Problem 1 (25 points)
On the Golub et al. (1999) data set, find the expression values for the GRO2 GRO2 oncogene and the GRO3 GRO3 oncogene. (Hint: Use grep() to find the gene rows in golub.gnames. Review module 2, or page 12 of the textbook on how to do this. Be careful to search *only in the column with gene names.*)
(a) Find the correlation between the expression values of these two genes.

(b) Find the parametric 90% confident interval for the correlation with cor.test(). (Hint: use ?cor.test to learn how to set the confidence level different from the default value of 95%.)

(c) Find the bootstrap 90% confident interval for the correlation.

(d) Test the null hypothesis that correlation = 0.64 against the one-sided alternative that correlation > 0.64 at the $\alpha = 0.05$ level. What is your conclusion? Explain you reasoning supported by the appropriate R outputs.

A)
Rscript:
```
library(multtest)
data(golub)
grep("GRO2 GRO2",golub.gnames[,2])
grep("GRO3 GRO3",golub.gnames[,2])
cor(golub[2714,],golub[2715,])
```

Answer:

Correlation between the expression values of these two genes are: 0.7966283

B)
Rscript:
```
library(multtest)
data(golub)
grep("GRO2 GRO2",golub.gnames[,2])
grep("GRO3 GRO3",golub.gnames[,2])
cor.test(golub[2714,],golub[2715,], conf.level=0.9)
```

Answer:
```
Pearson's product-moment correlation

data:  golub[2714, ] and golub[2715, ]
t = 7.9074, df = 36, p-value = 2.201e-09
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.6702984 0.8780861
sample estimates:
```

0.7966283

parametric 90% confident interval for the correlation with cor.test() is 0.6702984 0.8780861

3)
Rscript:

```
data(ALL,package="ALL");library(ALL)
nboot <- 2000
data<-cbind(golub[2714,],golub[2715,])
boot.cor<-matrix(0,nrow=nboot,ncol=1)
for (i in 1:nboot){
  dat.star<-data[sample(1:nrow(data),replace=TRUE),]
  boot.cor[i,] <- cor(dat.star[,1],dat.star[,2])
}
quantile(boot.cor[,1],c(0.05,0.95))
```

Answer:

```
5%      95%
0.5757196 0.8977350
```

bootstrap 90% confident interval for the correlation is 0.5757196 0.8977350

4)
Rscript:

```
data(ALL,package="ALL");library(ALL)
nboot <- 2000
data<-cbind(golub[2714,],golub[2715,])
boot.cor<-matrix(0,nrow=nboot,ncol=1)
for (i in 1:nboot){
  dat.star<-data[sample(1:nrow(data),replace=TRUE),]
  boot.cor[i,] <- cor(dat.star[,1],dat.star[,2])
}
quantile(boot.cor[,1],c(0.0,0.95))
```

Answer:
95% confidence interval for the correlation is 0.5177840 0.9051295. Since 0.64 lies in this confidence interval we will accept the null hypothesis.

Problem 2 (25 points)
On the Golub et al. (1999) data set, we consider the correlation between the Zyxin gene expression values and each of the gene in the data set.
(a) How many of the genes have correlation values less than negative 0.5? (Those genes are highly negatively correlated with Zyxin gene).

(b) Find the gene names for the top five genes that are most negatively correlated with Zyxin gene.

(c) Using the t-test, how many genes are negatively correlated with the Zyxin gene? Use a false discovery rate of 0.05. (Hint: use cor.test() to get the p-values then adjust for FDR. Notice that we want a one-sided test here.)

A)
Rscript:
```
data(golub,package='multtest')
Zyxin<-grep("Zyxin",golub.gnames[,2])
correlation<-apply(golub,1, function(x) cor(x,golub[Zyxin,]))
sum(correlation<(-0.5))
```

Answer:
85 genes have correlation values less than negative 0.5

B)
Rscript:
```
data(golub,package='multtest')
Zyxin<-grep("Zyxin",golub.gnames[,2])
correlation<-apply(golub,1, function(x) cor(x,golub[Zyxin,]))
ordered <-order(correlation,decreasing="False")
print(golub.gnames[ordered[1:5],2])
```

