**> # Question 1**

>

> pres<-read.table(file.choose(), header=TRUE, sep=",")

> attach(pres)

>

> # part A

>

> plot(TIME,COST\_CLM)

> 

> # we can see from the plot that it is increasing in the mean and also in the variance.

> # part B

>

> model1<-lm(COST\_CLM~TIME)

> summary(model1)

Call:

lm(formula = COST\_CLM ~ TIME)

Residuals:

Min 1Q Median 3Q Max

-1.64454 -0.55596 -0.00093 0.51839 1.74802

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.96830 0.18022 71.96 <2e-16 \*\*\*

TIME 0.23450 0.00454 51.65 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7349 on 66 degrees of freedom

Multiple R-squared: 0.9759, Adjusted R-squared: 0.9755

F-statistic: 2667 on 1 and 66 DF, p-value: < 2.2e-16

> # value of R^2 = 0.9759 (this is high, which is good)

> # value of R^2 adjusted = 0.9755 (this is high, which is good)

> # F p-value <2.2e-16 (very small, which means very good)

> # this means that TIME is a significant predictor of COST\_CLM

>

> #part C

>

> COST\_CLM1=COST\_CLM[-1]

> diff=COST\_CLM1-COST\_CLM[1:(length(COST\_CLM)-1)]

> plot(TIME[1:67],diff)

>



> # we can see that this plot does not have the linear trend.

> # The plot is stationary in the mean but is not stationary in the variance.

> #part D

>

> model2<-lm(diff~TIME[1:67])

> summary(model2)

Call:

lm(formula = diff ~ TIME[1:67])

Residuals:

Min 1Q Median 3Q Max

-1.45004 -0.37658 0.02294 0.36376 1.13098

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.097871 0.125097 0.782 0.437

TIME[1:67] 0.004196 0.003198 1.312 0.194

Residual standard error: 0.5063 on 65 degrees of freedom

Multiple R-squared: 0.0258, Adjusted R-squared: 0.01081

F-statistic: 1.721 on 1 and 65 DF, p-value: 0.1942

> # value of R^2 = 0.0258 (this is low, which is bad)

> # value of R^2 adjusted = 0.01081 (this is low, which is bad)

> # F p-value = 0.194 (high, which is bad)

> # part E

>

> sd(COST\_CLM)

[1] 4.693981

> sd(diff)

[1] 0.5090181

>

> # From parts B and D, I might think this is as a random walk model. The original data has a linear trend with time as the x, but the modified data does not. Also, as we can notice, the original data has higher sd than the modified data.

>

> # part F

>

> TIME2=TIME\*TIME

> model2<-lm(COST\_CLM~TIME+TIME2)

> summary(model2)

Call:

lm(formula = COST\_CLM ~ TIME + TIME2)

Residuals:

Min 1Q Median 3Q Max

-1.87358 -0.43601 0.04004 0.37880 1.24449

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.369e+01 2.511e-01 54.528 < 2e-16 \*\*\*

TIME 1.727e-01 1.679e-02 10.285 2.88e-15 \*\*\*

TIME2 8.959e-04 2.358e-04 3.799 0.000323 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6699 on 65 degrees of freedom

Multiple R-squared: 0.9802, Adjusted R-squared: 0.9796

F-statistic: 1612 on 2 and 65 DF, p-value: < 2.2e-16

> # the overall trend looks linear and I could not see any quadratic trend here.

But, the output that you have above has a significant squared term with a p-value = 0.000323.

> # part G

>

> COST\_CLM[68]+mean(diff)\*5

[1] 31.86506

>

>

> # part H

>

> l=5

> COST\_CLM[68]+mean(diff)\*l-sd(diff)\*2\*sqrt(l)

[1] 29.58866

> COST\_CLM[68]+mean(diff)\*l+sd(diff)\*2\*sqrt(l)

[1] 34.14145

>

> # 95% PI at time 73 = (29.58866, 34.14145)

>

> # part I

>

> lagCOST=COST\_CLM[-68]

>

> COST=COST\_CLM[-1]

>

> AR1=lm(lagCOST~COST)

