STAT 40001/MA 59800 Statistical Computing/ Computational Statistics Fall 2013
Homework 6- Solution

Name:

Due : November 14, 2013 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw R codes with Errors!). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

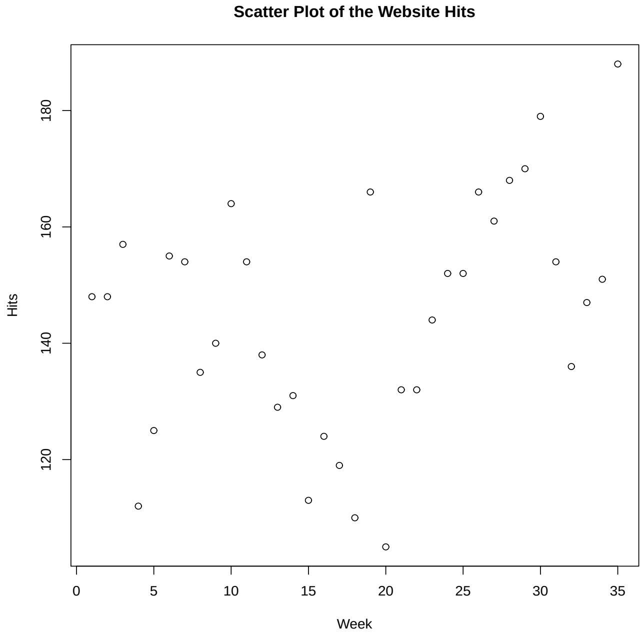
Q.N. 1) An author maintains a website on a particular book and using Google Analytics, records the number of visits on this particular website on each day of the year. As expected there are more hits during weekdays then on weekends. Since the book is used as a textbook for a statistics course there are more hits during the time when the classes are in session. Table below provides the data for 35 weeks from April through November 2009. To explore the week by week visit patterns of these

Week	Hits
1	148
2	148
3	157
4	112
5	125
6	155
7	154
8	135
9	140
10	164
11	154
12	138
13	129
14	131
15	113
16	124
17	119
18	110
19	166
20	105
21	132
22	132
23	144
24	152
25	152
26	166
27	161
28	168
29	170
30	179
31	154
32	136
33	147
34	151
35	188

a) Display the data using a scatterplot

Solution: We imported the data in the R readable format and plotted the scantier plot using R code below

```
> data=read.table("C://STAT 40001//Data sets//website.txt", header=T)
> plot(data, main="Scatter Plot of the Website Hits")
```



b) Calculate the rank correlation coefficient to measure the association between the week and the number of hits on the website.

Solution: Use R code below to calculate the rank correlation coefficient

```
> data=read.table("C://STAT 40001//Data sets//website.txt", header=T)
> cor.test(data$Week,data$Hits, method="spearman")

        Spearman's rank correlation rho

data:  data$Week and data$Hits
S = 4842.713, p-value = 0.05945
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3217489
```

Therefore, the rank correlation coefficient is 0.3217489.

c) Test for the significance of the correlation at **0.05** level.

Solution: We would like to test the hypothesis

$$\begin{aligned} \boldsymbol{H_0} &: \boldsymbol{\rho = 0} \\ \boldsymbol{H_a} &= \boldsymbol{\rho \neq 0} \end{aligned}$$

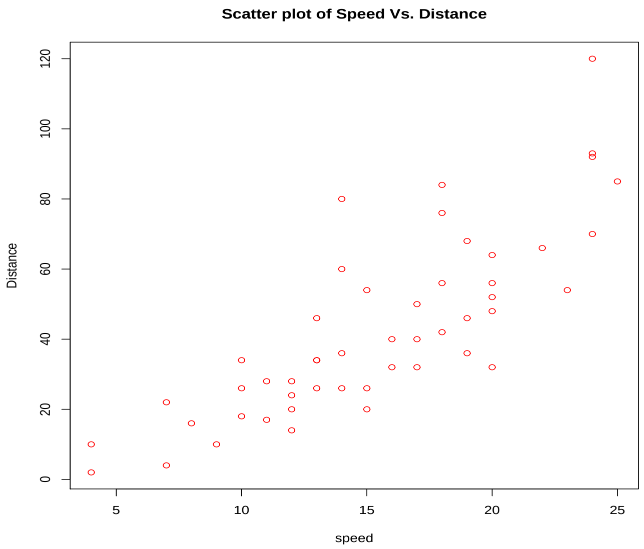
From the R output in part (b) we see that p-value = 0.05945 which is greater than 0.05, so we fail to reject the null hypothesis. This means we don't have enough evidence to say that the week and amount of website hits are correlated.

