CSCI E-106:Assignment 5

Due Date: October 12, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

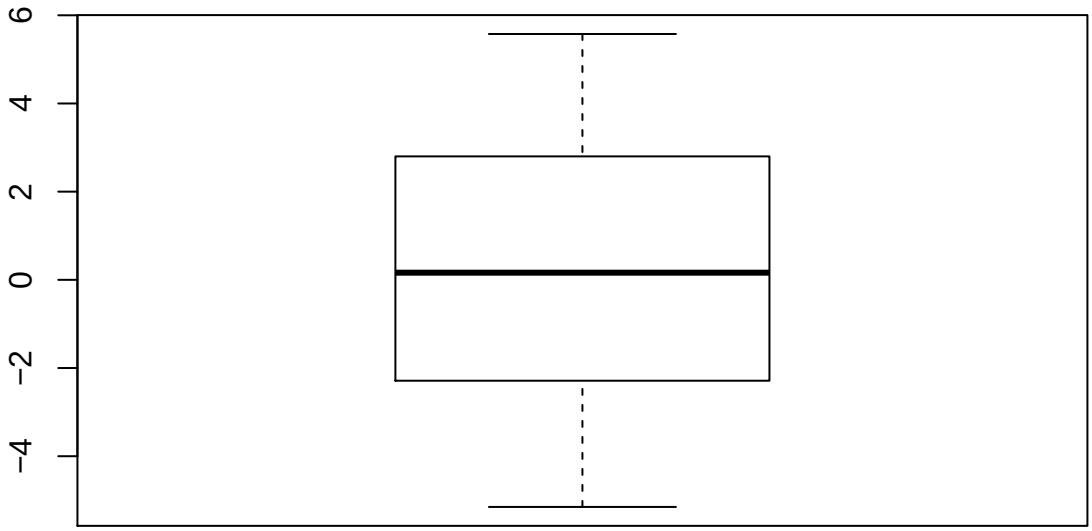
Refer to Plastic hardness data set. X is the elapsed time in hours? and Y is hardness in Brinell units. Build a model to predict Y. (30 points, 5 points each)

Solutions

a-) Obtain the residuals ei and prepare a box plot of the residuals. What information is provided by your plot?

The residuals are normally distributed and there are no outliers.

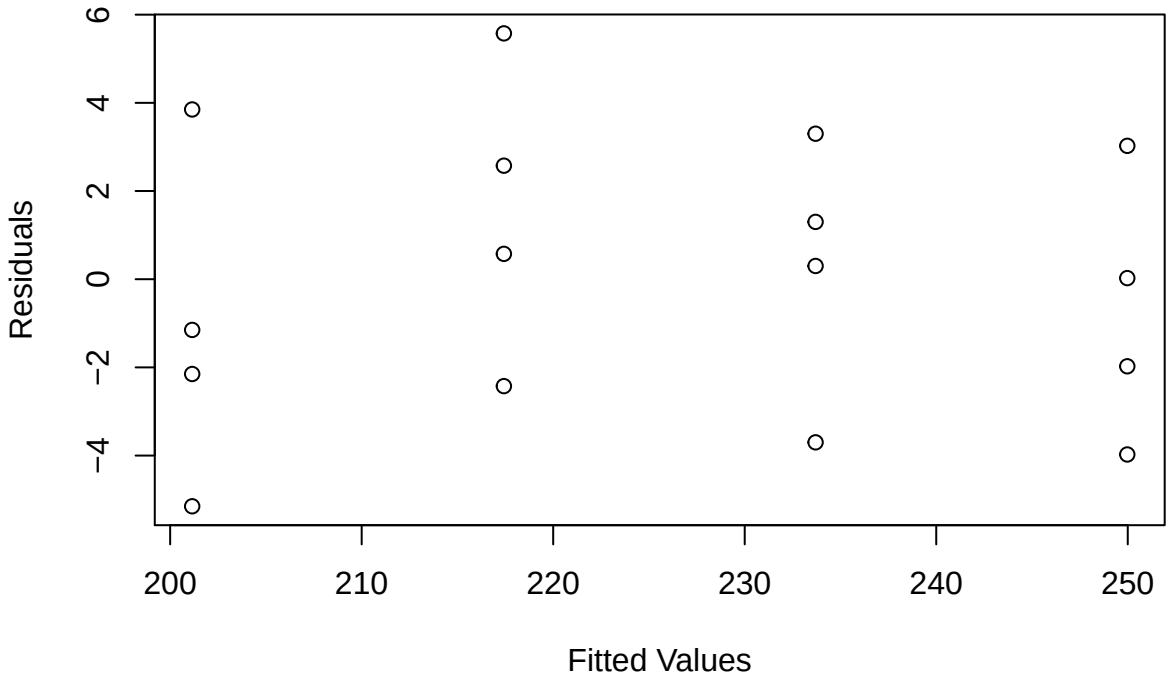
```
PH <- read.csv("/cloud/project/Fall 2020/Plastic Hardness Data.csv")
f<-lm(Y~X,data=PH)
ei<-f$residuals
boxplot(ei)
```



