STATS191 homework 3

Zolboo Chuluunbaatar

Question 1 (ALSM, 6.18)

```
useful_function = function(dataname) {
    return(paste("http://www.stanford.edu/class/stats191/data/", dataname, sep=''))
}
useful_function("math-salaries.table")
```

```
## [1] "http://www.stanford.edu/class/stats191/data/math-salaries.table"
```

```
h = read.table(useful_function("math-salaries.table"), header=TRUE, sep='')
```

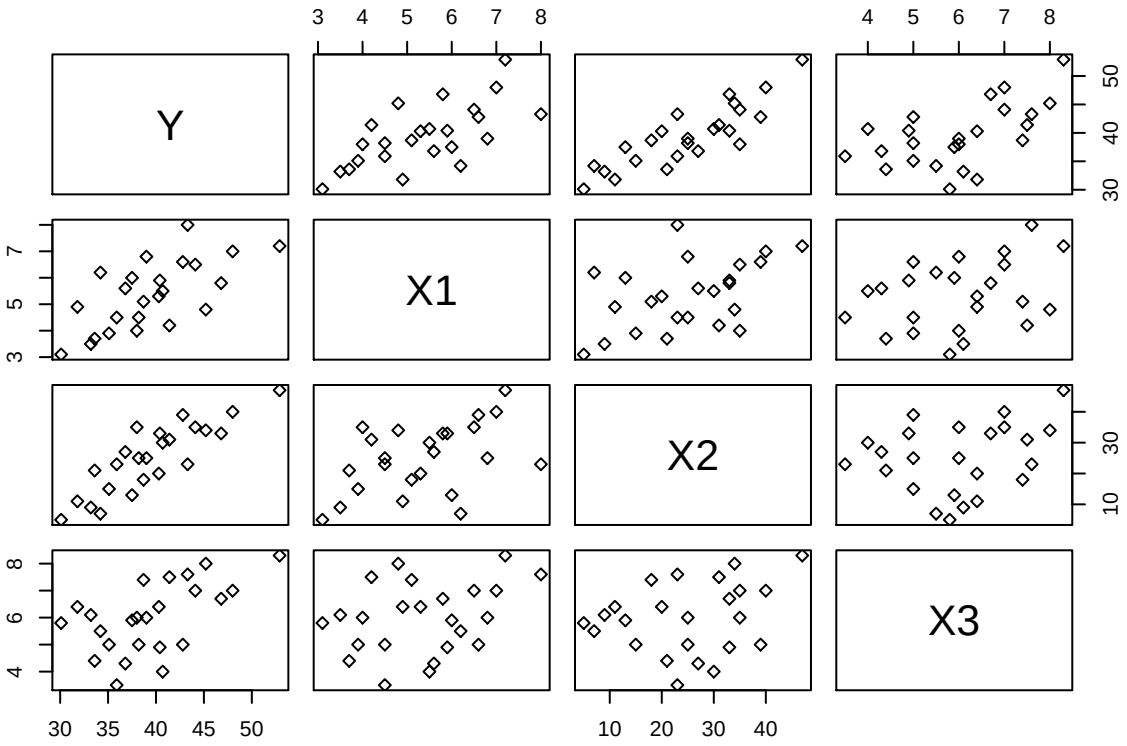
1.

Below, we have the correlation matrix and the scatter plot matrix. Based on the correlation matrix,(considering 0.7 as a threshold for strong correlation) "Y and X2" are the strongest correlated coefficients. Indeed in the scatter plot, we see the tight clustering around the "diagonal Y=X2" line. In other words, knowing X2 helps predict the value of Y. On the other hand, the least uncorrelated coefficients are "X2 and X3".

```
cor(h)
```

```
##            Y        X1        X2        X3
## Y  1.0000000 0.6670958 0.8585582 0.5581960
## X1 0.6670958 1.0000000 0.4669511 0.3227612
## X2 0.8585582 0.4669511 1.0000000 0.2537530
## X3 0.5581960 0.3227612 0.2537530 1.0000000
```

```
pairs(h[,1:4], pch=23)
```



2.

Fitted regression function is $Y = 1.10 * X1 + 0.32 * X2 + 1.29 * X3$

```
fit <- lm(h$Y ~ h$X1 + h$X2 + h$X3, data = h)
summary(fit)
```

```
##
## Call:
## lm(formula = h$Y ~ h$X1 + h$X2 + h$X3, data = h)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2463 -0.9593  0.0377  1.1995  3.3089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.84693    2.00188   8.915 2.10e-08 ***
## h$X1         1.10313    0.32957   3.347 0.003209 **
## h$X2         0.32152    0.03711   8.664 3.33e-08 ***
## h$X3         1.28894    0.29848   4.318 0.000334 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.753 on 20 degrees of freedom
## Multiple R-squared:  0.9109, Adjusted R-squared:  0.8975
## F-statistic: 68.12 on 3 and 20 DF,  p-value: 1.124e-10
```

3.

From the fit of the multilinear regression, $F = 68.12$

Using Goodness of Fit test,

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ H_a : at least one of β's is not zero.

Reject H_0 at level $\alpha = 0.10$ if

$$F > F_{(3,20,0.90)}$$

```
qf(.90, df1=20, df2=3)
```

```
## [1] 5.184482
```

Since $F = 68.12 > 5.18$**, we reject the null hypothesis.**

4.

90% **Confidence Intervals:**

$$CI_{\beta_1} = [0.53, 1.67]$$

$$CI_{\beta_2} = [0.26, 0.39]$$

$$CI_{\beta_3} = [0.77, 1.80]$$

```
confint(fit, level = 0.90,
adjust.method = "bonferroni")
```

```
##                     5 %       95 %
## (Intercept) 14.3942591 21.2996022
## h$X1         0.5347093  1.6715515
## h$X2         0.2575177  0.3855216
## h$X3         0.7741485  1.8037333
```

5.

From the summary of the fit, R-squared $= 0.9109$**, and Adjusted R-squared** $= 0.8975$**.**

R-squared is a statistical measure of how close the data are to the fitted regression line.