> summary(AR1)

Call:

lm(formula = lagCOST ~ COST)

Residuals:

Min 1Q Median 3Q Max

-1.06877 -0.35935 -0.03756 0.35018 1.45804

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.25897 0.28646 0.904 0.369

COST 0.97639 0.01323 73.811 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5008 on 65 degrees of freedom

Multiple R-squared: 0.9882, Adjusted R-squared: 0.988

F-statistic: 5448 on 1 and 65 DF, p-value: < 2.2e-16

> # for this model, the value of r1 should be = 0.97639. We must determine the cutoff = 0.244. Since b1 = 0.97639 which is more than 0.244, this means AR1 is not an appropriate model.

>

> # part J

>

> e=COST-as.numeric(AR1$fitted)

> s=sqrt(sum((e-mean(e))^2)/(length(COST)-2))

>

> s=sqrt(sum((e-mean(e))^2)/(length(COST)-2))plot(c(1:67),e,xlim=c(0,70),ylim=c(-.7,.7),xlab="COST", ylab="",type="o",pch=16,axes=F)

Error: unexpected symbol in "s=sqrt(sum((e-mean(e))^2)/(length(COST)-2))plot"

> plot(c(1:67),e,xlim=c(0,70),ylim=c(-.7,.7),xlab="COST", ylab="",type="o",pch=16,axes=F)

> abline(h=0,lty=1)

> abline(h=c(-3\*s,3\*s),lty=2)

> xat=seq(0,70,by=10)

> yat=seq(-.7,.7,by=0.3)

> axis(1,at=xat,labels=as.character(xat),las=1)

> axis(2,at=yat,labels=as.character(yat),las=1)

> box()

>



> # we can still see obvious trend that present in the residuals. Also, we still have some points outside the +- 3 sd limits. This means AR1 is not a good model.

**QUESTION 2**

> sp500<-read.table(file.choose(),header=TRUE,sep=",")

> attach(sp500)

>

> # part A

>

> plot(YEAR,SPINDEX)

>



> # this plot seems like not stationary. It is increasing in mean and also in the variance.

> # part B

>

> SPINDEX1=SPINDEX[-1]

> diff=SPINDEX1-SPINDEX[1:(length(SPINDEX)-1)]

> plot(YEAR[1:283], diff)

>



> # this plot is stationary in the mean but it is not stationary in the variance.

> # part C

> # it is not a random walk model because the differenced series is not stationary in the variance.

**Question 3**

> spdata = read.csv(file.choose(),header=TRUE)

> sub = subset(spdata,caldt<20050103)

> dim(sub)

[1] 1256 2

> Index=1:1256

> attach(sub)

>

> # part A

> plot(Index,vwretd)

>



> # we can see from the graph that the mean of the sequence is constant and does not vary over time, but the variance appear to be larger in the first half and smaller at the end

> # So the sequence can be said to be stationary in the mean, but not the variance.

> # part B

>

> summary(vwretd)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-5.949e-02 -7.310e-03 8.276e-05 -6.940e-06 6.973e-03 5.754e-02

>

> sd(vwretd)

[1] 0.01279667

>

> # Assuming white noise, the forecast of an observation in the future is its sample mean (which is -0.00000694)

> # This forecast is independent of the number of steps ahead.

> # part C

>

> autocorr= acf(vwretd,lag.max=10,type="correlation")

> autocorr

Autocorrelations of series ‘vwretd’, by lag

0 1 2 3 4 5 6 7 8 9 10

1.000 -0.023 -0.045 -0.013 0.018 -0.032 -0.028 -0.035 0.012 -0.003 -0.023

>

> # cutoff = 0.0564

> # since all of these autocorrelations are < 0.0564, this means none of them are statistically significant from 0.

**Question 4**

# Part A

S = 1.67

Y = 500

N = 23

Y1 = 137

c = (137-500)/23

= -15.78

Y2 = 137 + 23(-15.78)

= $ -226

# part B

= -226 + 2(1.67)\* = -209.98

= -226 - 2(1.67)\* = -242.01

= $ (-242.01, -209.98)