b-) Plot the residuals ei against the fitted values Y; to ascertain whether any departures from regression model (2.1) are evident. State your findings.

No sign of heterodasticty from this graph.

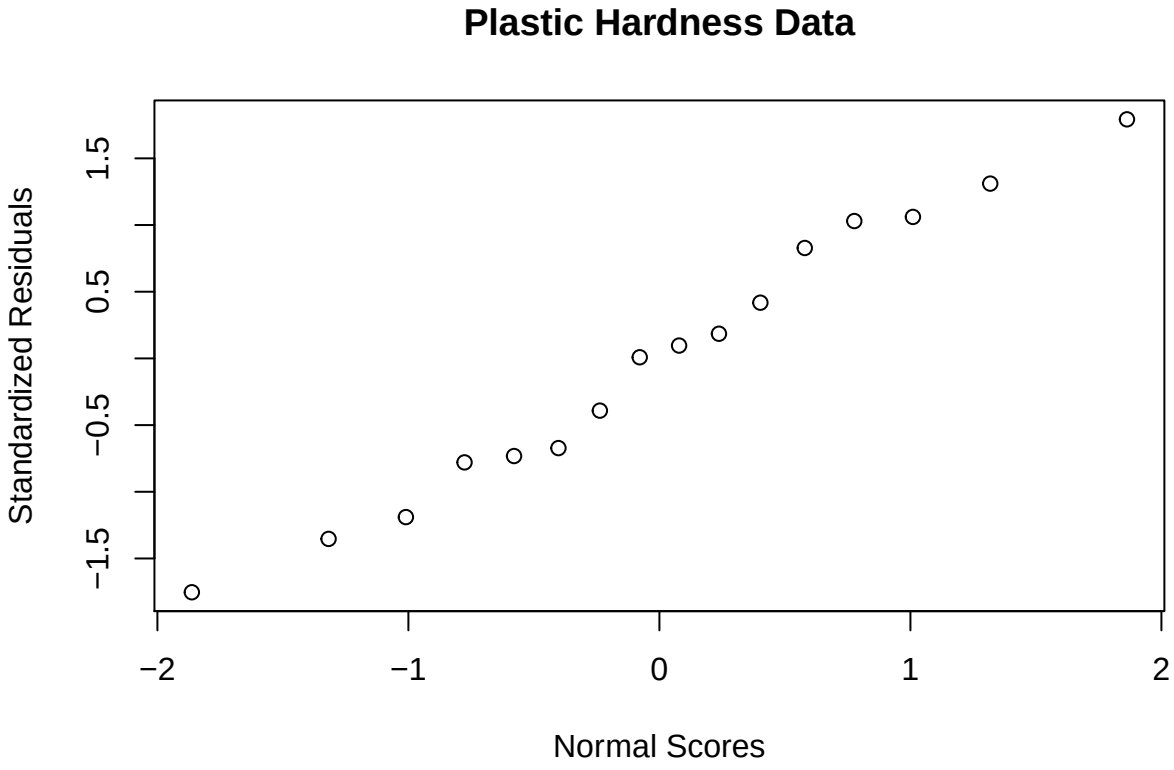
```
plot(f$fitted.values,ei,xlab="Fitted Values",ylab="Residuals")
```



c-) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here?

