Homework 6

Thu Nguyen

26 May, 2019

Problem 1.

Visit the Gene Expression Omnibus (GEO) website. Find a dataset of some interest to you with two groups to be compared in terms of their gene expression profiles. Specify what dataset you choose to work with, and describe it in your own words in a paragraph or so. Then apply a two-sample test to each gene (or nucleotide sequence), and correct for multiple testing by applying a method for FDR control. Set the desired level at 20% (which is not particularly large for FDR). Briefly comment on your findings.

Data set used is `GSE7621.csv`. Title: Parkinson's disease: substantia nigra. Summary: Analysis of substantia nigrae from postmortem brains of patients with Parkinson's disease (PD). Neurons in the substantia nigra, which produces dopamine, degenerate in PD. Results provide insight into the molecular pathogenesis of PD.

```
# Problem 1  ----------------------------------------------------
# Load data
data <- read.table('GSE7621.csv', header = TRUE)
# Remove Index column
data <- data[,-1]
```

Multiple testings for effects of each gene

```
# Number of observations
m <- nrow(data)
# Vector to store p-values for each observation
pval <- numeric(m)
# Pairwise t-test
for (i in 1:m) {
  pval[i] <- t.test(data[i,1:9], data[i, -(1:9)])$p.value
}
# Number of Rejects without Correction
R <- sum(pval <= .2)
mes <- 'Number of Rejects without Correction:'
print(cat(mes, R, '\n'))
```

```
Number of Rejects without Correction: 16102
NULL
```

p-value correction for 20% FDR control

```
# p-value corrections
# Benjamini-Hochberg correction
pval_bh <- p.adjust(pval, 'BH')
R_bh <- sum(pval_bh <= .2)
mes <- 'Number of Rejects with Benjamini-Hochberg Correction: '
print(paste0(mes, R_bh))
```

```
[1] "Number of Rejects with Benjamini-Hochberg Correction: 879"
```

```r
# Benjamini-Yekutieli correction
pval_by <- p.adjust(pval, 'BY')
R_by <- sum(pval_by <= .2)
mes <- 'Number of Rejects with Benjamini-Yukitieli Correction: '
print(paste0(mes, R_by))
```

```
[1] "Number of Rejects with Benjamini-Yukitieli Correction: 3"
```

Problem 2

Go to the following webpage. The dataset was extracted from the China Health and Nutrition Survey.1 Download the dataset. Let's focus on the relation between systolic and diastolic blood pressure.

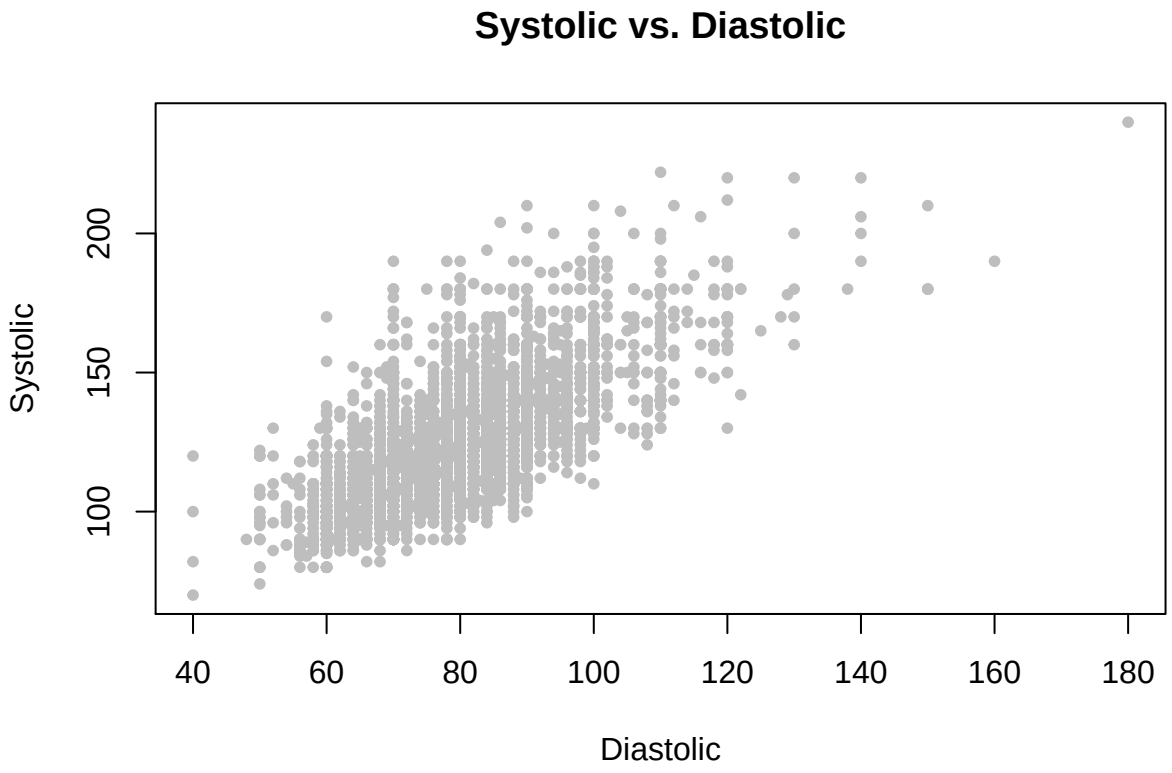
A. Plot the data. Make it nice.
B. Perform some test for association. Name the test you are performing and specify the null hypothesis that it is testing. Briefly comment on your findings.
C. Fit a line by least squares. Produce 90% confidence intervals for the slope and intercept. What assumptions do these rely on? Add the line to the plot above.
D. Would you recommend fitting a polynomial of degree 2 instead? Explain.

