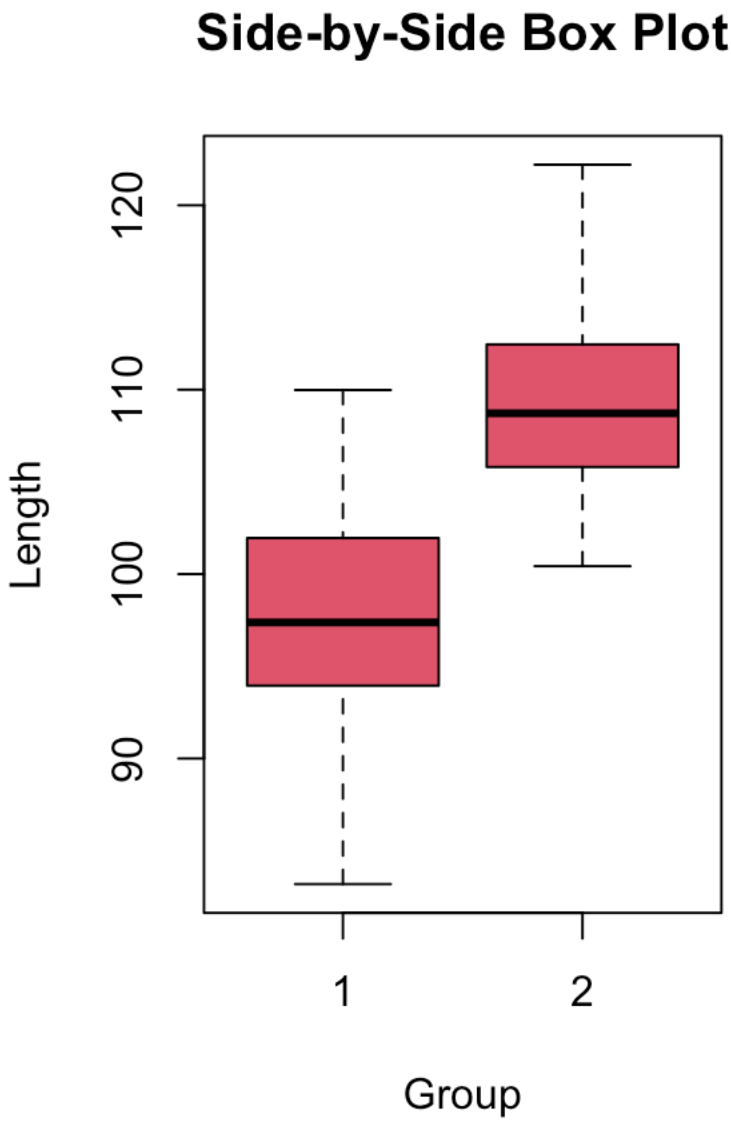
Homework 4 **Name:** Ivan Wang
(Parts 7-8; 80 pts)

1. The data set apnea.txt (UBLearns) contains data on a facial measurement thought to be related to a disruptive sleep condition. Thirty-one subjects with sleep apnea (group 1) were measured, along with thirty-nine subjects without sleep apnea (group 2). (18 pts)
   a) Produce side-by-side boxplots to compare the two samples. Comment on any visual differences.




```
boxplot(length ~ group, data = apnea,
        col = (length(unique(apnea$group))),
        main = "Side-by-Side Box Plot",
        xlab = "Group", ylab = "Length")
#Most of Group 1 data seems be around 100, while most of Group 2 data is around 110.
```

b) Investigators wish to test whether this facial measurement differs on average between healthy and apnea patients. Write the appropriate null and alternative hypotheses.

H0: $\mu_1 = \mu_2$(Facial measurement is same between healthy and apnea patients)

Ha: $\mu_1 \neq \mu_2$(Facial measurement is different between healthy and apnea patients)

c) Carry out the test ($\alpha = 0.05$) assuming that the groups have equal variance. Give the test statistic, p-value, and conclusion **in context**.