QQ plot does not indicate any significant departure from the normal distribution.

```
error.std = rstandard(f)
qqnorm(error.std,ylab="Standardized Residuals",xlab="Normal Scores",main="Plastic Hardness Data")
```



d-) Compare the frequencies of the residuals against the expected frequencies under normality, using the 25th, 50th, and 75th percentiles of the relevant distribution. Is the information provided by these comparisons consistent with the findings from the normal probability plot in part (c)?

From the regression output, $\sqrt{MSE}$=3.234 and n=16. Shapiro-Wilk normality test:

Ho: Data is normally distributed Ha: Data is NOT normally distributed

P value is 0.89 >0.05, Accept Ho. Data is normally distributed. 25% and 75% percentiles are (-1.97 and 1.97) under the normal distribution, under the t distribution, the percentiles are (-0.69, 0.69). 50th percentile or median value is 0 under both approach.

```
shapiro.test(ei)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.97348, p-value = 0.8914
```

```
ki<-rank(ei)
n<-length(ei)
summary(f)
```

```
##
## Call:
## lm(formula = Y ~ X, data = PH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1500 -2.2188  0.1625  2.6875  5.5750
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
## X             2.03438    0.09039   22.51 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.234 on 14 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9712
## F-statistic: 506.5 on 1 and 14 DF,  p-value: 2.159e-12
```

```
exp.z<-3.234*qnorm((ki-0.375)/(n+0.25))
#under the expected normal distribution
summary(exp.z)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -5.720  -1.997   0.000   0.000   1.997   5.720
```

```
#under T distribution
cbind(qt(0.25,14),qt(0.50,14),qt(0.75,14))
```

```
##            [,1] [,2]      [,3]
## [1,] -0.6924171    0 0.6924171
```

e-) Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X \leq 24$, $X > 24$, and use $\alpha = 0.01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (b)?

Ho: Error Variance is constant Ha: Error Variance is Not constant

p-value of the test is 0.796, accept the null. Error variance is constant.

```
install.packages("onewaytests")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.5'
## (as 'lib' is unspecified)
```

```
library(onewaytests)
ind=I(PH$X<=24)*1
bf.dat<-data.frame(ei=f$residuals,yhat=f$fitted.values,ind=as.factor(I(PH$X<=24)*1))
bf.test(ei~ind,data=bf.dat)
```

```
##
##   Brown-Forsythe Test (alpha = 0.05)
## ------------------------------------------------------------
##   data : ei and ind
##
##   statistic  : 0.06942228
##   num df     : 1
##   denom df   : 13.1945
##   p.value    : 0.7962498
##
##   Result     : Difference is not statistically significant.
## ------------------------------------------------------------
```

f-) conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = 0.01$. State the alternatives. decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (b)? From the regression model ${e_i}^2 = \gamma_0 + \gamma_1 X_i$

the pvalue for γ_1 is 0.44 and the R^2 is 0.05. Indicating that there is no linear relationship between X and the residuals. Furthermore, the chi square test for

Ho: $\gamma_i = 0$ Ha: $\gamma_i \neq 0$

indicates that the variances are equal.

```
ei2<-(f$residuals)^2
g<-lm(ei2~PH$X )
summary(g)
```

```
##
## Call:
## lm(formula = ei2 ~ PH$X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.562  -6.728  -2.975   3.566  21.018
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.5281     7.7384   2.007   0.0645 .
## PH$X         -0.2277     0.2633  -0.865   0.4016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.419 on 14 degrees of freedom
## Multiple R-squared:  0.05074,    Adjusted R-squared:  -0.01707
## F-statistic: 0.7483 on 1 and 14 DF,  p-value: 0.4016
```

```
anova(g)
```

```
## Analysis of Variance Table
##
## Response: ei2
##           Df  Sum Sq Mean Sq F value Pr(>F)
## PH$X       1   66.38  66.385  0.7483 0.4016
## Residuals 14 1242.01  88.715
```

```
anova(f)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1 5297.5  5297.5  506.51 2.159e-12 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSE.F=146.4
SSR.R=1242.01
tc<-(SSR.R/2)/((SSE.F/16)^2)
pchisq(tc,1)
```

```
## [1] 0.9935405
```