Answer:
Gene names for the top five genes that are most negatively correlated with Zyxin gene:
[1] "Macmarcks"
[2] "Inducible protein mRNA"
[3] "C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds"
[4] "Oncoprotein 18 (Op18) gene"
[5] "54 kDa protein mRNA"

C)

Rscript:

```
data(golub,package='multtest')
Zyxin<-grep("Zyxin",golub.gnames[,2])
pvalue<-apply(golub,1, function(x) cor.test(x,golub[2124,],alternative="less")$p.value)
sum(pvalue<0.05)
pfdr<-p.adjust(pvalue, method='fdr')
sum(pfdr<0.05)
```

Answer:

572 genes are negatively correlated with the Zyxin gene. 142 genes are negatively correlated with zyxin gene after fdr adjustments.

Problem 3 (30 points)
On the Golub et al. (1999) data set, regress the expression values for the GRO3 GRO3 oncogene on the expression values of the GRO2 GRO2 oncogene.
(a) Is there a statistically significant linear relationship between the two genes' expression? Use appropriate statistical analysis to make the conclusion. What proportion of the GRO3 GRO3 oncogene expression's variation can be explained by the regression on GRO2 GRO2 oncogene expression?

(b) Test if the slope parameter is less than 0.5 at the $\alpha = 0.05$ level.

(c) Find an 80% prediction interval for the GRO3 GRO3 oncogene expression when GRO2 GRO2 oncogene is not expressed (zero expression value).

(d) Check the regression model assumptions. Can we trust the statistical inferences from the regression fit?

A)
Rscript:
```
library(multtest)
data(golub)
grep("GRO2 GRO2",golub.gnames[,2])
grep("GRO3 GRO3",golub.gnames[,2])
summary(lm(golub[2715,] ~ golub[2714,]))
```

Answer:
```
Call:
lm(formula = golub[2715, ] ~ golub[2714, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.78038 -0.10639 -0.00553  0.14225  0.96298

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.84256    0.05941 -14.182 2.62e-16 ***
golub[2714, ]  0.35820    0.04530   7.907 2.20e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3201 on 36 degrees of freedom
Multiple R-squared:  0.6346,            Adjusted R-squared:  0.6245
F-statistic: 62.53 on 1 and 36 DF,  p-value: 2.201e-09
```

As we see in the above summary for H0: $\beta_0=0$ and H0: $\beta_1=0$ p values are 2.62e-16 and 2.20e-09 respectively. Both the p values are less than 0.05 so we reject the null hypothesis and conclude that there is a statistically significant linear relationship between the two genes' expression.

Proportion of the GRO3 GRO3 oncogene expression's variation can be explained by the regression on GRO2 GRO2 oncogene expression by the R squared value, which is 0.6346.

B)
Rscript:

```
library(multtest)
data(golub)
grep("GRO2 GRO2",golub.gnames[,2])
grep("GRO3 GRO3",golub.gnames[,2])
reg.fit<-lm(golub[2715,] ~ golub[2714,])
reg.fit
confint(reg.fit, level=0.95)
```

Answer:
Slope parameter is 0.3582, which is less than 0.5. 95 % confidence interval is 0.2663291 0.4500727 and .3582 lies in this confidence interval. So we can conclude that slope parameter is less than 0.5 at the α = 0.05 level.

C)
Rscript:

```
library(multtest)
data(golub)
grep("GRO2 GRO2",golub.gnames[,2])
grep("GRO3 GRO3",golub.gnames[,2])
GRO3<-golub[2715,]
GRO2<-golub[2714,]
reg.fit<-lm(GRO3 ~ GRO2)
predict(reg.fit,newdata=data.frame(GRO2=0),interval="prediction",level=.8)
```

Answer:

```
        fit        lwr        upr
1 -0.842559 -1.267563 -0.4175553
```

80% prediction interval for the GRO3 GRO3 oncogene expression when GRO2 GRO2 oncogene is not expressed is -1.267563 -0.4175553