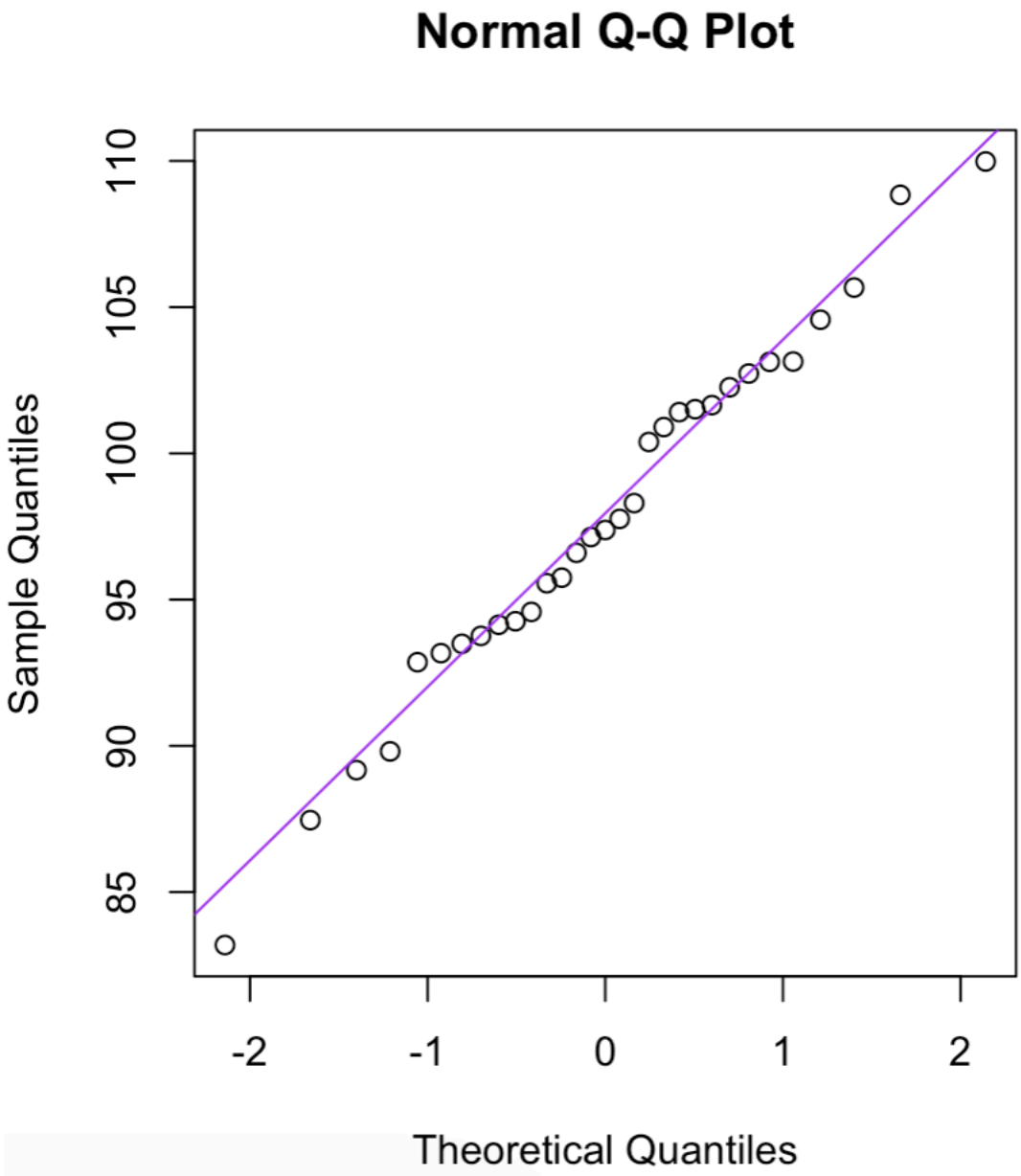
```
> t.test(length ~ group, data = apnea, var.equal = TRUE)

        Two Sample t-test

data:  length by group
t = -8.5514, df = 68, p-value = 2.187e-12
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
 -14.432086  -8.970991
sample estimates:
mean in group 1 mean in group 2
        97.7600        109.4615

> #t = -8.5514, df = 68, p-value = 2.187e-12
> #With P-value = 2.187e-12 < 0.05, we reject H0. There is significant evidence that facial measurement is different between healthy and apnea patients
>
```

d) Check for equal variance using Bartlett's test - see the help file for bartlett.test(.). Give the hypotheses, p-value, and conclusion.

```
> #H0: var1 = var2(there's equal variance between the healthy and apnea group)
> #Ha: Var1 != var2(there's unequal variance between the two groups)
> bartlett.test(length ~ group, data = apnea)

        Bartlett test of homogeneity of variances

data:  length by group
Bartlett's K-squared = 0.65871, df = 1, p-value = 0.417

>
> # Bartlett's K-squared = 0.65871, df = 1, p-value = 0.417
> #With p-value = 0.417 > 0.05, we fail to reject H0. There is significant evidience that there's equal variance between the healthy and apnea group.
>
```