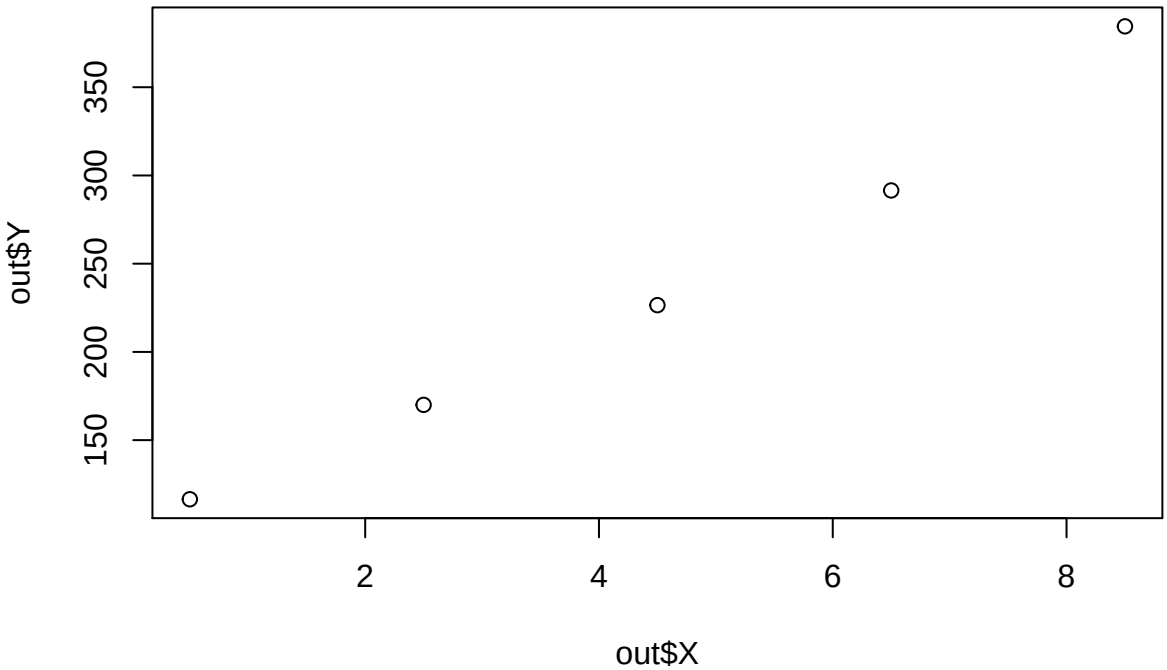
Problem 2

Refer to Sales growth Data. (30 points, 10 points each)

a-) Divide the range of the predictor variable (coded years) into five bands of width 2.0, as follows: Band 1 ranges from X = -.5 to X = 1.5; band 2 ranges from X = 1.5 to X = 3.5; and so on. Determine the median value of X and the median value of Y in each band and develop the band smooth by connecting the five pairs of medians by straight lines on a scatter plot of the data. Does the band smooth suggest that the regression relation is linear? Discuss.

Yes, it does suggest the regression relation is linear. Please see below.

```
SG <- read.csv("/cloud/project/Fall 2020/Sales Growth Data.csv")
ind=SG$X
for (i in 1:10){
if (SG$X[i]<=1.5){ind[i]=1} else if (SG$X[i]<=3.5){ind[i]=2} else if (SG$X[i]<=5.5){ind[i]=3} else if (S
q2.dat<-data.frame(SG,ind)
out<-aggregate(q2.dat[,1:2], list(q2.dat[,3]), median)
plot(out$X,out$Y)
```



Yes, it does suggest the regression relation is linear. Please see below.

b-) Create a series of seven overlapping neighborhoods of width 3.0 beginning at X = -.5. The first neighborhood will range from X = -.5 to X = 2.5; the second neighborhood will range from X = .5 to X = 3.5; and so on. For each of the seven overlapping neighborhoods, fit a linear regression function and obtain the fitted value $\hat{Y}_c$ at the center X_c of the neighborhood. Develop a simplified version of the lowess smooth by connecting the seven $(X_c, \hat{Y}_c)$ pairs by straight lines on a scatter plot of the data.