Since R-squared $= 0.91$ **is quite closer to 1, the fit is pretty good.**

6.

```
function_s = function(dataname) {
    return(paste("https://web.stanford.edu/class/stats191/data/", dataname, sep=''))
}
function_s("salary_levels.table")
```

```
## [1] "https://web.stanford.edu/class/stats191/data/salary_levels.table"
```

```
salaries = read.table(function_s("salary_levels.table"), header=TRUE, sep='')

salaries$L1
```

```
## [1] 5 6 4
```

```
newdat <- data.frame(X1 = 5, X2 = 6, X3 = 4)
predict(fit, newdat, se.fit=TRUE, interval="confidence", level=0.90)
```

```
## Warning: 'newdata' had 1 row but variables found have 24 rows
```

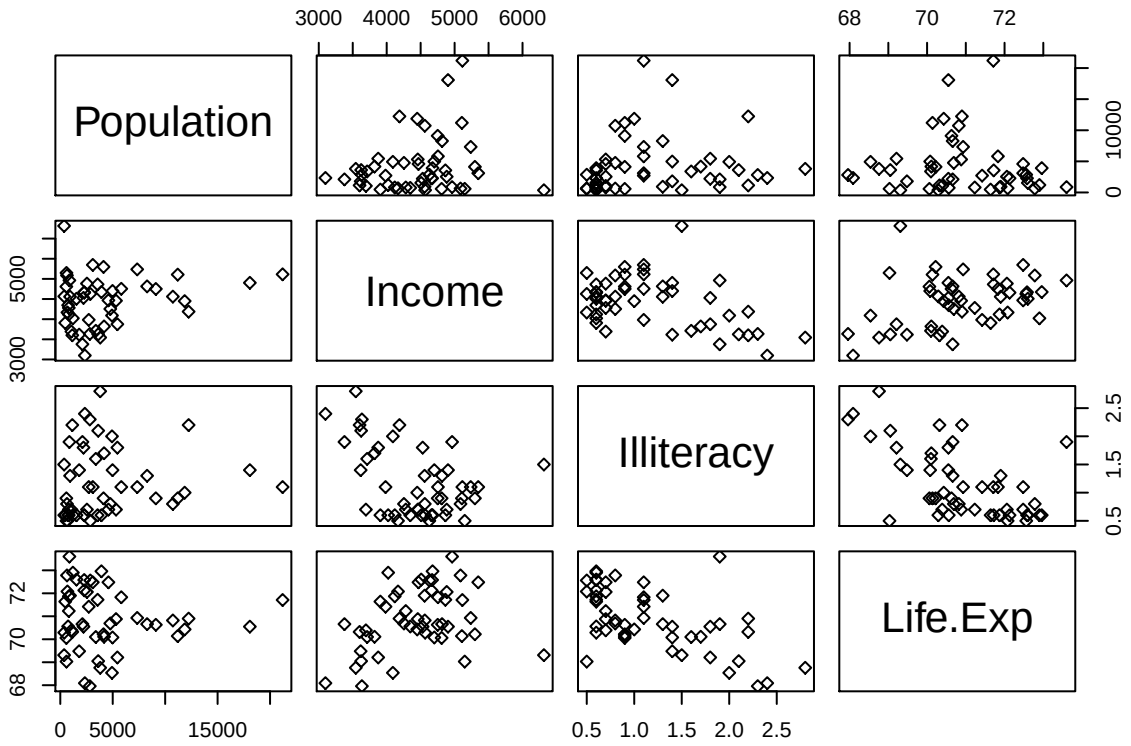
```
## $fit
##         fit      lwr      upr
## 1  32.46410 31.16820 33.76001
## 2  38.37314 37.63912 39.10715
## 3  38.79841 37.70063 39.89620
## 4  43.49114 42.68860 44.29369
## 5  42.11425 40.71605 43.51245
## 6  36.25022 35.09446 37.40598
## 7  41.11985 40.09614 42.14357
## 8  38.71550 37.43605 39.99495
## 9  30.35009 28.86659 31.83359
## 10 51.59910 49.97647 53.22174
## 11 37.29371 36.42181 38.16562
## 12 35.03821 33.95692 36.11950
## 13 43.86288 42.15148 45.57427
## 14 45.29305 44.34872 46.23738
## 15 44.11156 42.80952 45.41360
## 16 34.35177 33.17636 35.52717
## 17 34.02615 32.46328 35.58902
## 18 47.45222 46.29529 48.60915
## 19 41.24629 39.90138 42.59119
## 20 34.71726 33.34433 36.09020
## 21 41.28136 40.24508 42.31763
## 22 38.24794 37.13383 39.36204
## 23 44.38515 42.95211 45.81820
## 24 33.41664 32.41310 34.42018
##
## $se.fit
##         1         2         3         4         5         6         7
## 0.7513718 0.4255855 0.6365019 0.4653197 0.8106831 0.6701159 0.5935567
##         8         9        10        11        12        13        14
## 0.7418331 0.8601384 0.9408134 0.5055328 0.6269379 0.9922741 0.5475272
##        15        16        17        18        19        20        21
## 0.7549283 0.6815052 0.9061599 0.6707925 0.7797833 0.7960345 0.6008373
##        22        23        24
## 0.6459633 0.8308880 0.5818592
```

```
##
## $df
## [1] 20
##
## $residual.scale
## [1] 1.752755
```

Question 2

```
state.data = data.frame(state.x77)
```

```
pairs(state.data[,1:4], pch=23)
```




```
state.lm <- lm(state.data$Income ~ state.data$Population + state.data$Illiteracy + state.data$HS.Grad)

summary(state.lm)
```