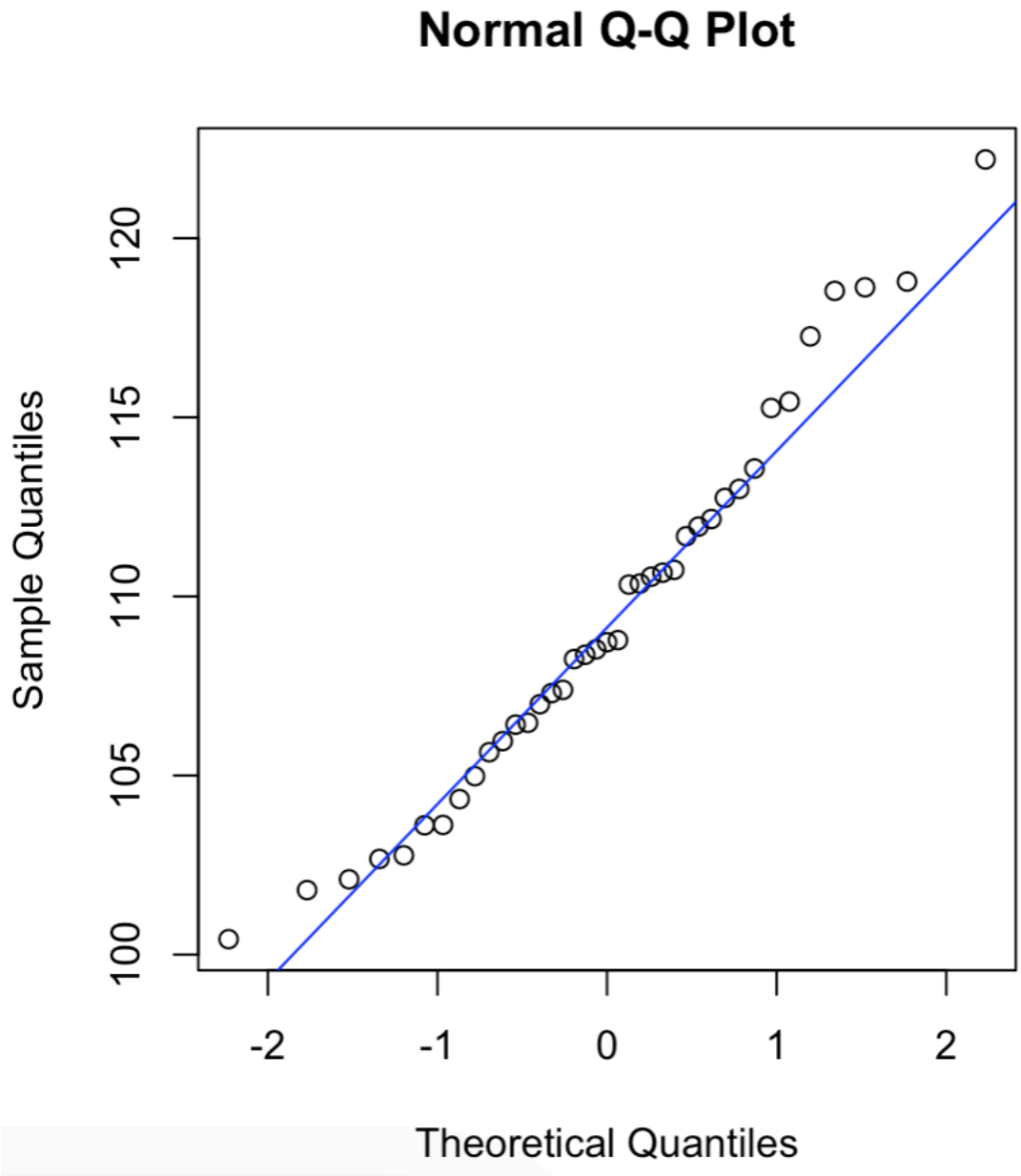
e) Test for normality in both groups.

For Apnea:



The points closely follow the 45 degree reference line, which could indicate normality. Lets check Shapiro-Wilk Test

For healthy:



The points closely follow the 45 degree reference line aside from a couple of outlier points, which could indicate normality. Lets check Shapiro-Wilk Test

```
> shapiro.test(apnea_pateients)

        Shapiro-Wilk normality test

data:  apnea_pateients
W = 0.98584, p-value = 0.9456

> shapiro.test(healthy_patients)

        Shapiro-Wilk normality test

data:  healthy_patients
W = 0.97145, p-value = 0.4157
```

With p-values both >0.05, we fail to reject H0. There is evidence that the data is normal.

f) Give a 95% confidence interval for the difference between the two population means.

