**Module 9 Home Work**

**Problem 1: (25 points)**

On the Golub et al. (1999) data set, find the expression values for the GRO2 GRO2oncogene 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 α = 0.05 level. What is your conclusion? Explain you reasoning supported by the appropriate R outputs.

**Answer:**

data(golub,package="multtest")

grep("GRO2 GRO2",golub.gnames[,2])

grep("GRO3 GRO3",golub.gnames[,2])

GRO2a <- golub[2714,]

GRO3a <- golub[2715,]

(a)

cor.results <- cor.test(GRO2a, GRO3a)

cor.results

**Output:**

Correlation between the expression values of these two genes=0.7966283

(b)

cor.results.conf <- cor.test(GRO2a, GRO3a, conf.level = 0.90)

cor.results.conf

**Output:**

The parametric 90% confident interval = 0.6702984 0.8780861

(c)

nboot <- 2000

boot.cor <- matrix(0, nrow=nboot, ncol = 1)

data <- cbind(GRO2a, GRO3a)

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))

**Output:**

The bootstrap 90% confident interval for the correlation = 0.5907881 0.8949947

(d)

nboot <- 2000

boot.cor <- matrix(0, nrow=nboot, ncol = 1)

data <- cbind(GRO2a, GRO3a)

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(0.025,0.975))

**Output:**

2.5% 97.5%

0.5268934 0.9060638

Here we can accept the null hypothesis for the correlation, as 0.64 is within the Confidence Interval (alpha=0.05)

**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.

Answer:

(a)

data(golub,package="multtest")

grep("Zyxin",golub.gnames[,2])

Zyxin\_data <- golub[2124,]

cor.data <- apply(golub, 1, function(x) cor.test(x, Zyxin\_data)$estimate)

results.a <- sum(cor.data < -0.5)

results.a

**Output:**

There are 85 genes that have correlation values less than negative 0.5

(b)

order.cor <- order(cor.data, decreasing=FALSE)

results.b <- golub.gnames[order.cor[1:5],2]

results.b

**Output:**

The gene names for the top five genes that are most negatively correlated with Zyxin gene are

[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)

cor.ttest <- apply(golub, 1, function(x) cor.test(x, Zyxin\_data, alternative = "l")$p.value)

results.c <- sum(cor.ttest < 0.05)

results.c

cor.fdr <- p.adjust(p = cor.ttest, method = "fdr")

results.fd <- sum(cor.fdr < 0.05)

results.fd

**Output:**

Using the t-test, it was found that there are 572 genes that are negatively correlated

with the Zyxin gene

After the FDR adjustment there are: 142

**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 α = 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?

Answer:

data(golub,package="multtest")

GRO2a<- golub[2714,]

GRO3a<- golub[2715,]

(a)

reg.fit <- lm(GRO3a ~ GRO2a)

reg.fit

summary(reg.fit)

**Output:**

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 \*\*\*

GRO2a 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 can see the p-values are < 0.05, we can reject the Null hypothesis i.e., beta=0. Hence, the relationship between the 2 genes expressions is statistically significant.

2. Proportion of the GRO3 GRO3 oncogene expression’s variation can be explained by the regression on GRO2 GRO2 oncogene expression = 0.6346.

(b)

data(golub,package="multtest")

GOR2a<- golub[2714,]

GOR3a<- golub[2715,]

lin1<- lm(GOR3a ~ GOR2a)

confint(lin1, level = 0.95)

**Output:**

GOR2a 0.2663291 0.4500727

Here, the slope parameter is less than 0.05

(c)

data(golub,package="multtest")

GOR2a<- golub[2714,]

GOR3a<- golub[2715,]

lin1<- lm(GOR3a ~ GOR2a)

data1 <- data.frame(GOR2a = 0)

predict(lin1,data1,interval="prediction",level= 0.80)

**Output:**

fit lwr upr

1 -0.842559 -1.267563 -0.4175553

80% prediction interval = -1.267563 -0.4175553

(d)

data(golub,package="multtest")

GOR2a<- golub[2714,]

GOR3a<- golub[2715,]

lin1<- lm(GOR3a ~ GOR2a)

plot(lin1,which=2)

shapiro.test(resid(lin1))

plot(lin1,which= 1)

**Output:**



from the plot we can see that the Q-Q line is normal (mainly central data)

Shapiro test to confirm p-value:

Shapiro-Wilk normality test

data: resid(lin1)

W = 0.94779, p-value = 0.07532

Here the p-value < 0.05. Null hypothesis of normal distribution can be accepted. The assumptions are not always true so I say that we can’t completely trust the statistical inference.

From the plot it can be observed that there is no Non-linear mean patterns are observed. And also the variance seem to be different at different points. It is violating Homoscedasticity.



**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.

**Answer:**

(a)

data(stackloss)

str(stackloss)

stackloss.data <- as.data.frame(stackloss[,c('Air.Flow', 'Water.Temp', 'Acid.Conc.', 'stack.loss')])

lin.re <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc., data=stackloss.data)

summary(lin.re)

**Output:**

The fitted regression is:-­‐39.92+0.72 Air.Flow +1.3 Water.Temp­‐0.15 Acid.Conc

(b)

From the output of 4(a)

1. P-­‐values for Air.flow=5.8e-­‐05 and Water.Temp =0.00263, they have significant effect on stack.loss as p<0.05.

2. P-­‐value for Acid.Conc =0.34405.This has a significant effect on stack.loss as the p-­‐value>0.05

3. The proportion of variation In stack.loss that can be explained by the regression on the other three values= R-­‐squared=0.9136

(c)

g.data <- data.frame(Air.Flow = 60, Water.Temp = 20, Acid.Conc. = 90)

predict(lin.re, given.data, interval="confidence", level = 0.90)

predict(lin.re, given.data, interval="prediction", level = 0.90)

**Output:**

fit lwr upr

1 15.23343 13.50069 16.96617

90% Confidence Interval =13.50069 16.96617

fit lwr upr

1 15.23343 9.331184 21.13568

90% Prediction Interval = 9.331184 21.13568