Part A

Plot of the data

```
# Problem 2  ----------------------------------------------------
# Part A     ----------------------------------------------------
data <- read.csv('dataset-chns-2006-subset3.csv')
# Rename column 1
colnames(data)[1] <- 'Age'

# Plot
plot(data$diastolic, data$systolic, cex=1, pch=20, col='grey',
     main='Systolic vs. Diastolic',
     xlab = 'Diastolic', ylab = 'Systolic')
```



Part B: Test for Association

χ^2-test

$$H_0 : \text{Distribution of Systolic is independent of that of Diastolic}$$
$$H_1 : \text{otherwise}$$

```
mytable <- table(data$systolic, data$diastolic)
chisq.test(mytable)
```

```
    Pearson's Chi-squared test

data:  mytable
X-squared = 76300, df = 7242, p-value < 2.2e-16
```

Given a very small p-value, for any reasonable value of α, we would reject H_0: it appears that they are not independent, and that there is an association between them.

However, given that the contigency table contains a lot of 0, the χ^2-test might not have been the best. Instead, we can look at if the averages of Systolic and Diastolic are the same.

t-test

$$H_0 : \text{the averages are the same}$$
$$H_1 : \text{otherwise}$$

```
t.test(data$systolic, data$diastolic, alternative = 'two.sided')
```

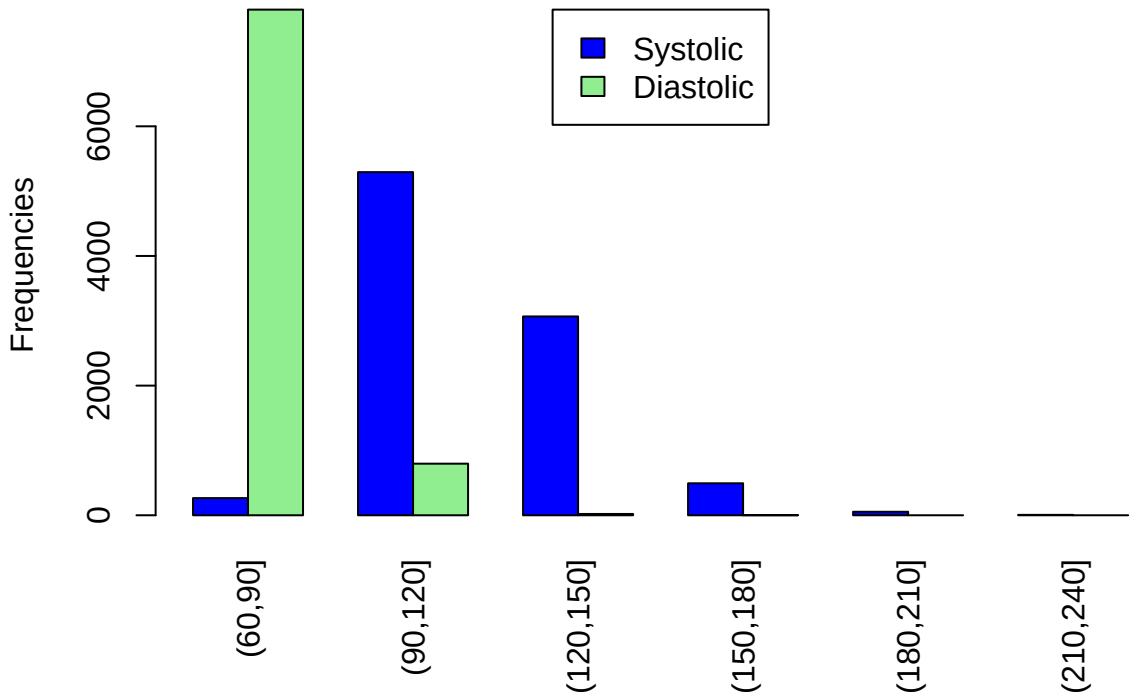
```
    Welch Two Sample t-test

data:  data$systolic and data$diastolic
t = 193.35, df = 15221, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 42.50951 43.38022
sample estimates:
mean of x mean of y
121.79375  78.84888
```

The t-test also returns a very small p-value. Hence, it appears that there are differences between the two.