Q.N. 2) The data set cars is one of the data sets installed with R and is available in base package. The data set contains 50 observations of speed(mph) and dist(stopping distance in feet).

a) Display the data using scatter plot.

Solution: We will read and plot the scatter plot of the data using R code below:

```
> data=cars
> attach(cars)
> names(cars)
[1] "speed" "dist"
> plot(speed,dist, main= "Scatter plot of Speed Vs. Distance", ylab="Distance", col=2)
```



b) Fit a simple regression model using speed as a predictor variable.

Solution: Use R code below to estimate the parameters

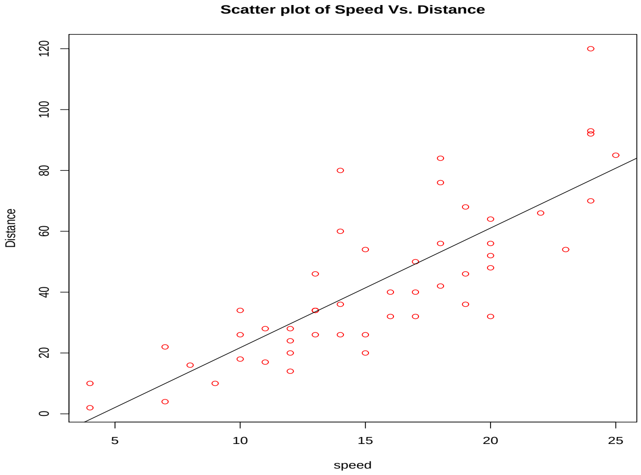
```
> model=lm(dist˜speed)
> model
Call:
lm(formula = dist ˜ speed)
Coefficients:
(Intercept)        speed
    -17.579        3.932
```

Therefor the fitted model is $\boldsymbol{distance = -17.579 + 3.932 \times speed}$

c) Add the fitted line to the scatter plot.

Solution: We can add the fitted line to the scatter plot using the code below:

```
> plot(speed,dist, main= "Scatter plot of Speed Vs. Distance", ylab="Distance", col=2)
> model=lm(dist˜speed)
> abline(model)
```

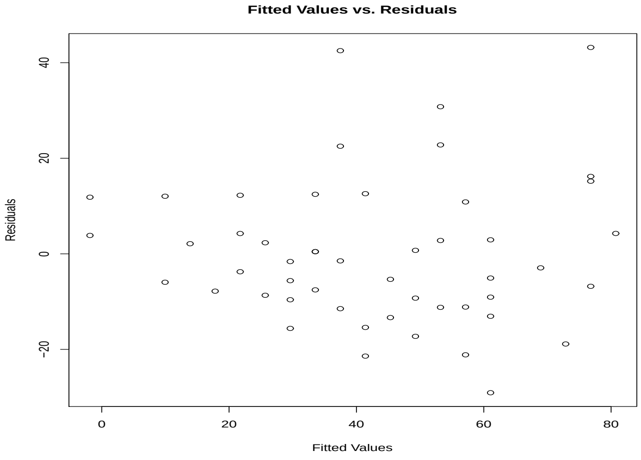


d) Calculate the residuals and fitted values and print only first five observations of the residuals and fitted values.
Solution: The fitted value and residuals of the model are calculated using the R code below:

```
> model=lm(dist~speed)
> fitted=fitted(model)
> residuals=resid(model)
> head(fitted,5)
        1         2         3         4         5
-1.849460 -1.849460  9.947766  9.947766 13.880175
> head(residuals,5)
        1         2         3         4         5
 3.849460 11.849460 -5.947766 12.052234  2.119825
```