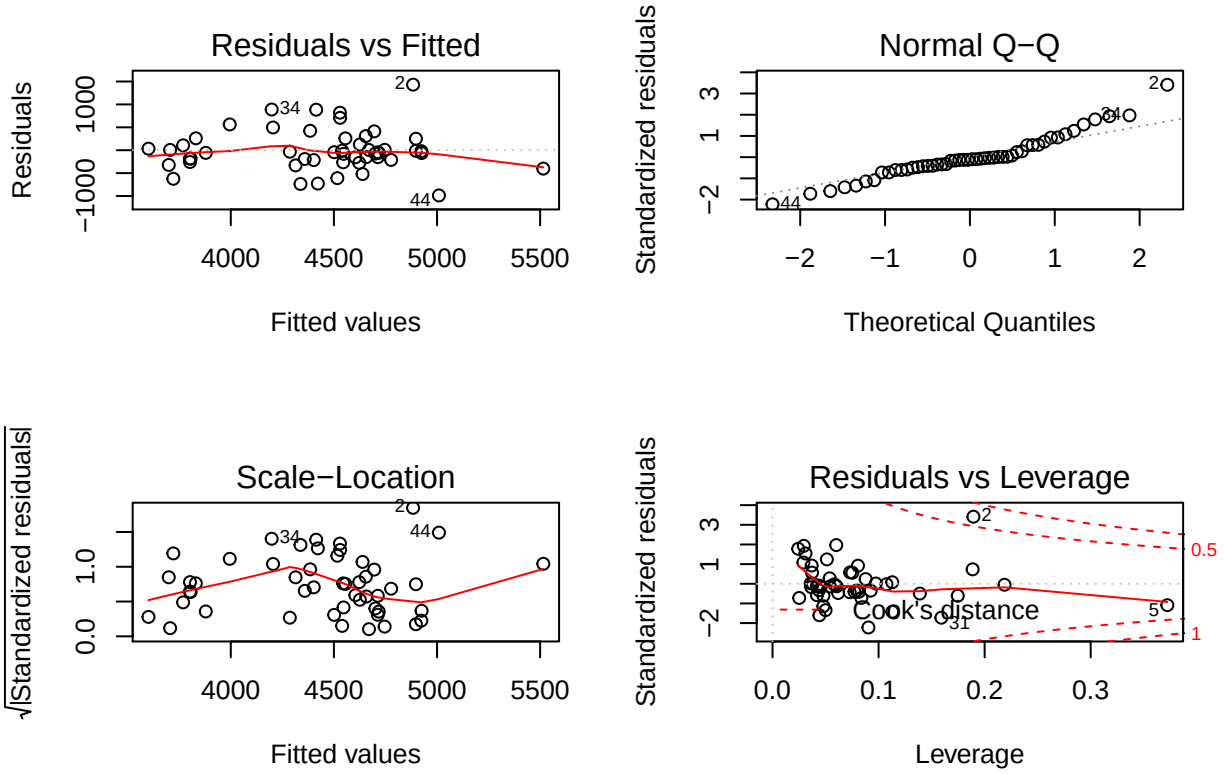
```
##
## Call:
## lm(formula = state.data$Income ~ state.data$Population + state.data$Illiteracy +
##     state.data$HS.Grad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -987.30 -213.67  -50.68  219.74 1430.99
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1940.74115  710.64526   2.731 0.008924 **
## state.data$Population     0.03786    0.01500   2.524 0.015129 *
## state.data$Illiteracy   -73.57563  145.07584  -0.507 0.614470
## state.data$HS.Grad       45.57445   10.93778   4.167 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 465.8 on 46 degrees of freedom
## Multiple R-squared:  0.4606, Adjusted R-squared:  0.4254
## F-statistic: 13.09 on 3 and 46 DF,  p-value: 2.612e-06
```

2. Based on the summary,

the most significant variable is HS.Grad (0.0001), then Population (0.015), and lastly Illiteracy has 0.61 significance. (Here the intercept is also significant.)

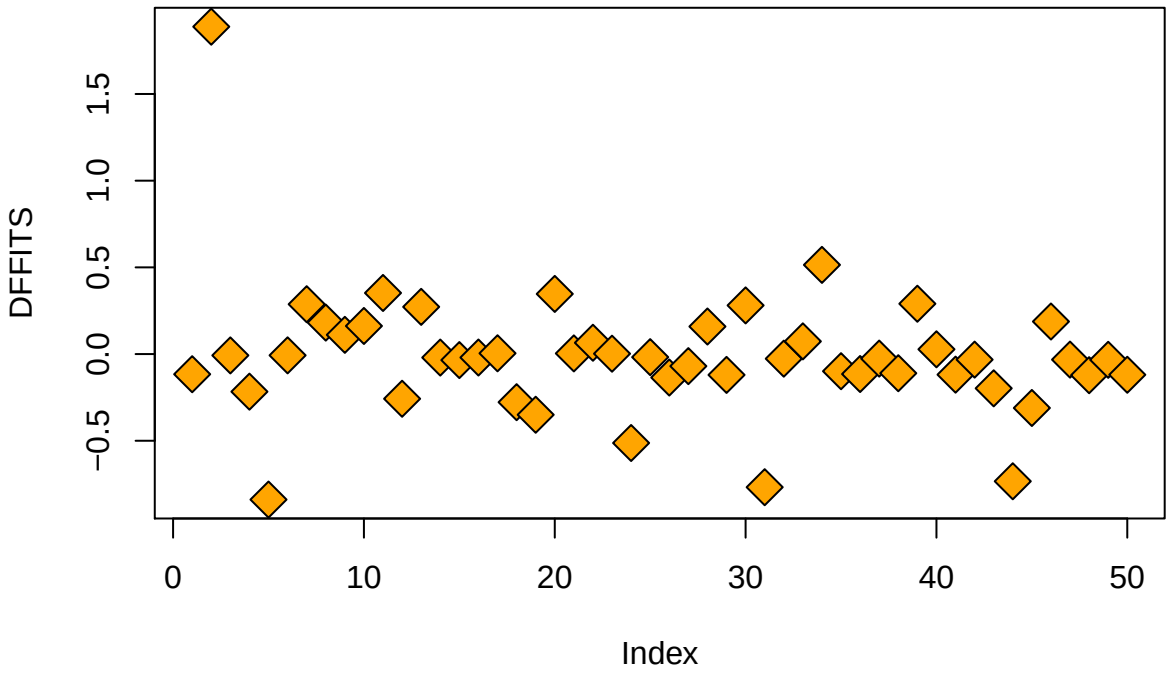
3.

```
par(mfrow = c(2, 2))
plot(state.lm)
```