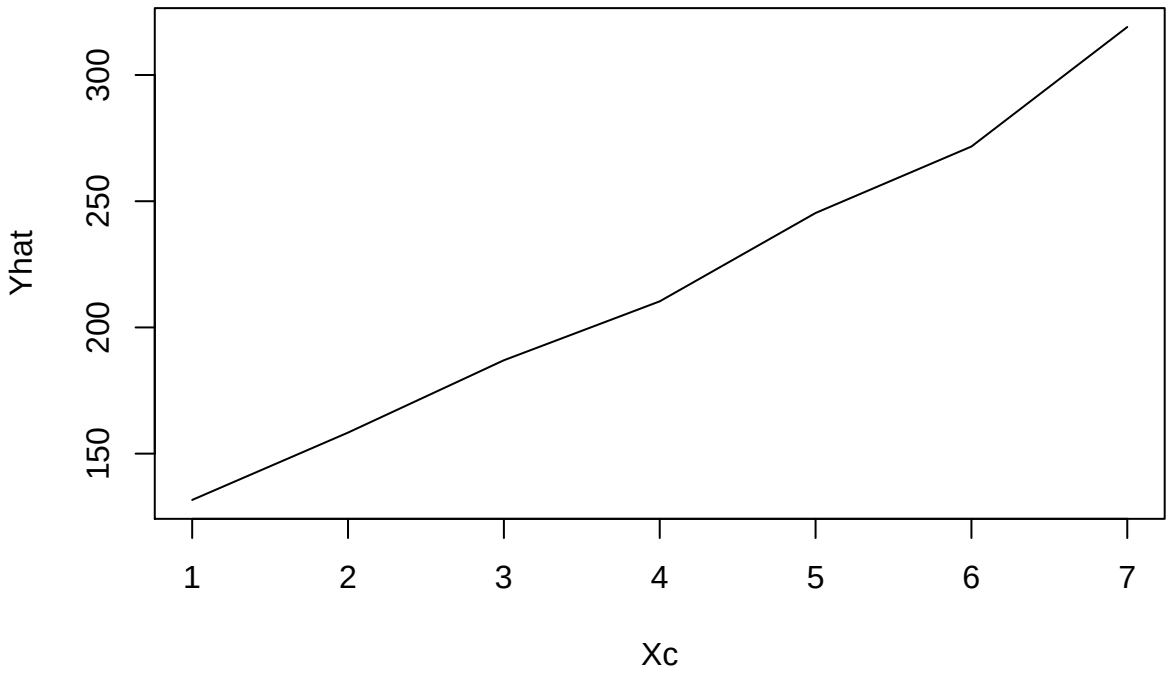
```
#1 X = -.5 to X = 2.5;
f1<-lm(Y~X,data=SG[SG[,2]>=-0.5 & SG[,2]<=2.5 ,])
f1.pred<-f1$fitted.values
#2 X = 0.5 to X = 3.5;
f2<-lm(Y~X,data=SG[SG[,2]>=0.5 & SG[,2]<=3.5 ,])
f2.pred<-f2$fitted.values
#3 X = 1.5 to X = 4.5;
f3<-lm(Y~X,data=SG[SG[,2]>=1.5 & SG[,2]<=4.5 ,])
```

```r
f3.pred<-f3$fitted.values
#4 X = 2.5 to X = 5.5;
f4<-lm(Y~X,data=SG[SG[,2]>=2.5 & SG[,2]<=5.5 ,])
f4.pred<-f4$fitted.values
#5 X = 3.5 to X = 6.5;
f5<-lm(Y~X,data=SG[SG[,2]>=3.5 & SG[,2]<=6.5 ,])
f5.pred<-f5$fitted.values
#6 X = 4.5 to X = 7.5;
f6<-lm(Y~X,data=SG[SG[,2]>=4.5 & SG[,2]<=7.5 ,])
f6.pred<-f6$fitted.values
#7 X = 5.5 to X = 8.5;
f7<-lm(Y~X,data=SG[SG[,2]>=5.5 & SG[,2]<=8.5 ,])
f7.pred<-f7$fitted.values

Yhat=c(f1.pred[2],f2.pred[2],f3.pred[2],f4.pred[2],f5.pred[2],f6.pred[2],f7.pred[2])
Xc=c(1:7)
data.frame(Xc,Yhat)
```

```
##   Xc     Yhat
## 2  1 131.6667
## 3  2 158.3333
## 4  3 187.0000
## 5  4 210.3333
## 6  5 245.3333
## 7  6 271.6667
## 8  7 319.0000
```

```r
plot(Xc,Yhat,type="l")
```



c-) Obtain the 95 percent confidence band for the true regression line and plot it on the plot prepared in part (b). Does the simplified lowess smooth fall entirely within the confidence band for the regression line? What does this tell you about the appropriateness of the linear regression function?