95 CI: (-14.432086 , -8.970991)

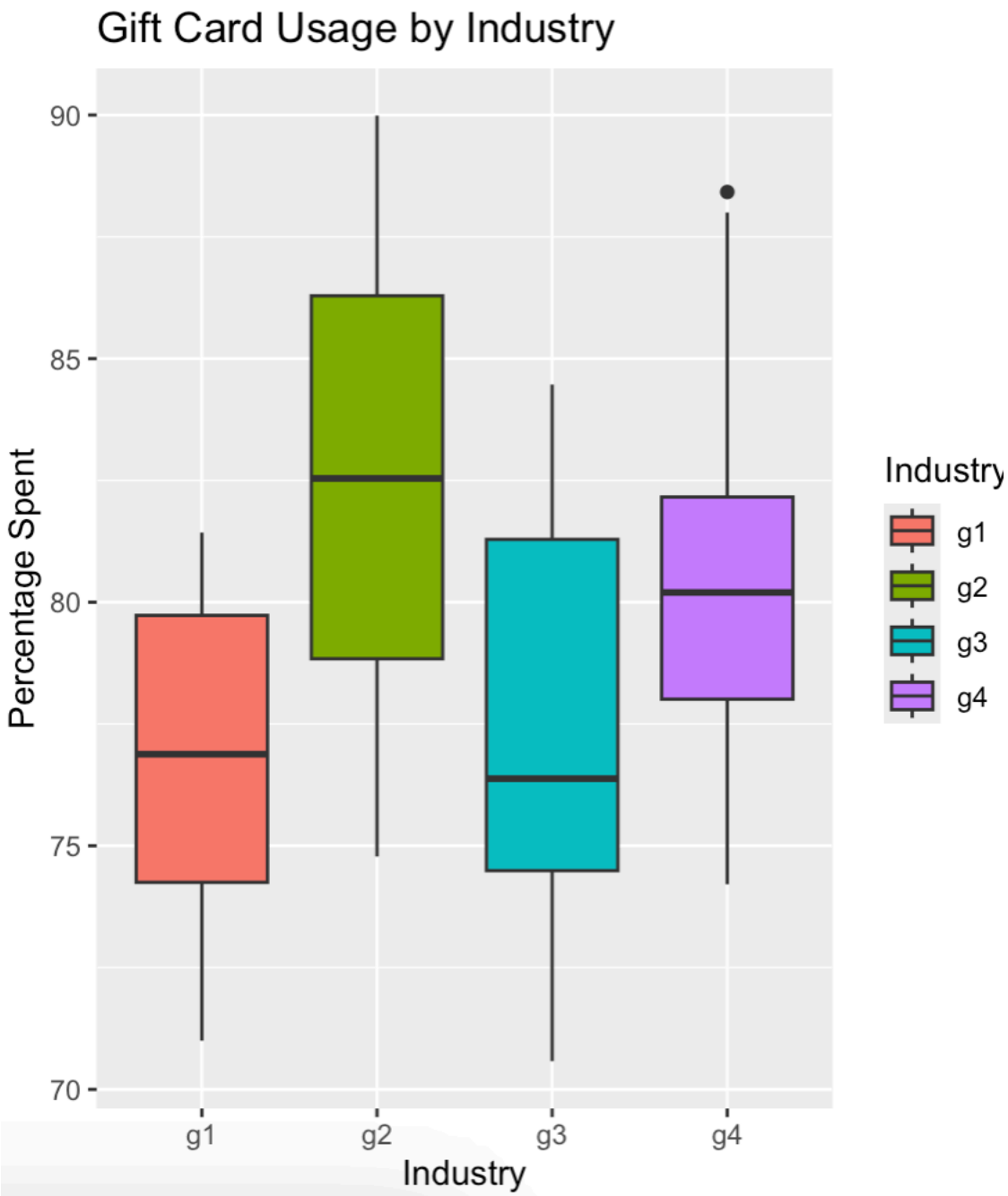
```
> t.test(length ~ group, data = apnea, var.equal = TRUE)

        Two Sample t-test

data:  length by group
t = -8.5514, df = 68, p-value = 2.187e-12
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
 -14.432086  -8.970991
```

2. A rookie statistician views the output from Bartlett's test and believes the group variances are different. Re-analyze the sleep apnea data using a nonparametric alternative to the two-sample t-test. Give the hypotheses, p-value, and conclusion in context. (4 pts)

```
        Wilcoxon rank sum test with continuity correction

data:  length by group
W = 81, p-value = 6.261e-10
alternative hypothesis: true location shift is not equal to 0

> #H0: The distributions of facial measurements for apnea and healthy groups are the same.
> #Ha: The distribution is not the same.
> #W = 81, p-value = 6.261e-10
>
> #with p-value = 6.261e-10 < 0.05, we reject H0. There is significant difference in facial measurements for apnea and healthy groups.
>
```

3. Businesses love to sell gift cards because many of the cards sold end up being lost, discarded, or used incompletely. The data set gift_card.txt (UBLearns) contains results from a year-long study of gift card usage in four different industries (labeled g1-g4). Each observation is the percentage of total gift card dollars spent at a business within a year of the gift card's initial sale. Take care that the data are read in correctly during the initial import phase. (16 pts)

a) Summarize the data using side-by-side boxplots. Also give a table containing group-specific sample means and standard deviations. This may require you to reorganize the data into a stacked form; one option is to explore the stack(.) function.



	Industry	Mean	SD
g1	g1	76.36294	3.268675
g2	g2	82.34059	4.366062
g3	g3	77.12941	4.165119
g4	g4	80.41294	3.848002

b) Fit an ANOVA model to compare the group means. Give the hypotheses, test statistic, p-value, and conclusion in context.

```
> summary(aov(Percentage_Spent ~ Industry, data = data_stacked))
            Df Sum Sq Mean Sq F value   Pr(>F)
Industry     3  401.1  133.70   8.639 6.76e-05 ***
Residuals   64  990.4   15.48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> #H0: u1 = u2 = u3 = u4(the means of all the industries are the same)
> #Ha: not all the same
>
> #F-value = 8.639, P-Value = 6.76e-05
> #With P-value = 6.76e-05 < 0.05, we reject H0. There is significant evidence that atleast one of the indutries means is different.
```

c) If the previous null hypothesis was rejected, test for pairwise differences between group means using a Bonferroni correction. Summarize which pairs are found to be statistically different.