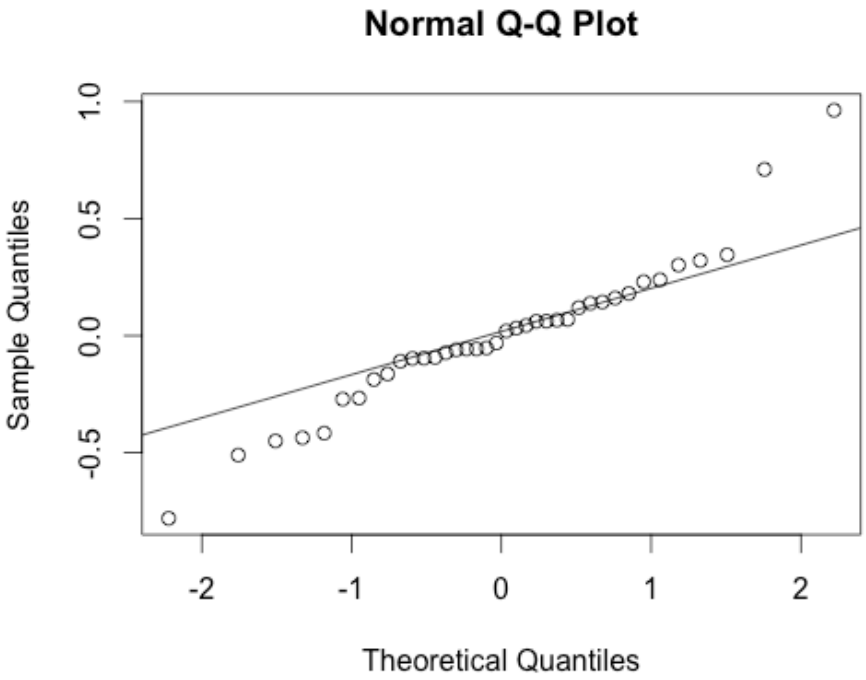
D)
Rscript:

```
library(multtest)
data(golub)
GRO3<-golub[2715,]
GRO2<-golub[2714,]
reg.fit<-lm(GRO3 ~ GRO2)
qqnorm(resid(reg.fit))
qqline(resid(reg.fit))

shapiro.test(resid(reg.fit))
```

plot(reg.fit,which=1)
var.test(GRO3,GRO2)

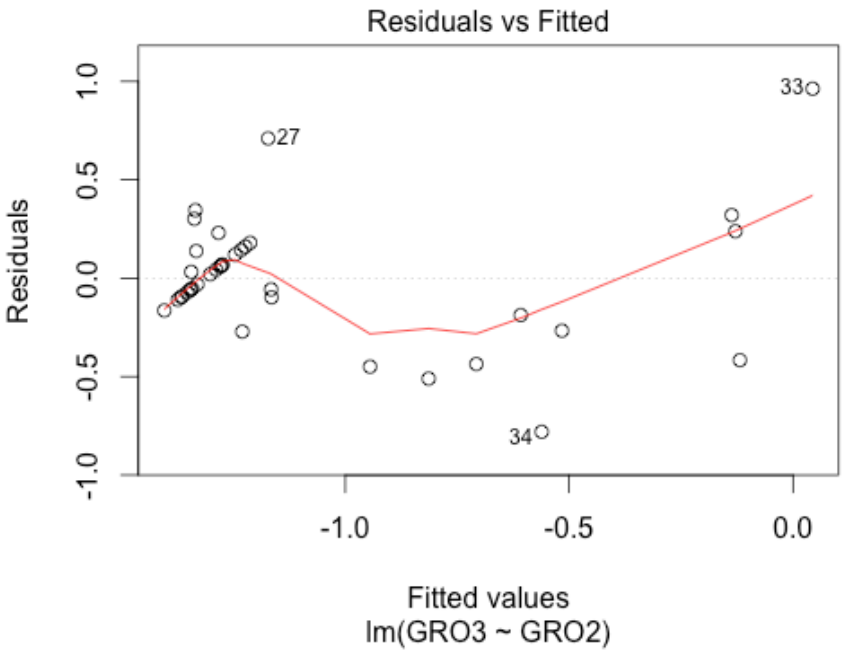
Answer:



Shapiro-Wilk normality test

data: resid(reg.fit)
W = 0.9478, p-value = 0.07532

As p value here is greater than 0.05 we conclude that the data is normally distributed. So first assumption is met.



In this residuals vs fitted value graph we can check that residuals values have increased going right which means variance is increasing thus we can conclude homoscedasticity assumption is not meeting.