Also, below is the histogram of the distributions of Systolic and Diastolic

```
cuts <- seq(60, 240, by = 30)
counts <- matrix(NA, 2, length(cuts)-1)
counts[1,] <- as.numeric(table(cut(data$systolic, cuts)))
counts[2,] <- as.numeric(table(cut(data$diastolic, cuts)))
rownames(counts) <- c('Systolic', 'Diastolic')
colnames(counts) <- names(table(cut(data$systolic, cuts)))
barplot(counts, legend=TRUE, beside=TRUE, args.legend = list(x='top'),
        col = c('blue', 'lightgreen'), las = 3,
        ylab = 'Frequencies')
```



Part C: Simple Linear Regression Model

Linear Regression 90% Confidence Interval

```
# Linear Regression Model
linreg <- lm(data$systolic ~ data$diastolic)
confint(linreg, level = .9)
```

```
                   5 %      95 %
(Intercept)    28.16103 31.276281
data$diastolic  1.14818  1.187302
```

```
# 90% Confidence Interval for the Slope
mes <- '90% Confidence interval for the slope: '
ci <- signif(confint(linreg, level = .9)[2,], 3)
print(paste0(mes, '(', ci[1], ',', ci[2], ')'))
```

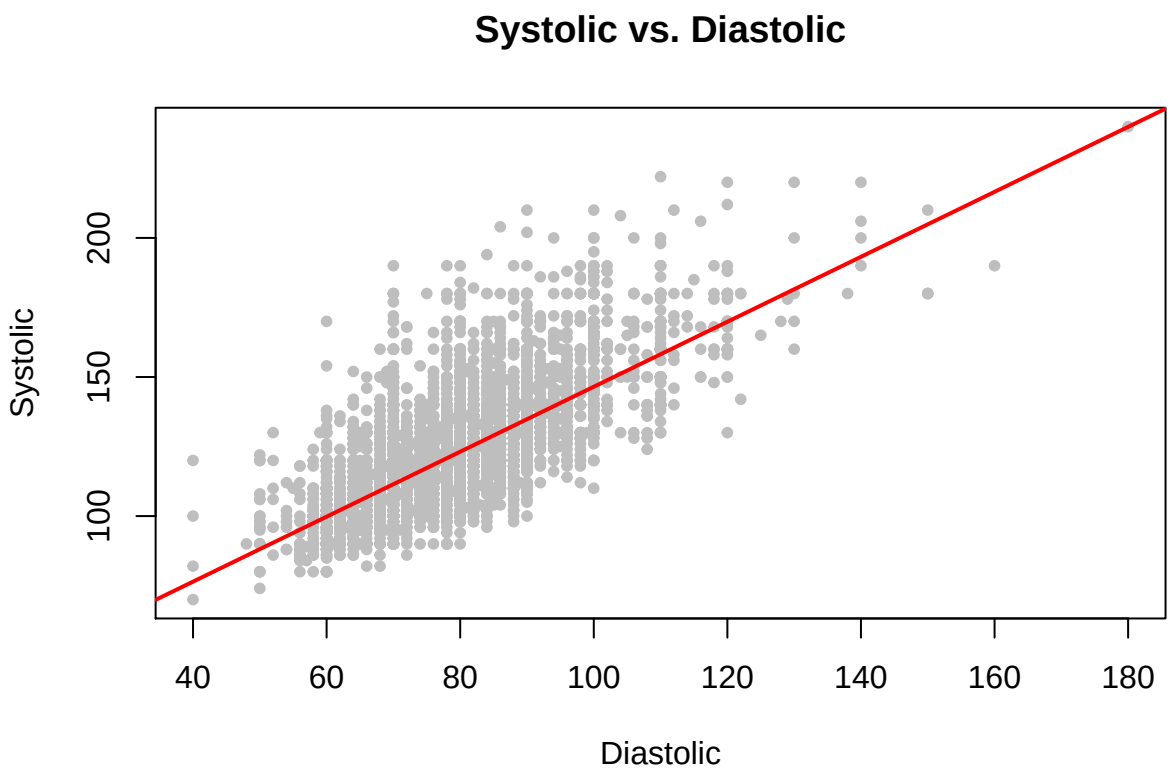
```
[1] "90% Confidence interval for the slope: (1.15,1.19)"
```

```
# 90% Confidence Interval for the Intercept
mes <- '90% Confidence interval for the Intercept: '
ci <- signif(confint(linreg, level = .9)[1,], 4)
print(paste0(mes, '(', ci[1], ',', ci[2], ')'))
```

```
[1] "90% Confidence interval for the Intercept: (28.16,31.28)"
```