4.

```
plot(dffits(state.lm), pch=23, bg='orange', cex=2, ylab="DFFITS")
```

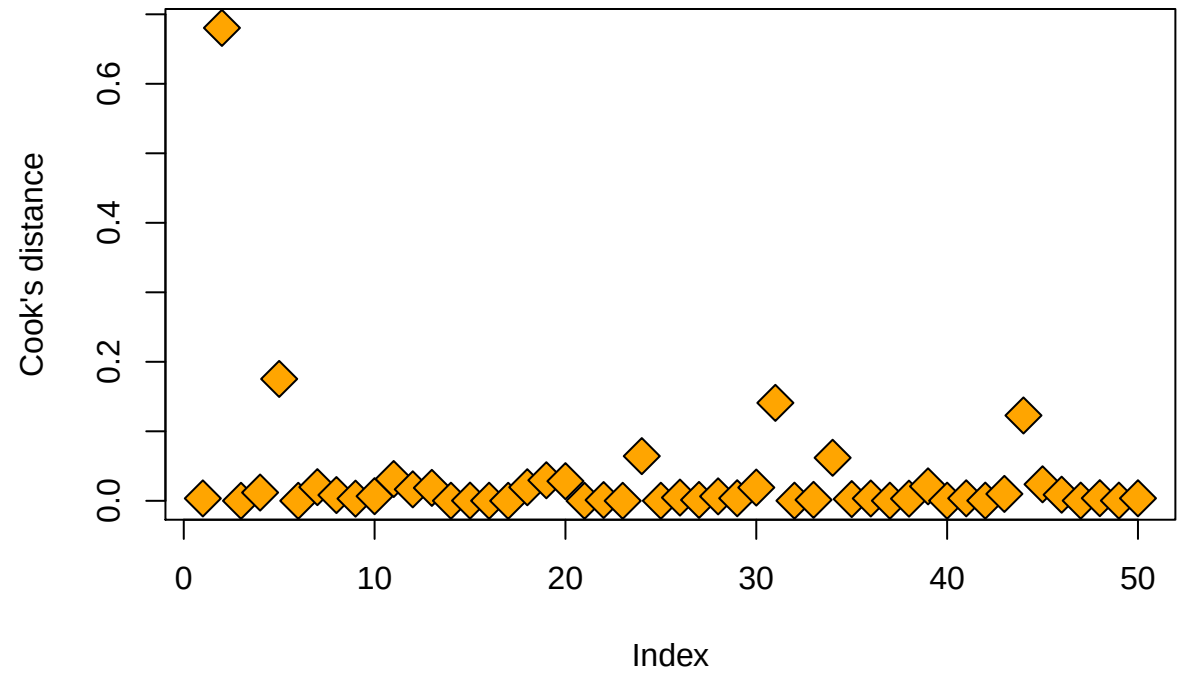



```
state.data[which(dffits(state.lm) > 0.5),]
```

```
##              Population Income Illiteracy Life.Exp Murder HS.Grad Frost
## Alaska              365   6315        1.5    69.31   11.3    66.7   152
## North Dakota        637   5087        0.8    72.78    1.4    50.3   186
##                Area
## Alaska       566432
## North Dakota  69273
```

Alaska and North Dakota have the highest influence. (> 0.5)

```
plot(cooks.distance(state.lm), pch=23, bg='orange', cex=2, ylab="Cook's distance")
```

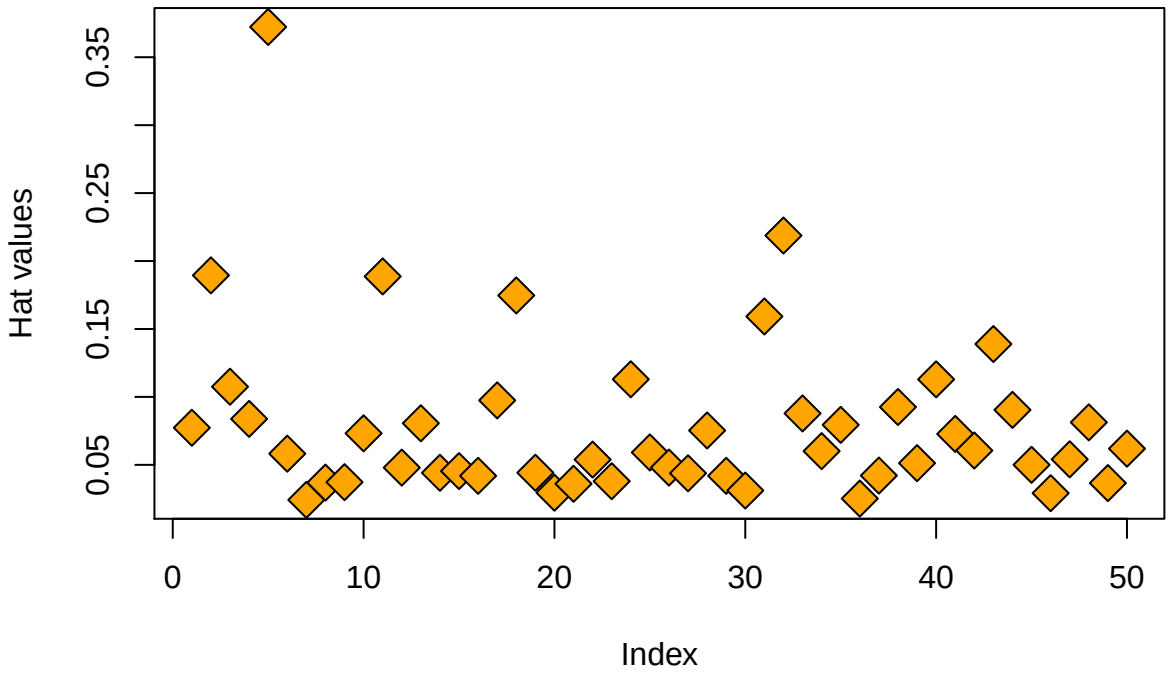


```
state.data[which(cooks.distance(state.lm) > 0.1),]
```

```
##            Population Income Illiteracy Life.Exp Murder HS.Grad Frost
## Alaska            365   6315        1.5    69.31   11.3    66.7   152
## California      21198   5114        1.1    71.71   10.3    62.6    20
## New Mexico       1144   3601        2.2    70.32    9.7    55.2   120
## Utah             1203   4022        0.6    72.90    4.5    67.3   137
##              Area
## Alaska     566432
## California 156361
## New Mexico 121412
## Utah        82096
```