e) Create a scatter plot of the residuals and fitted values.
Solution: R code below is used to create the scatter plot of the fitted value and residuals.

```
> model=lm(dist~speed)
> fitted=fitted(model)
> residuals=resid(model)
> plot(fitted,residuals,xlab="Fitted Values",ylab="Residuals",main="Fitted Values vs. Residuals")
```



f) Assuming that no intercept model is appropriate fit a simple linear regression model.
Solution: No-intercept model can be fitted using the R code below:

```
> model1=lm(dist˜-1+speed)
> model1
Call:
lm(formula = dist ˜ -1 + speed)
Coefficients:
speed
2.909
```

Hence, the fitted model is $\boldsymbol{Distance} = \mathbf{2.909} \times \boldsymbol{Speed}$
g) Calculate and compare the coefficient of determination for both the with intercept and no-intercept models.
Solution: In order to calculate the coefficient of determination we use the R code below:

```
> summary(model)
> summary(model1)
```

Note that we have the following values for the coefficient of determination:

Model	R^2	AdjustedR^2
Intercept Model	0.6511	0.6438
No-intercept	0.8963	0.8942

h) Using your fitted model predict the stopping distance for a car with an speed of 21 mph.

Solution: We can use R code below to predict the stopping distance for a car with an speed of 21 mph

```
> model=lm(dist˜speed)
> model1=lm(dist˜-1+speed)
> xval=as.data.frame(21)
> colnames(xval)="speed"
> predict(model,xval)
       1
65.00149
> predict(model1,xval)
       1
61.09178
```

Note that that the intercept model predicts a stopping distance of 65.00149 feet and the no-intercept model predicts a stopping distance of 61.09178 feet.

Q.N. 3) The mammals data set in the MASS package records brain size and body size of 62 different mammals.
a) Fit a regression model to describe the relation between brain size and body size.
Solution: We can use R code below to fit the regression model

```
> library(MASS)
> attach(mammals)
> model=lm(brain˜body)
> model
lm(formula = brain ˜ body)
Coefficients:
(Intercept)         body
    91.0044       0.9665
```

Hence, the fitted model is

$$\textit{Brain size} = \mathbf{91.0044} + \mathbf{0.9665} \times \textit{body size}\textbf{.}$$

b) Calculate the **95%** confidence interval for the slope parameter of the model.
Solution: We can use R code below to calculate the **95%** *confidence interval for parameters.*

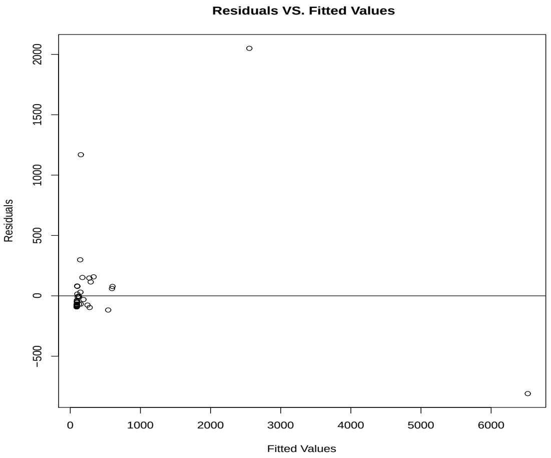
```
> confint(model,level=.95)
                2.5 %     97.5 %
(Intercept) 3.8862623 178.122530
body        0.8711564   1.061836
```

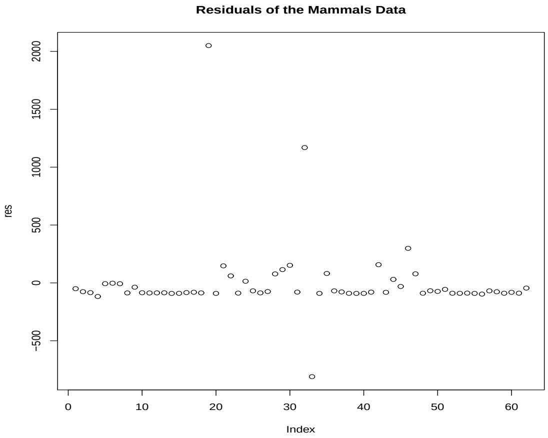
c) Calculate the **90%** confidence interval for the slope parameter of the model.
Solution: We can use R code below to calculate the **90%** *confidence interval for parameters.*

```
> confint(model,level=.90)
                    5 %       95 %
(Intercept) 18.2433254 163.765467
body         0.8868684   1.046124
```