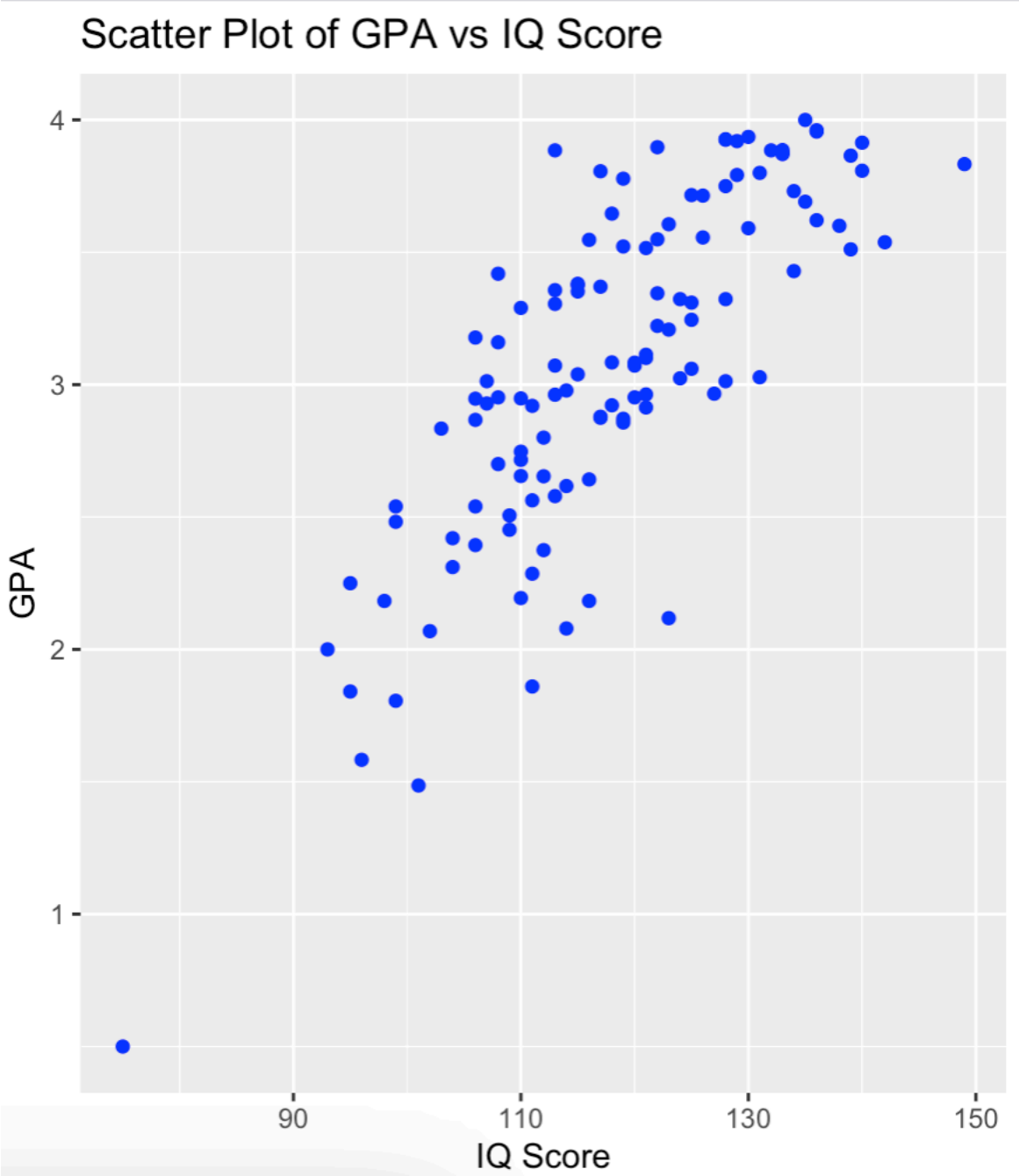
```
> pairwise.t.test(data_stacked$Percentage_Spent, data_stacked$Industry, p.adjust.method = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  data_stacked$Percentage_Spent and data_stacked$Industry

   g1      g2      g3
g2 0.00023 -       -
g3 1.00000 0.00159 -
g4 0.02296 0.94787 0.10654

P value adjustment method: bonferroni
>
> #Pairs (g1 and g2), (g1 and g4), (g2 and g3) are statistically different because their p-value < 0.05.
```

d) Apply Levene’s test. Give the hypotheses, p-value, and conclusion.

```
> leveneTest(Percentage_Spent ~ Industry, data = data_stacked)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.5063 0.6793
      64
> #H0: var of the industries is equal (homogeneous).
> #Ha: atleast one of the industry variance is different(heterogeneous).
>
> #F-value - 0.5063, P-Value = 0.6793
> #With P-Value = 0.6793 > 0.05, we fail to reject Ho. There is significant evidence that the variance across all of the industries is equal and there's homogeneous.
```

4. The data set gpa.txt contains data on 120 college students who had just completed their first year. We wish to investigate whether performance measures collected while these students were still in high school can be used to predict their first-year GPA (grade point average). (20 pts)

a) One of the measures recorded while the students were in high school was an IQ test. Produce a scatter plot involving the variables "gpa" and "IQ_score." Be sure that you can properly identify which of these should be treated as the predictor variable and which should be treated as the response variable.



b) Fit the simple linear regression model. Provide the table of regression estimates.

```
> model <- lm(gpa ~ IQ_score, data = data)
> summary(model)

Call:
lm(formula = gpa ~ IQ_score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1672 -0.2402 -0.0225  0.2977  1.0193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.873921   0.345709  -5.421  3.2e-07 ***
IQ_score     0.041944   0.002915  14.389  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3899 on 118 degrees of freedom
Multiple R-squared:  0.637,     Adjusted R-squared:  0.6339
F-statistic:   207 on 1 and 118 DF,  p-value: < 2.2e-16
```

c) A previous study suggests that a 1-unit increase in IQ score is associated with a 0.05-unit increase in college GPA. Test whether the regression slope is different than 0.05.