5. Alaska, California, New Mexico, and Utah have the highest influence. (> 0.1)

```
plot(hatvalues(state.lm), pch=23, bg='orange', cex=2, ylab='Hat values')
```




```
state.data[which(hatvalues(state.lm) > 0.3),]
```

```
##            Population Income Illiteracy Life.Exp Murder HS.Grad Frost
## California      21198   5114        1.1    71.71   10.3    62.6    20
##              Area
## California 156361
```

6. California has the outlying predictors.

7.

```
library(car)
```

```
## Loading required package: carData
```

```r
outlierTest(state.lm)
```

```
##   rstudent unadjusted p-value Bonferroni p
## 2 3.905347         0.00031274     0.015637
```

```r
row.names(state.data)[2]
```

```
## [1] "Alaska"
```

By the built-in Outlier Test, Alaska is the outlier.

```r
new.data <- state.data[-c(2), ]
head(new.data)
```

```
##             Population Income Illiteracy Life.Exp Murder HS.Grad Frost
## Alabama           3615   3624        2.1    69.05   15.1    41.3    20
## Arizona           2212   4530        1.8    70.55    7.8    58.1    15
## Arkansas          2110   3378        1.9    70.66   10.1    39.9    65
## California       21198   5114        1.1    71.71   10.3    62.6    20
## Colorado          2541   4884        0.7    72.06    6.8    63.9   166
## Connecticut       3100   5348        1.1    72.48    3.1    56.0   139
##               Area
## Alabama      50708
## Arizona     113417
## Arkansas     51945
## California  156361
## Colorado    103766
## Connecticut   4862
```

```r
new.lm <- lm(new.data$Income ~ new.data$Population + new.data$Illiteracy + new.data$HS.Grad)
summary(new.lm)
```

```
##
## Call:
## lm(formula = new.data$Income ~ new.data$Population + new.data$Illiteracy +
##     new.data$HS.Grad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -806.64 -223.85  -94.83  228.49  895.23
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2922.00458  669.84639   4.362 7.42e-05 ***
## new.data$Population    0.04463    0.01322   3.375  0.00153 **
## new.data$Illiteracy -248.72103  134.46160  -1.850  0.07092 .
## new.data$HS.Grad      29.75007   10.38054   2.866  0.00630 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 407 on 45 degrees of freedom
## Multiple R-squared:  0.4997, Adjusted R-squared:  0.4663
## F-statistic: 14.98 on 3 and 45 DF,  p-value: 6.723e-07
```

Fitting the model without the outlier, we get a new model and the significance of the independent variable Illiteracy changed to 0.07.

```
outlierTest(new.lm)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 6 2.337367           0.024034           NA
```

The Outlier Test says there are no more outliers in the newer model.

8.

```
inflm <- influence.measures(state.lm)
which(apply(inflm$is.inf, 1, any))
```

```
##  2  5 11 18 32
##  2  5 11 18 32
```

```
state.data[which(apply(inflm$is.inf, 1, any)), ]
```

```
##            Population Income Illiteracy Life.Exp Murder HS.Grad Frost
## Alaska            365   6315        1.5    69.31   11.3    66.7   152
## California      21198   5114        1.1    71.71   10.3    62.6    20
## Hawaii            868   4963        1.9    73.60    6.2    61.9     0
## Louisiana        3806   3545        2.8    68.76   13.2    42.2    12
## New York        18076   4903        1.4    70.55   10.9    52.7    82
##              Area
## Alaska     566432
## California 156361
## Hawaii       6425
## Louisiana   44930
## New York    47831
```

If we haven't removed the influential points, (we find all the influential states using the original fitted model), then

the influential states are Alaska, California, Hawaii, Louisiana, and New York.

Question 3

```
data(iris)
```

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
iris.lm <- lm(iris$Sepal.Length ~ iris$Sepal.Width + iris$Petal.Length + iris$Petal.Width, data = iris)

iris.reduced.lm <- lm(iris$Sepal.Length ~ iris$Petal.Width, data = iris)

anova(iris.lm, iris.reduced.lm)
```

```
## Analysis of Variance Table
##
## Model 1: iris$Sepal.Length ~ iris$Sepal.Width + iris$Petal.Length + iris$Petal.Width
## Model 2: iris$Sepal.Length ~ iris$Petal.Width
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    146 14.445
## 2    148 33.815 -2   -19.369 97.884 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova test says F-stat for the reduced model is 97.88 and the P=value is statistically significant. Therefore we do not reject the null hypothesis at significance level $\alpha = 0.05$.

3.

Test

$$H_0 : \beta_{sepalwidth} = \beta_{petallength}$$

```
iris.1.lm <- lm(iris$Sepal.Length ~ I( iris$Sepal.Width + iris$Petal.Length) + iris$Petal.Width, data =

anova(iris.lm, iris.1.lm)
```

```
## Analysis of Variance Table
##
## Model 1: iris$Sepal.Length ~ iris$Sepal.Width + iris$Petal.Length + iris$Petal.Width
## Model 2: iris$Sepal.Length ~ I(iris$Sepal.Width + iris$Petal.Length) +
##     iris$Petal.Width
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    146 14.445
## 2    147 14.508 -1 -0.062559 0.6323 0.4278
```

Anova test says F-stat for the reduced model is 0.63 and the P-value= 0.43 is NOT statistically significant. Therefore we reject the null hypothesis at significance level $\alpha = 0.05$.

The H_0 Null hypothesis is equivalent to the following:

$$H_0 : abs(\beta_{petallengh} - \beta_{sepalwidth}) + \beta_{sepalwidth} = \beta_{petallengh}$$

```
z <- iris$Petal.Length
abs <- abs(z - iris$Sepal.Width)
Z <- iris$Sepal.Width + iris$Petal.Length

iris.2.lm <- lm(iris$Sepal.Length ~ Z + abs + iris$Petal.Width, data = iris)

anova(iris.2.lm, iris.lm)
```

```
## Analysis of Variance Table
##
## Model 1: iris$Sepal.Length ~ Z + abs + iris$Petal.Width
## Model 2: iris$Sepal.Length ~ iris$Sepal.Width + iris$Petal.Length + iris$Petal.Width
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    146 13.999
## 2    146 14.445  0  -0.44602
```

```
sum(iris$Petal.Length - iris$Sepal.Width)
```

```
## [1] 105.1
```

Question 4.