d) Display a residual plot using the plot method for the results of the lm function.
Solution: We can use R code below to display the

```
> model=lm(brain~body)
> fitted=fitted(model)
> residuals=resid(model)
> plot(fitted,residuals,xlab="Fitted Values", ylab="Residuals", main=" Residuals VS. Fitted Values")
> abline(h=0)
```





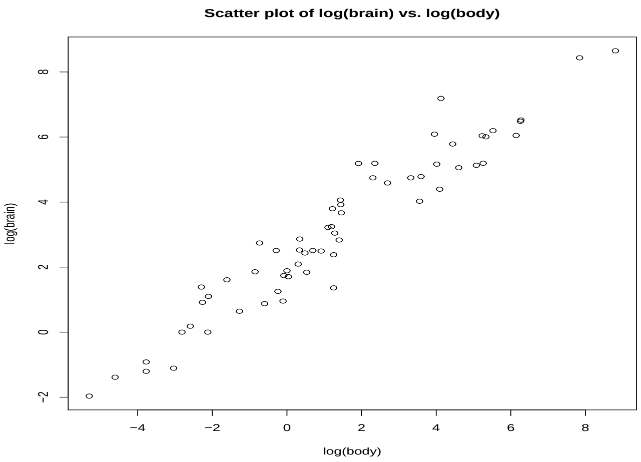
e) Which observation(mammal) has the largest residual in your fitted model?
Solution: We can identify the mammal having the largest residual using R code below

```
> model=lm(brain~body)
> res=residuals(model)
> which.max(res)
19
19
> mammals[19,]
               body brain
Asian elephant 2547  4603
```

It appears that the Asian elephant gives the largest residual.

Q.N. 4) The mammals data set in the MASS package records brain size and body size of 62 different mammals.
a) Display a scatter plot of the log(brain) vs. log(body).
Solution: Please note that we will be using natural `log` *i.e* `ln` *. It should be noted data log in R means ln*
We will use the R code below to convert the data and plot the scatter plot of the data

```
> library(MASS)
> attach(mammals)
> x1=log(body)
> y1= log(brain)
> plot(x1,y1, xlab="log(body)", ylab="log(brain)", main="Scatter plot of log(brain) vs. log(body)")
```



b) Fit a simple linear regression model to the transformed data.
Solution: We can use R code below to fit the regression model

```
> library(MASS)
> attach(mammals)
> x1=log(body)
> y1= log(brain)
> > model=lm(y1~x1)
> model
lm(formula = y1 ~ x1)
Coefficients:
(Intercept)           x1
     2.1348       0.7517
```

Hence, the fitted model is

$$\boldsymbol{log(brain) = 2.1348 + 0.7517 \times log(body).}$$

c) What is the equation of the fitted model.
Solution: We have the fitted model $\boldsymbol{log(brain) = 2.1348 + 0.7517 \times log(body)}$. *which can be simplified as below*

$$\begin{aligned}
\boldsymbol{log(brain)} &= \boldsymbol{2.1348 + 0.7517 \times log(body)} \\
\boldsymbol{log(brain) - 0.7517 \times log(body)} &= \boldsymbol{2.1348} \\
\boldsymbol{log\left(\frac{brain}{(body)^{0.7517}}\right)} &= \boldsymbol{2.1348} \\
\boldsymbol{\frac{brain}{(body)^{0.7517}}} &= \boldsymbol{e^{2.1348}} \\
\boldsymbol{brain} &= \boldsymbol{8.455 \times (body)^{0.7517}}
\end{aligned}$$