Plot Update

```
plot(data$diastolic, data$systolic, cex=1, pch=20, col='grey',
     main='Systolic vs. Diastolic',
     xlab = 'Diastolic', ylab = 'Systolic')
abline(linreg, col='red', lwd=2)
```

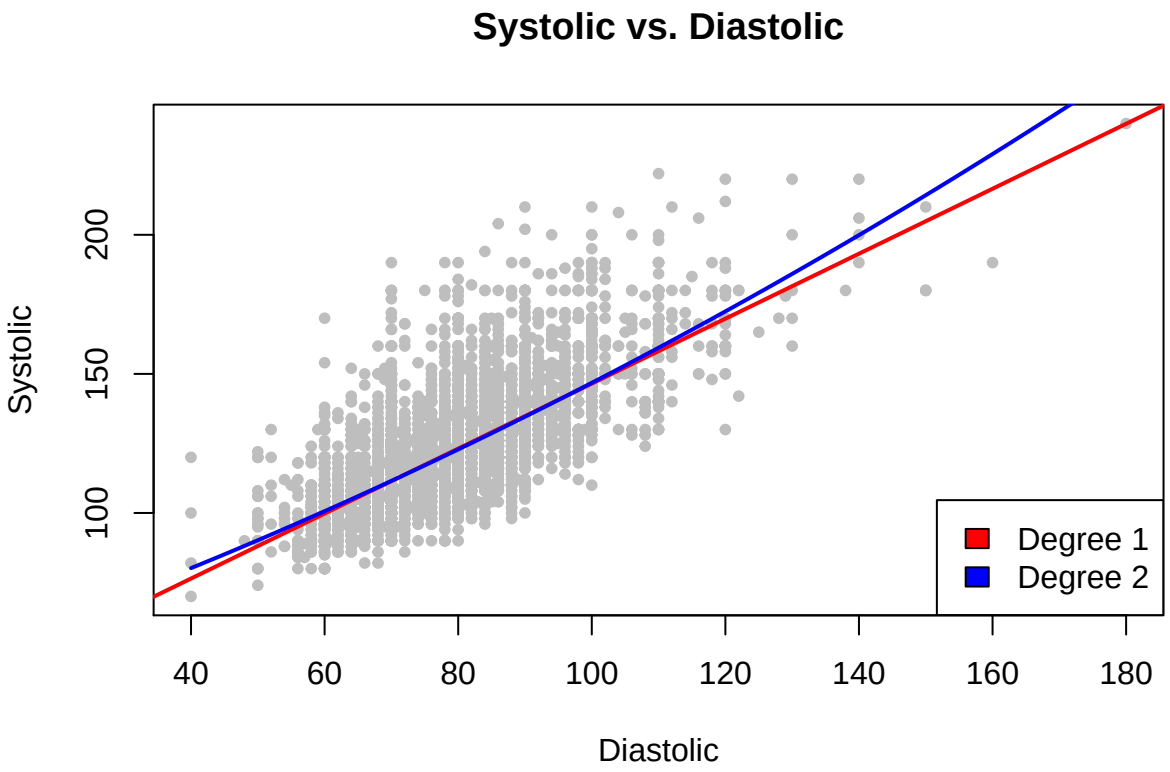


Part D: Polynomial Regression Model

```
# Polynomial Regression of 2nd power
polyreg2 <- lm(systolic ~ poly(diastolic, 2, raw = TRUE), data = data)

# Update Data: x and y values
grid <- seq(40, 180, length=1000)
# Predicted values based on Polynomial Regression Model
values <- predict(polyreg2, data.frame(diastolic = grid))

# Base plot
plot(data$diastolic, data$systolic, cex=1, pch=20, col='grey',
     main='Systolic vs. Diastolic',
     xlab = 'Diastolic', ylab = 'Systolic')
# Linear Model Regression line
abline(linreg, col='red', lwd=2)
# Polynomial of 2nd degree Model Regression line
lines( grid, values, col='blue', lwd=2)
legend('bottomright', c('Degree 1', 'Degree 2'), fill = c('red', 'blue'))
```



Adjusted R-squared

```
linregR <- summary(linreg)$adj.r.squared
polyreg2R <- summary(polyreg2)$adj.r.squared
# Summary table of Adjusted R-squared
data.frame(cbind('Linear Regression' = linregR,
                 'Polynomial Regression' = polyreg2R))
```

```
  Linear.Regression Polynomial.Regression
1         0.5123777             0.5132843
```

From the summary above, adding a polynomial term of 2^{nd} degree does not increase the Adjusted R-squared value significantly. Hence, for better interpretability, it is preferred to have a model as simple as possible, it would be recommended to use the Simple Linear Regression model.