We can confirm this by var.test.
F test to compare two variances

data: GRO3 and GRO2
F = 0.2022, num df = 37, denom df = 37, p-value = 4.123e-06
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1050779 0.3890205
sample estimates:
ratio of variances
       0.2021817

As we see here p value is less than 0.05 so we reject the null hypothesis of equal variances and conclude that variances are not equal in this data.

I think that because one of the assumption is not meeting here so we can say that statistical inferences from the regression fit cannot be trusted.

Problem 4 (20 points)
For this problem, work with the data set stackloss that comes with R. You can get help on the data set with ?stackloss command. That shows you the basic information and source reference of the data set. Note: it is a data frame with four variables. The variable stack.loss contains the ammonia loss in a manufacturing (oxidation of ammonia to nitric acid) plant measured on 21 consecutive days. We try to predict it using the other three variables: air flow (Air.Flow) to the plant, cooling water inlet temperature (C) (Water.Temp), and acid concentration (Acid.Conc.)
(a) Regress stack.loss on the other three variables. What is the fitted regression equation?

(b) Do all three variables have statistical significant effect on stack.loss? What proportion of variation in stack.loss is explained by the regression on the other three variables?

(c) Find a 90% confidence interval and 90% prediction interval for stack.loss when Air.Flow=60, Water.Temp=20 and Acid.Conc.=90.

A)
Rscript:
```
data(stackloss)
the.data <- as.data.frame(stackloss[,c('Air.Flow', 'Water.Temp', 'Acid.Conc.', 'stack.loss')])
lin.reg<-lm(stack.loss~Air.Flow +Water.Temp+Acid.Conc., data=the.data)
lin.reg
```

Answer:
```
Coefficients:
(Intercept)    Air.Flow   Water.Temp   Acid.Conc.
  -39.9197      0.7156      1.2953     -0.1521
```

Fitted regression equation is:
Stack.loss= -39.9197 + 0.7156 Air.Flow + 1.2953 Water.Temp – 0.1521 Acid.Conc.

B)
Rscript:
```
data(stackloss)
the.data <- as.data.frame(stackloss[,c('Air.Flow', 'Water.Temp', 'Acid.Conc.', 'stack.loss')])
summary(lm(stack.loss~Air.Flow +Water.Temp+Acid.Conc., data=the.data))
```

Answer:
```
Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
   data = the.data)

Residuals:
   Min     1Q  Median     3Q    Max
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-39.9197	11.8960	-3.356	0.00375	**
Air.Flow	0.7156	0.1349	5.307	5.8e-05	***
Water.Temp	1.2953	0.3680	3.520	0.00263	**
Acid.Conc.	-0.1521	0.1563	-0.973	0.34405	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared: 0.9136, Adjusted R-squared: 0.8983
F-statistic: 59.9 on 3 and 17 DF, p-value: 3.016e-09

P value for Air.Flow and Water.Temp is 5.8e-05 and 0.00263 respectively, which is less than 0.05 hence we reject the null hypothesis and conclude that Air.Flow and Water.Temp do have statistically significant effect on stack.loss

On the other hand pvalue for Acid.Conc. is 0.34405, which is greater than 0.5 hence we fail to reject null hypothesis and conclude that Acid.Conc do not have statistically significant effect on stack.loss.

proportion of variation in stack.loss is explained by the regression on the other three variables is 0.9136.

C)

Rscript:

```
data(stackloss)
the.data <- as.data.frame(stackloss[,c('Air.Flow', 'Water.Temp', 'Acid.Conc.', 'stack.loss')])
lin.reg<-lm(stack.loss~Air.Flow +Water.Temp+Acid.Conc., data=the.data)
predict(lin.reg,newdata=data.frame(Air.Flow=60,Water.Temp=20,Acid.Conc.=90),
interval="confidence",level=.9)
predict(lin.reg,newdata=data.frame(Air.Flow=60,Water.Temp=20,Acid.Conc.=90),
interval="prediction",level=.9)
```

Answer:

CI

	fit	lwr	upr
1	15.23343	13.50069	16.96617

PI

fit	lwr	upr
1 15.23343	9.331184	21.13568

90% confidence interval = 13.50069 16.96617

90% prediction interval = 90% prediction interval