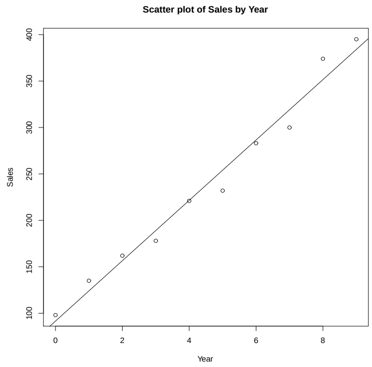
Therefore the equation of the fisted model in the original unit of measurements is

$$\boldsymbol{brain = 8.455 \times (body)^{0.7517}}$$

Q.N. 5) A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where x is the year coded and y is the sales in thousands of units:

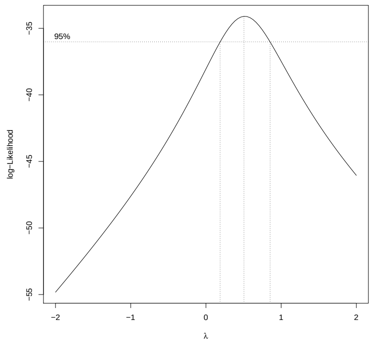
$\boldsymbol{x_i}$:	0	1	2	3	4	5	6	7	8	9
$\boldsymbol{y_i}$:	98	135	162	178	221	232	283	300	374	395

a) Prepare a scatter plot of the data. Does a linear relation appear adequate?
Solution:
a) The scatter plot of the data is as below



b) Use Box-Cox procedure to find an appropriate transformation of $\boldsymbol{y}$.

Solution: b=boxcox(model) in R indicates that $\boldsymbol{\lambda = 0.5}$ *produces the minimum SSE so is the best choice.*

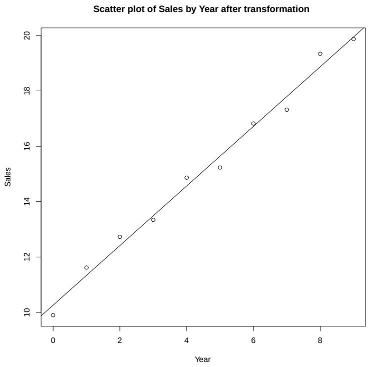


c) Plot the estimated regression line for the transformed data.
Solution: The estimated regression line for the transformed data is

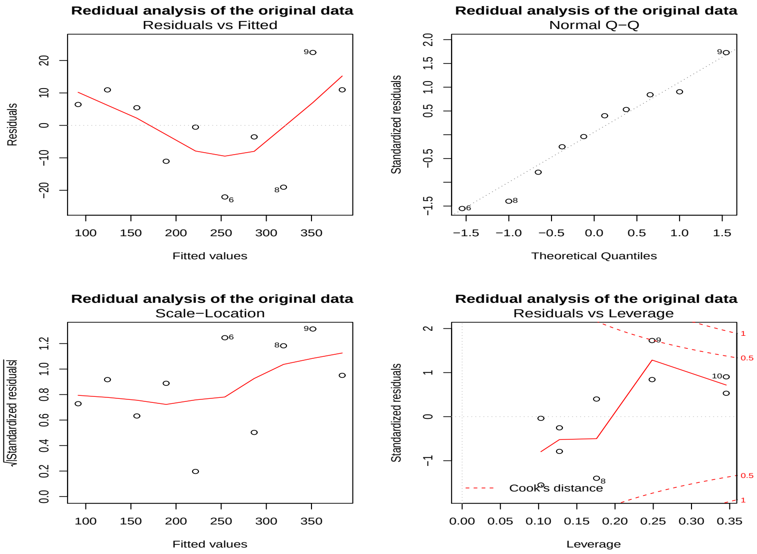
$$\boldsymbol{\hat{y} = 10.261 + 1.076x}$$

and its plot is as below

d) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plot show?



Solution: The residual plot is structureless and the normal probability plot reveals that the transformed data has low standard deviation so the transformed data better fits a simple linear regression model.



e) Express the estimated regression function in the original units.

Solution: Since we have used the square root transformation we will have

$$\boldsymbol{\hat{y} = (10.261 + 1.076x)^2.}$$