looking back at model summary.
F-statistic: 207 on 1 and 118 DF, p-value: < 2.2e-16
With p-value: < 2.2e-16 < 0.05, we reject Ho. There is significant evidence that the regression slope is different than 0.05

d) Interpret the R^2 statistic, and show how it is calculated (i.e. use a formula).

```
> anova(model)
Analysis of Variance Table

Response: gpa
           Df Sum Sq Mean Sq F value    Pr(>F)
IQ_score    1 31.470  31.470  207.04 < 2.2e-16 ***
Residuals 118 17.936   0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Multiple R-squared: 0.637, Adjusted R-squared: 0.6339
SSTO = SSE + SSR = 1

$R^2 = \frac{SSR}{SSTO}$ = 31.470 / (31.470 + 17.936) = 0.637

It represents the proportion of variance in dependent variable that is explained by the independent variables in a regression model or how well the line fits the model.

e) Use the fitted model to estimate the first-year GPA for a student whose high school IQ score was 120. Also provide a 95% confidence interval for the mean first-year GPA when IQ score is 120.

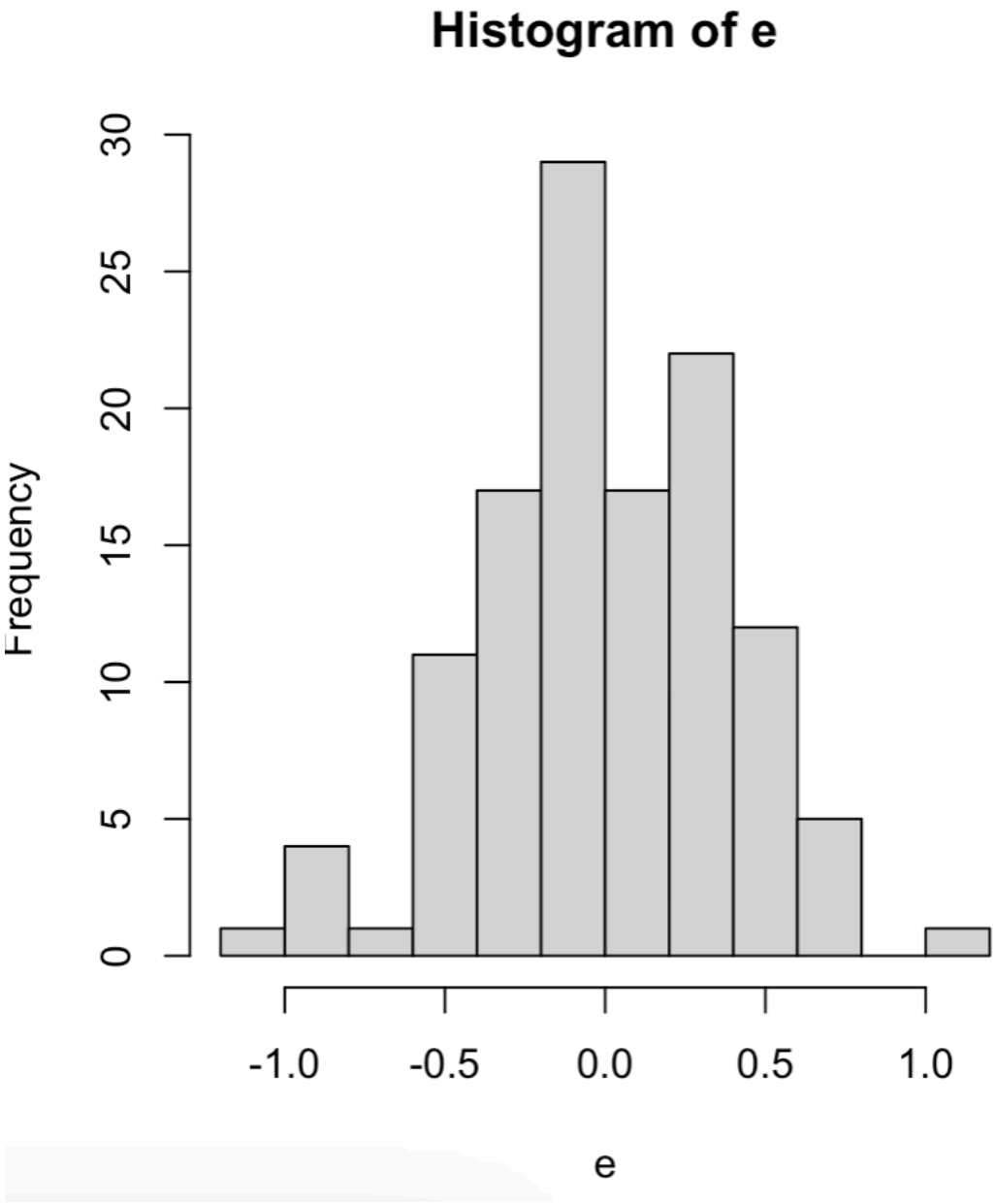
```
> predicted <- predict(model, newdata = smart)