1.

The model "R.lm" below includes an interaction between log(Lscd) and Cluster. These lines have the same intercept but possibly different slopes within the Cluster groups -R and -D.

```
tomasetti = read.csv("https://stats191.stanford.edu/data/Tomasetti.csv")
attach(tomasetti)
log.risk <- log(tomasetti$Risk)
log.lscd <- log(tomasetti$Lscd)
cluster <- tomasetti$Cluster

tomasetti.lm <- lm(log.risk ~ log.lscd, data = tomasetti)

R.lm <- lm(log.risk ~ log.lscd + log.lscd:cluster , data = tomasetti)
summary(R.lm)
```

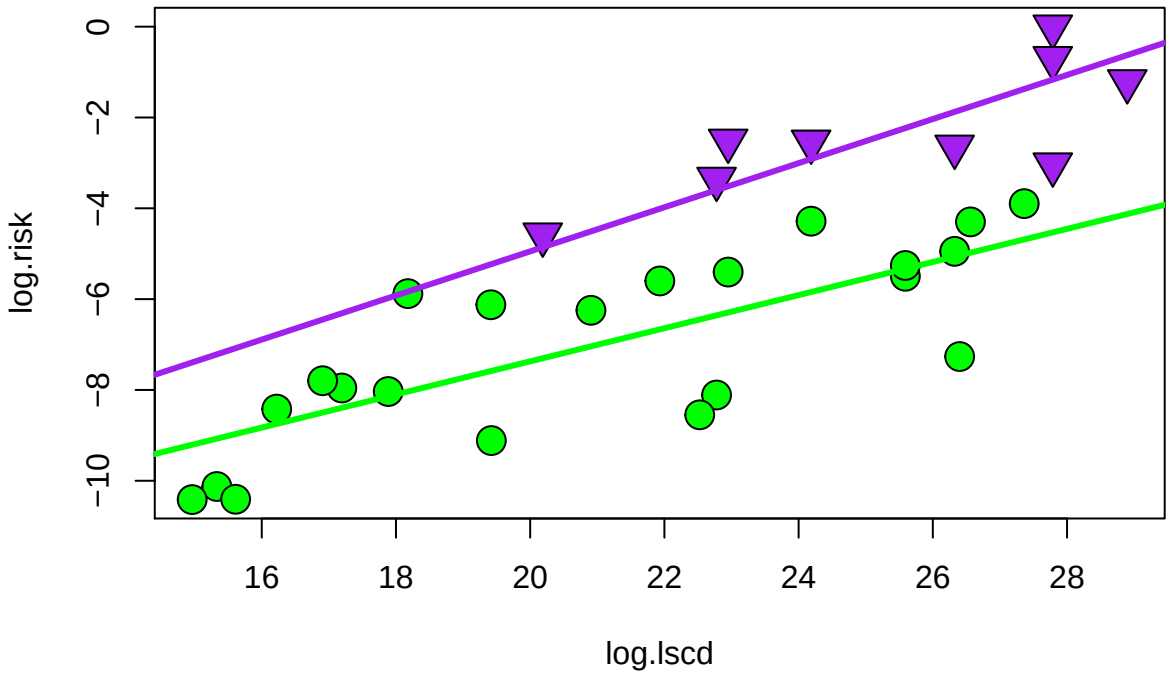
```
##
## Call:
## lm(formula = log.risk ~ log.lscd + log.lscd:cluster, data = tomasetti)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2281 -0.9116  0.2778  0.7911  2.1565
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -14.66296    1.26590 -11.583 3.42e-12 ***
## log.lscd                      0.48574    0.05165   9.405 3.66e-10 ***
## log.lscd:clusterReplicative  -0.12109    0.02149  -5.634 4.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.202 on 28 degrees of freedom
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.8224
## F-statistic: 70.46 on 2 and 28 DF,  p-value: 1.182e-11
```

2.

The scatter plot and the two regression lines (assuming the slopes are different, and the intercepts are the same.)

```
plot(log.lscd, log.risk, type='n')
points(log.lscd[(cluster == "Replicative")], log.risk[(cluster == "Replicative")], pch=21, cex=2, bg='g
points(log.lscd[(cluster == "Deterministic")], log.risk[(cluster == "Deterministic")], pch=25, cex=2, b

abline(R.lm$coef['(Intercept)'], R.lm$coef['log.lscd'], lwd=3, col='purple')
abline(R.lm$coef['(Intercept)'], R.lm$coef['log.lscd'] + R.lm$coef['log.lscd:clusterReplicative'], lwd=
```



3.

P-value is $4.922e-06$ **and is statistically significant.**

```
anova(tomasetti.lm, R.lm)
```

```
## Analysis of Variance Table
##
## Model 1: log.risk ~ log.lscd
## Model 2: log.risk ~ log.lscd + log.lscd:cluster
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 86.275
## 2     28 40.436  1     45.84 31.742 4.922e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.

Since the P-value is statistically significant, the model with classification taken into account is different from the model which doesn't account for this classification. We also see that in the plot, the two regression lines look completely different from each other. Therefore, the p-value from part 3 makes sense.

Question 5.

1.

```
myFunction <- function() {
n<- 100
X <- matrix(rnorm(1000), nrow = n, ncol = 10)
Y <- 1 + 0.1 * X[,1] + rnorm(n)

M <- cbind(X, Y)
colnames(M) <- c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "Y")
return(data.frame(M))
}
```

myFunction() creates a sample table with column names X1, . . . , X10, and Y, and it has 100 entries. Below we show the first 3 rows of the table.

