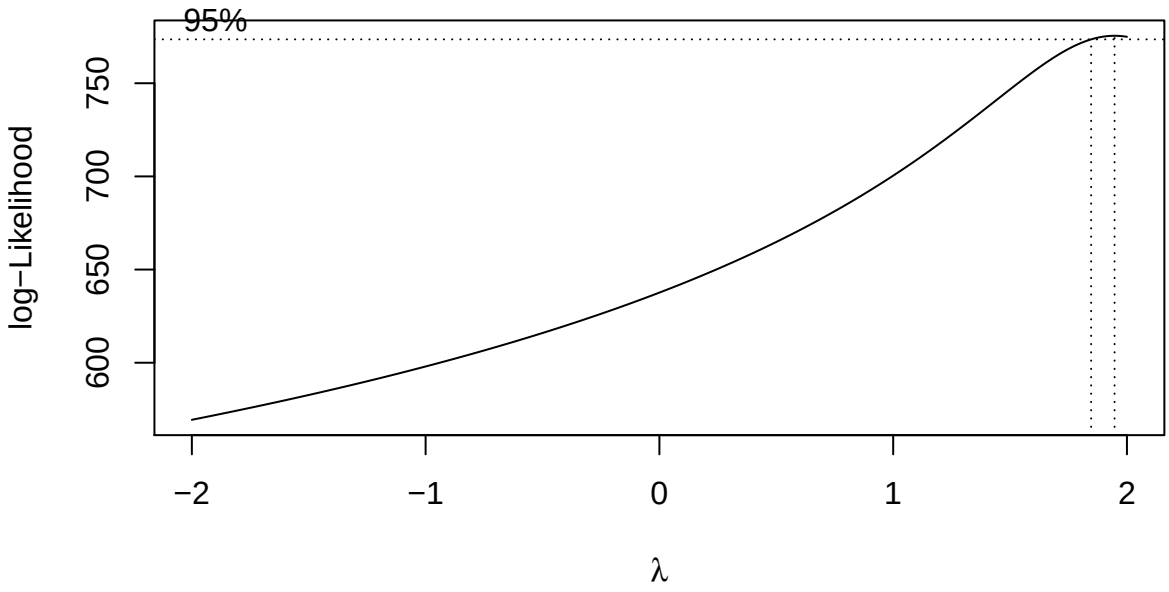
Below is the plot for the residuals without the outlier. As we can see, the model follows the line quite well, leading us to believe that the outlier was causing all of the problems in the previous model (surprise).

```
plot(model3_nout, which=1, labels.id='')
```



e)

```
bc=boxcox(y2~x2,lambda=seq(-2,2,by=.001))
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 1.947
```

The above code tells us that the value of λ that maximizes the log-likelihood function is $\lambda = 1.947$.