> predicted
       1
3.159336
>
>
> predict(model, smart, interval="confidence", level=0.95)
       fit      lwr      upr
1 3.159336 3.087887 3.230784
>
```

f) Use the model residuals to assess the assumptions of normality and constant variance.

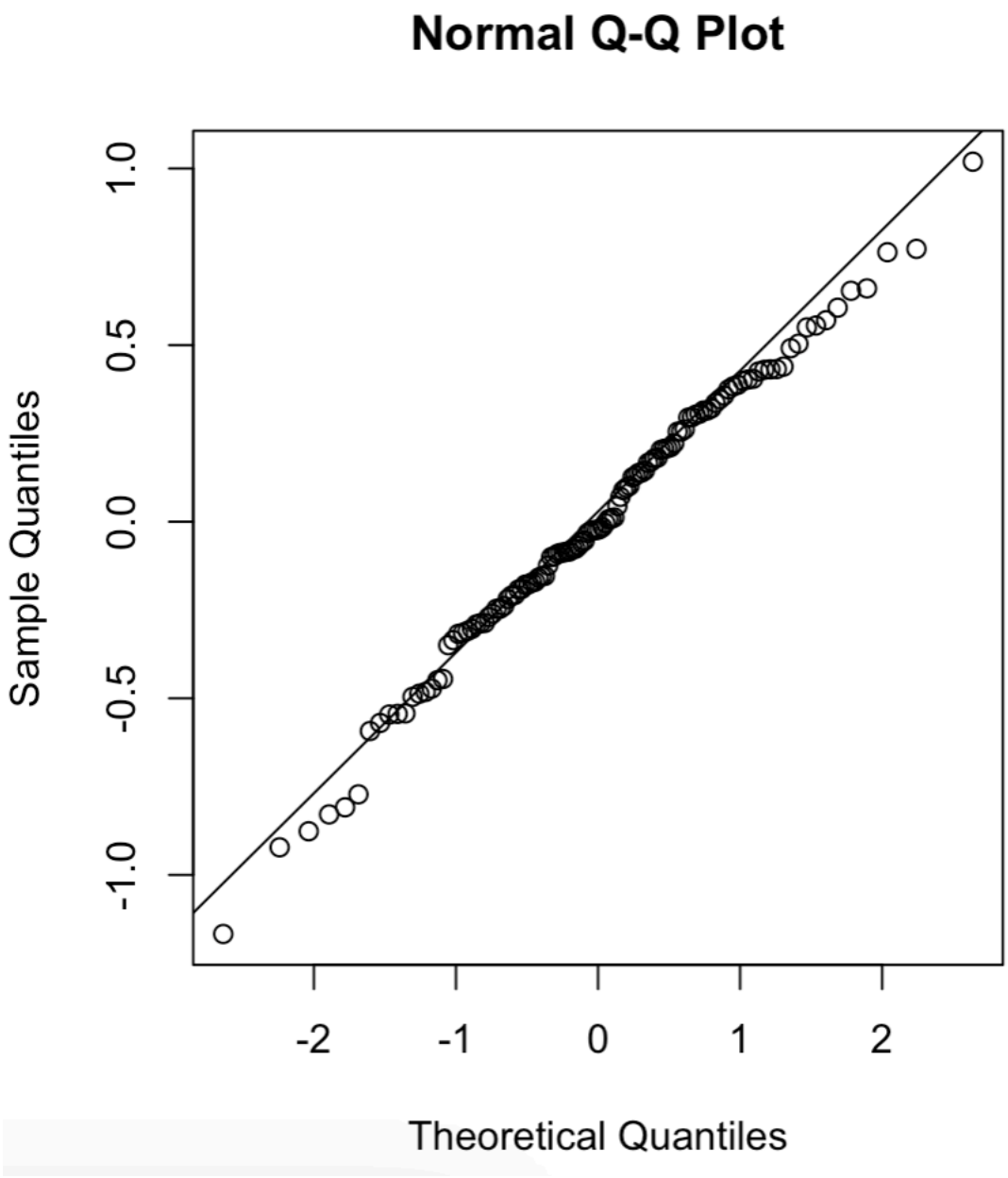
Shapiro-Wilk normality test

data: e
W = 0.99132, p-value = 0.6566

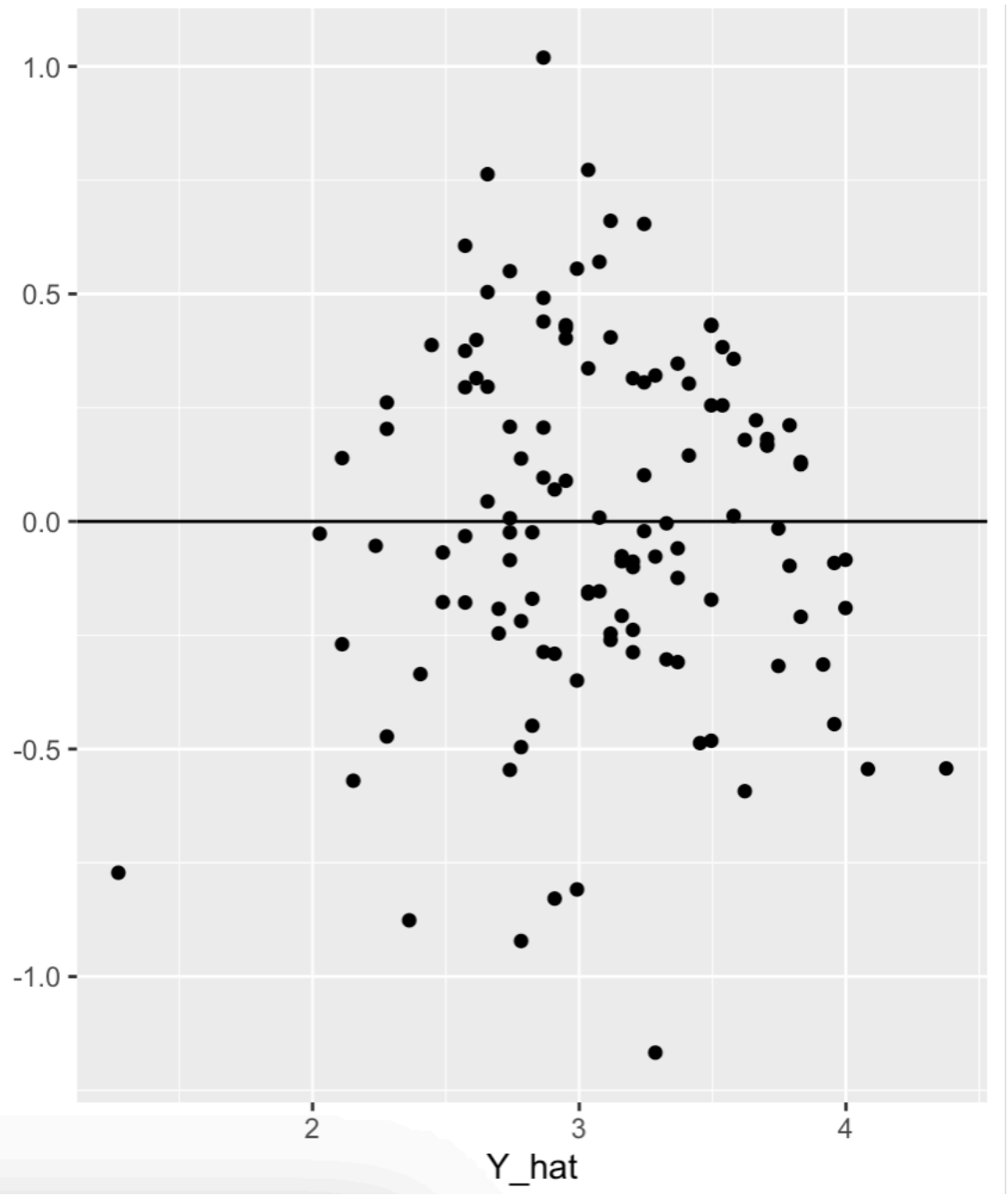
With p-value = 0.6566 > 0.05, we fail to reject Ho. There is significant evidence that the residuals have a normal distribution.



looks bell shaped, which could suggest normality.



most data points are along the 45-degree reference line, which suggest normality



There is linearity because the points seem to be randomly scattered. There may be non-constant variance because Y-hat values 2-4 seem to contain the most points.

5. We will again use the GPA data. Two additional predictors (ACT score and high school class rank) are available. (22 pts)

a) Use the cor(.) function to obtain sample correlations between all pairs of variables in the data set. Which variable is most strongly correlated with first-year GPA?