```
sampleData <- myFunction()
head(sampleData, 3)
```

```
##           X1         X2         X3          X4          X5         X6
## 1 -1.5789825 0.02325698  0.3319958  0.96091825 -0.01979968 -1.4525286
## 2 -1.8052733 0.03073790 -0.1210163 -0.04898713 -0.85941167  0.2771803
## 3 -0.5859566 0.19645346  1.1591618  1.11023804  0.72485428  1.0054119
##           X7         X8        X9         X10         Y
## 1  1.3716389 -0.4089755  1.902124 -0.2419698 1.0112462
## 2 -0.3067082  0.8781090 -2.270546 -1.9543477 1.0206061
## 3  1.4081711  0.3745949  1.074139  0.3205065 0.1675841
```

2. Fit a model lm(Y ~ X), computing the features for which the p-value is less than 10% and returning 95% confidence intervals for those selected coefficients. What number should each of these numbers cover? That is, if we form a 95% confidence interval for the effect of X3 what should the interval cover? (Note that there are 11 coefficients so we want 11 different numbers.) How often do your intervals cover what they should? A hint for computation: write a function that returns a vector of length 11 as follows: if a feature is selected return 1 if the interval covers and 0 otherwise; if the feature is not covered set the value to be NA. Store these results as rows in a matrix and compute the mean of each column (removing NA).

```
sample.lm <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10, data = sampleData)
summary(sample.lm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
##     X10, data = sampleData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75479 -0.73443  0.03004  0.69760  2.62276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.02850    0.11788   8.725 1.39e-13 ***
## X1          -0.09470    0.11286  -0.839   0.4036
## X2           0.16989    0.13051   1.302   0.1964
```

```
## X3           0.01544    0.11124   0.139   0.8899
## X4          -0.06915    0.11366  -0.608   0.5445
## X5           0.01994    0.11788   0.169   0.8661
## X6          -0.02249    0.13375  -0.168   0.8669
## X7          -0.10005    0.12770  -0.783   0.4354
## X8          -0.01981    0.12689  -0.156   0.8763
## X9          -0.01814    0.11268  -0.161   0.8725
## X10          0.23782    0.11815   2.013   0.0472 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.139 on 89 degrees of freedom
## Multiple R-squared:  0.08544,    Adjusted R-squared:  -0.01732
## F-statistic: 0.8315 on 10 and 89 DF,  p-value: 0.5996
```

```
pvalues <- summary(sample.lm)$coef[,4]
```

The which function below will return features (including intercept if it has pvalue <0.10) whose p value is less than 0.10.

```
which(summary(sample.lm)$coef[,4] < 0.1)
```

```
## (Intercept)         X10
##           1          11
```

The following will return corresponding confidence intervals for those features.

```
confint(sample.lm)[which(summary(sample.lm)$coef[,4] < 0.1),]
```

```
##                   2.5 %    97.5 %
## (Intercept) 0.794277776 1.2627248
## X10         0.003057858 0.4725852
```

Question 6. (ALSM 19.14)

1.

```
useful_function = function(dataname) {
    return(paste("http://stats191.stanford.edu/data/", dataname, sep=''))
}
useful_function("hayfever.table")
```

```
## [1] "http://stats191.stanford.edu/data/hayfever.table"
```

```
hf.data = read.table(useful_function("hayfever.table"), header=TRUE, sep='')

hf.data$A.factor <- factor(hf.data$A)
hf.data$B.factor <- factor(hf.data$B)
hf.lm <- lm(hours ~ A.factor + B.factor + A.factor:B.factor, data = hf.data)
summary(hf.lm)
```

```
##
## Call:
## lm(formula = hours ~ A.factor + B.factor + A.factor:B.factor,
##     data = hf.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0000 -0.2188 -0.0875  0.2313  2.2000
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.5000     0.3318   7.534 4.19e-08 ***
## A.factor2              2.8750     0.4693   6.126 1.52e-06 ***
## A.factor3              3.8000     0.4693   8.097 1.07e-08 ***
## B.factor2              2.1000     0.4693   4.475 0.000125 ***
## B.factor3              2.0750     0.4693   4.422 0.000144 ***
## A.factor2:B.factor2    1.3250     0.6637   1.996 0.056058 .
## A.factor3:B.factor2    2.3750     0.6637   3.579 0.001334 **
## A.factor2:B.factor3    1.7000     0.6637   2.561 0.016323 *
## A.factor3:B.factor3    5.6250     0.6637   8.476 4.35e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6637 on 27 degrees of freedom
## Multiple R-squared:  0.9725, Adjusted R-squared:  0.9643
## F-statistic: 119.3 on 8 and 27 DF,  p-value: < 2.2e-16
```

Below, the estimated mean value of hours is 5.375 when factor A is 2 and factor B is 1.

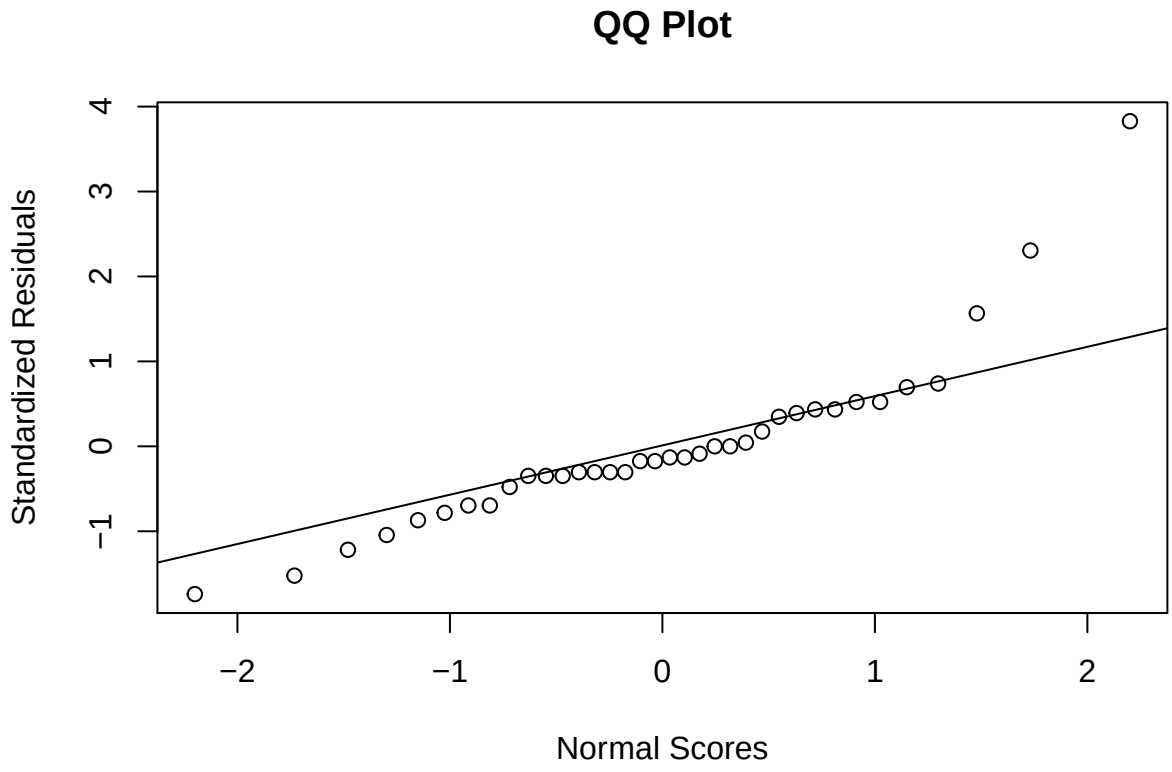
```
predict(hf.lm, list(A.factor=factor(2),B.factor=factor(1)))
```