Yes, all values are within the confidence interval. The regression model is appropriate.

```
W <- sqrt( 2 * qf(0.95,2,8))
f.sg<-lm(Y~X,data=SG)
pred<-predict(f.sg,se.fit=TRUE)
out=cbind(SG$X,pred$fit-W*pred$se.fit,pred$fit + W*pred$se.fit )
out[2:8,]
```

```
##   [,1]     [,2]     [,3]
## 2    1 101.7362 146.3850
## 3    2 137.7823 175.3329
## 4    3 173.0775 205.0316
## 5    4 207.1764 235.9267
## 6    5 239.6733 268.4236
## 7    6 270.5684 302.5225
## 8    7 300.2671 337.8177
```

Problem 3

Refer to Plastic hardness Problem and data.(10 points, 5 points each)

a-) Obtain Bonferroni joint confidence intervals for β_0 and β_1, using a 90 percent family confidence coefficient. Interpret your confidence intervals.

Joint confidence intervals are

$162.90 \leq \beta_0 \leq 174.30$ $1.84 \leq \beta_1 \leq 2.23$

```
f<-lm(Y~X,data=PH)
confint(f,level=1-0.1/2)
```

```
##                2.5 %    97.5 %
## (Intercept) 162.9013 174.29875
## X             1.8405   2.22825
```

b-) What is the meaning of the family confidence coefficient in part (a)?

we conclude that β_0 is between 162.90 and 174.30 and β_1, is between 1.84 and 2.23. The family confidence coefficient is at least .90 that the procedure leads to correct pairs of interval estimates.

Problem 4

Refer to Plastic hardness Problem and data. (25 points)

a-) Management wishes to obtain interval estimates of the mean hardness when the elapsed time is 20, 30, and 40 hours, respectively. Calculate the desired confidence intervals. using the Bonferroni procedure and a 90 percent family confidence coefficient. What is the meaning of the family confidence coefficient here? (9 points)

Please see below for the joint confidence interval.

```
f<-lm(Y~X,data=PH)
X<-c(20,30,40)
predict.lm(f,data.frame(X = c(X)),interval = "confidence", level = 1-0.1/3)
```

```
##        fit      lwr      upr
## 1 209.2875 206.7277 211.8473
## 2 229.6312 227.6762 231.5863
## 3 249.9750 246.7824 253.1676
```

b-) Is the Bonferroni procedure employed in part (a) the most efficient one that could be employed here? Explain. (8 points)

No, Working Hotelling it more efficient. (B=2.36 and W=2.33)

```
B=qt(1-0.1/(2*3),14)
W=sqrt(2 *qf(0.90,2,14))
cbind(B,W)
```

```
##            B        W
## [1,] 2.35982 2.335152
```

c-) The next two test items will be measured after 30 and 40 hours of elapsed time, respectively. Predict the hardness for each of these two items, using the most efficient procedure and a 90 percent family confidence coefficient. (8 points)

For this problem g=2, Bonferoni is 2.14 and Scheefe is 2.34. Bonferoni is more efficent. Please see below

```
B=qt(1-0.1/(2*2),14)
S=sqrt(2*qf(0.90,2,14))
cbind(B,S)
```

```
##             B        S
## [1,] 2.144787 2.335152
```

```
X<-c(30,40)
predict.lm(f,data.frame(X = c(X)),interval = "confidence", level = 1-0.1/2)
```

```
##        fit      lwr      upr
## 1 229.6312 227.8544 231.4081
## 2 249.9750 247.0733 252.8767
```

```
pred<-predict.lm(f,data.frame(X = c(X)),se.fit=TRUE)
s.pred<-sqrt(pred$se.fit^2+pred$residual.scale^2)
cbind(pred$fit-B*s.pred,pred$fit+B*s.pred)
```

```
##       [,1]     [,2]
## 1 222.4710 236.7915
## 2 242.4562 257.4938
```