Math7340 HW9

Chengbo Gu

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.

(a) Find the correlation between the expression values of these two genes.

```
data(golub, package="multtest")
GRO2index <- grep("GRO2 GRO2 oncogene", golub.gnames[,2])
GRO3index <- grep("GRO3 GRO3 oncogene", golub.gnames[,2])
GRO2 <- golub[GRO2index,]
GRO3 <- golub[GRO3index,]
cor(GRO2, GRO3)
```

```
## [1] 0.7966283
```

(b) Find the parametric 90% confident interval for the correlation with cor.test().

```
cor.test(x=GRO2, y=GRO3, conf.level = 0.90)
```

```
##
##  Pearson's product-moment correlation
##
## data:  GRO2 and GRO3
## 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:
##       cor
## 0.7966283
```

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

```
nboot <- 2000
boot.cor <- rep(NA, nboot)
data <- cbind(GRO2, GRO3)
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, c(0.05,0.95))
```

```
##        5%       95%
## 0.6069760 0.9010593
```

(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.

```
#cor.test(x=GRO2, y=GRO3, alternative = "greater", conf.level = 0.95)
quantile(boot.cor, 0.05)
```

```
##        5%
## 0.606976
```

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).

```
data(golub, package="multtest")
Zyxindex <- grep("Zyxin", golub.gnames[,2])
Zyx <- golub[Zyxindex,]
cor.res <- apply(golub, 1, function(x) cor(Zyx, x))
length(cor.res[cor.res < -0.5])
```

```
## [1] 85
```

```
#sum(apply(golub, 1, function(x) cor(Zyx, x)) < -0.5)
```

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

```
golub.gnames[,2][order(cor.res)][1:5]
```

```
## [1] "Macmarcks"
## [2] "Inducible protein mRNA"
## [3] "C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternati
## [4] "Oncoprotein 18 (Op18) gene"
## [5] "54 kDa protein mRNA"
```

(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.)

```
p.values <- apply(golub, 1, function(x) cor.test(Zyx, x, alternative = "less")$p.value)
p.fdr <- p.adjust(p=p.values, method="fdr")
sum(p.fdr<0.05)
```

```
## [1] 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.

```
data(golub, package="multtest")
GRO2index <- grep("GRO2 GRO2 oncogene", golub.gnames[,2])
GRO3index <- grep("GRO3 GRO3 oncogene", golub.gnames[,2])
GRO2 <- golub[GRO2index,]
GRO3 <- golub[GRO3index,]
reg.fit <- lm(GRO3 ~ GRO2)
summary(reg.fit)
```

```
##
## Call:
## lm(formula = GRO3 ~ GRO2)
##
## 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 ***
## GRO2         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
```

(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?

The summary provides the two-sided t-statistics and two-sided p-values for testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. The p-values of 2.62e-16 and 2.20e-09 are very small, so that we conclude both the intercept β_0 and the slope β_1 are nonzero. Thus, there is statistically significant linear relationship between the two genes' expression.

R-squared here is 0.6346, so 63.46% 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.

The one-sided uppder confidence interval of β_1 is $(-\infty, \beta_1 + t_{1-\alpha,n-2} SE_{\beta_1})$.

We can calculate it both using built-in function and manully.

```
# Using confint
confint(reg.fit, level=0.9)
```

```
##                    5 %       95 %
## (Intercept) -0.9428580 -0.7422600
## GRO2         0.2817217  0.4346801
```

```
# Using formula
0.35820 + qt(0.95, length(GRO2 <- golub[GRO2index,])-2)*0.04530
```