```
##     1
## 5.375
```

2.

In the qq-plot, we see some outliers that are away from the qqline. Therefore, it violates the normality of the qqline.

```
hf.stdres <- rstandard(hf.lm)
qqnorm(hf.stdres,ylab="Standardized Residuals", xlab="Normal Scores", main="QQ Plot")
qqline(hf.stdres)
```



Moreover, Shapiro-Wilk test of normality implies that the p-value is statistically significant, so the distribution is significantly different from normal distribution. In other words, we cannot assume the normality.

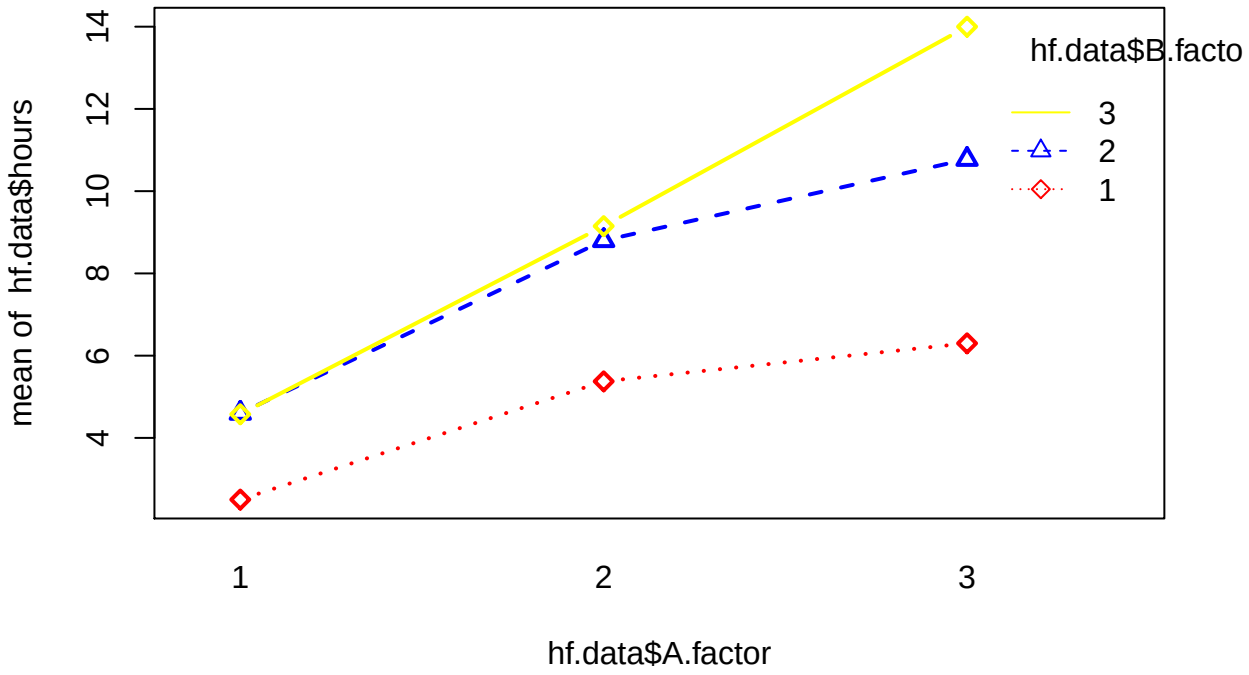
```
shapiro.test(hf.stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hf.stdres
## W = 0.85973, p-value = 0.0003214
```

3.

Below we plot the interaction plot. Since these broken lines are not parallel, there is evidence of an interaction.

```
interaction.plot(hf.data$A.factor, hf.data$B.factor, hf.data$hours, type='b', col=c('red',
                 'blue', 'yellow'), lwd=2, pch=c(23,24))
```



4.

Based on anova(hf.lm) result below, we see that the Interaction of Factors A and B is statistically significant. So we can REJECT the null hypothesis that there is an interaction between Factors A and B.

```
hf1.lm <- lm(hours ~ A.factor + B.factor, data = hf.data)
anova(hf1.lm, hf.lm)
```

```
## Analysis of Variance Table
##
## Model 1: hours ~ A.factor + B.factor
## Model 2: hours ~ A.factor + B.factor + A.factor:B.factor
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     31 46.294
## 2     27 11.892  4    34.402 19.526 1.186e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(hf.lm)
```

```
## Analysis of Variance Table
##
## Response: hours
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## A.factor           2 254.287 127.143 288.658 < 2.2e-16 ***
## B.factor           2 131.647  65.823 149.441 2.495e-15 ***
## A.factor:B.factor  4  34.402   8.600  19.526 1.186e-07 ***
## Residuals         27  11.892   0.440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.

We can see that Main Effects of Factors A and B are all statistically significant at level 0.05. We can REJECT the null hypothesis that there are main effects of Factor A and Factor B.