```
## [1] 0.4346799
```

Thus, the 95% CI for the slope β_1 is $(-\infty, 0.43468)$ which indicates that 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).

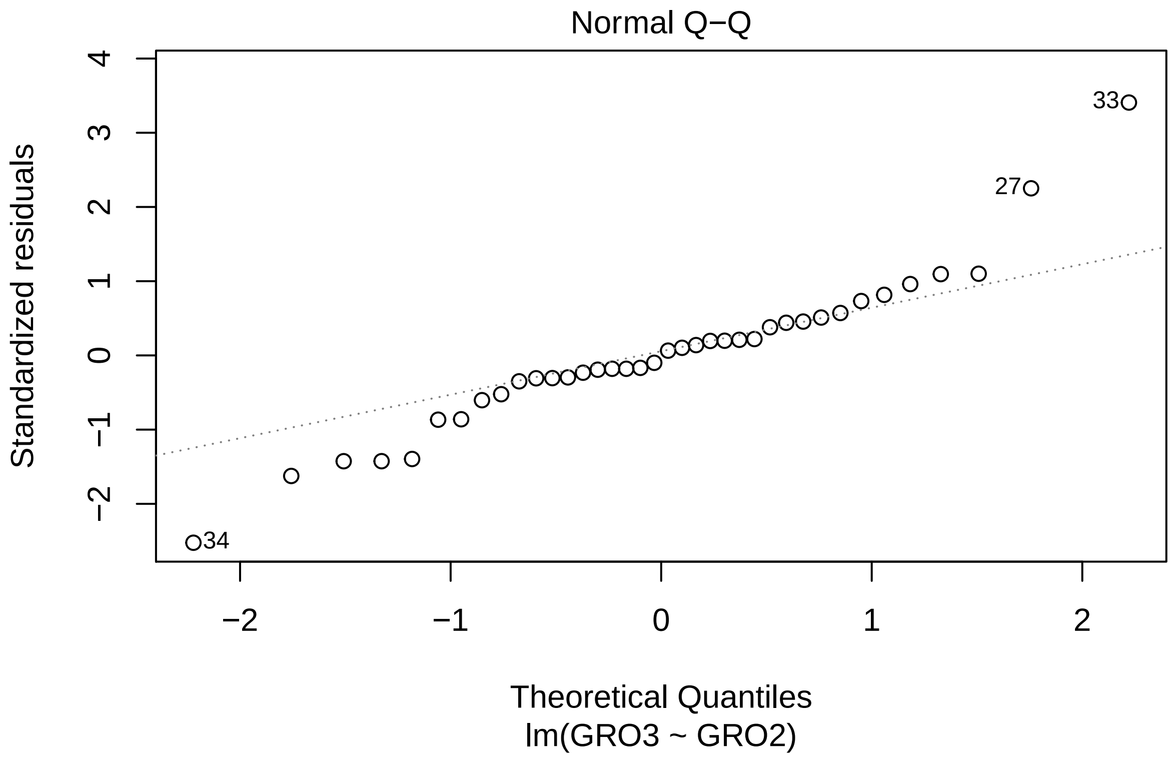
```
predict(reg.fit, newdata=data.frame(GRO2=0), interval="prediction", level = 0.80)
```

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

Thus, the 80% PI for GRO3 given GRO2 = 0 is $(-1.267563, -0.4175553)$.

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

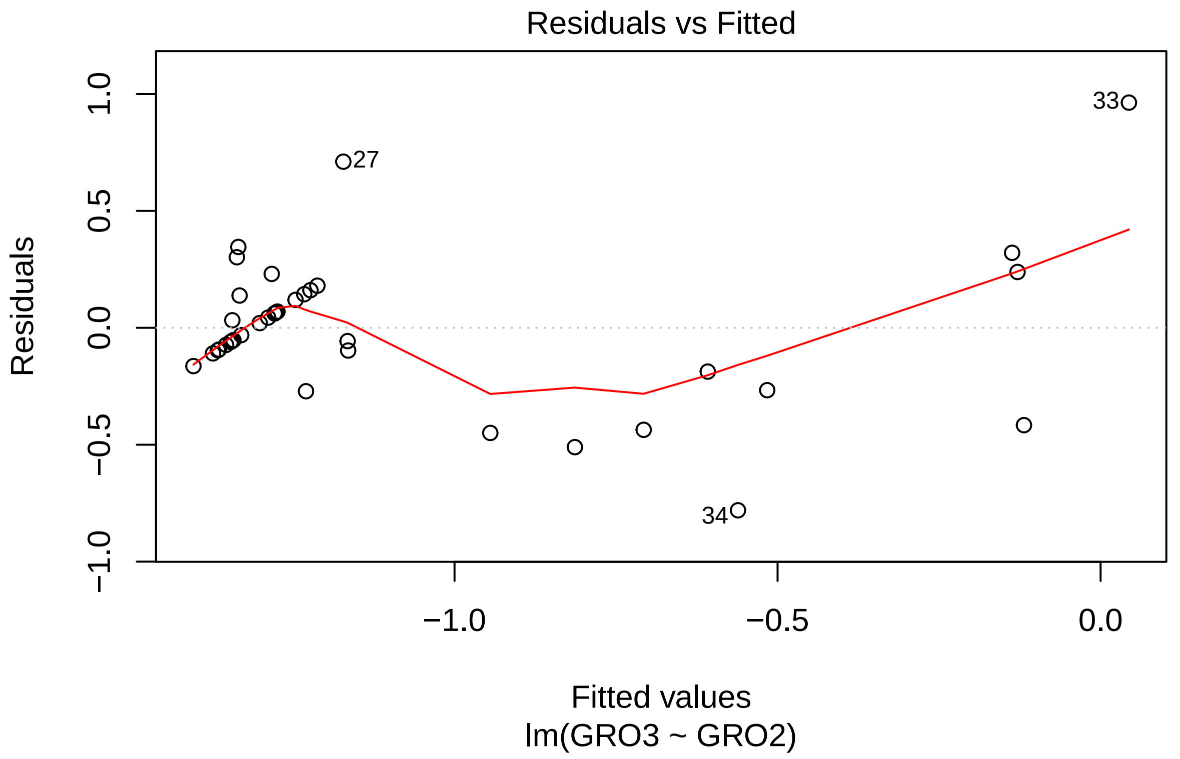
```
plot(reg.fit, which=2)
```



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

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(reg.fit)
## W = 0.94779, p-value = 0.07532
```

```
#library(lmtest)
plot(reg.fit, which=1)
```



```
bptest(reg.fit, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  reg.fit
## BP = 23.926, df = 1, p-value = 1.001e-06
```

Normality assumption holds because the p-value of shapiro test is greater than 0.05. Homoscedasticity assumption is violated which could be concluded both from the p-value of bp-test and the plot.

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?

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

```
##
## Call:
## lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
##     data = stackloss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2377 -1.7117 -0.4551  2.3614  5.6978
##
## 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
```

The fitted regression equation is stack.loss = -39.9197 + 0.7156Air.Flow + 1.2953Water.Temp - 0.1521Acid.Conc.

(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?

Air.Flow and Water.Temp have statistical significant effect on stack.loss because of their small p-values which are less than 0.05. Acid.Conc. doesn't have such effect because p-value = 0.34405.

91.36% 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.

```r
predict(lin.reg, newdata=data.frame(Air.Flow=60, Water.Temp=20, Acid.Conc.=90), interval="confidence")
```

```
##        fit      lwr      upr
## 1 15.23343 13.13195 17.33492
```

```r
predict(lin.reg, newdata=data.frame(Air.Flow=60, Water.Temp=20, Acid.Conc.=90), interval="prediction")
```

```
##        fit      lwr      upr
## 1 15.23343 8.075115 22.39175
```

90% CI is (13.13195, 17.33492).

90% PI is (8.075115, 22.39175).