```
> cor(data)
                gpa act_score  IQ_score class_rank
gpa        1.0000000 0.2694818 0.7981043  0.3702315
act_score  0.2694818 1.0000000 0.3092707  0.4581364
IQ_score   0.7981043 0.3092707 1.0000000  0.3673556
class_rank 0.3702315 0.4581364 0.3673556  1.0000000
```

IQ_score is most strongly correlated with first-year GPA.

b) Fit a multiple regression model using all three available predictors. This will be known as the full model. Present a table containing parameter estimates.

```
Call:
lm(formula = gpa ~ act_score + IQ_score + class_rank, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03062 -0.24020 -0.01855  0.28195  1.04770

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.837546   0.362397  -5.071 1.52e-06 ***
act_score   -0.001509   0.009114  -0.166    0.869
IQ_score     0.040314   0.003177  12.689  < 2e-16 ***
class_rank   0.002623   0.001820   1.441    0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3894 on 116 degrees of freedom
Multiple R-squared:  0.6439,    Adjusted R-squared:  0.6347
F-statistic: 69.92 on 3 and 116 DF,  p-value: < 2.2e-16
```

c) Which of the predictor variables are found to have significant association with first-year GPA? Provide p-values when you identify the significant predictors.

gpa is the significant predictor because its p-value <2e-16 < 0.05, therefore the predictor is significant.

d) Use the fitted model to estimate the first-year GPA for a student with ACT score 20, IQ score 100, and class rank 65.

```
> predict(model, smart, interval="confidence", level=0.95)
       fit      lwr      upr
1 3.159336 3.087887 3.230784
> student <- data.frame(act_score = 20, IQ_score = 100, class_rank = 65)
> predicted_gpa <- predict(model2, newdata = student)
> predicted_gpa
     1
2.3342
```

e) The current model has $g = 4$ regression parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$, and the confint(.) function produces four 95% confidence intervals. The probability that all four intervals contain their target parameters is $0.95^4 = 0.8145$. A Bonferroni correction can be applied to a set of g confidence intervals so that their joint probability is at least $(1 - \alpha)$ by computing each interval at level $\left(1 - \frac{\alpha}{g}\right)100\%$. Provide a set of Bonferroni-adjusted confidence intervals that maintains joint coverage probability 95%.

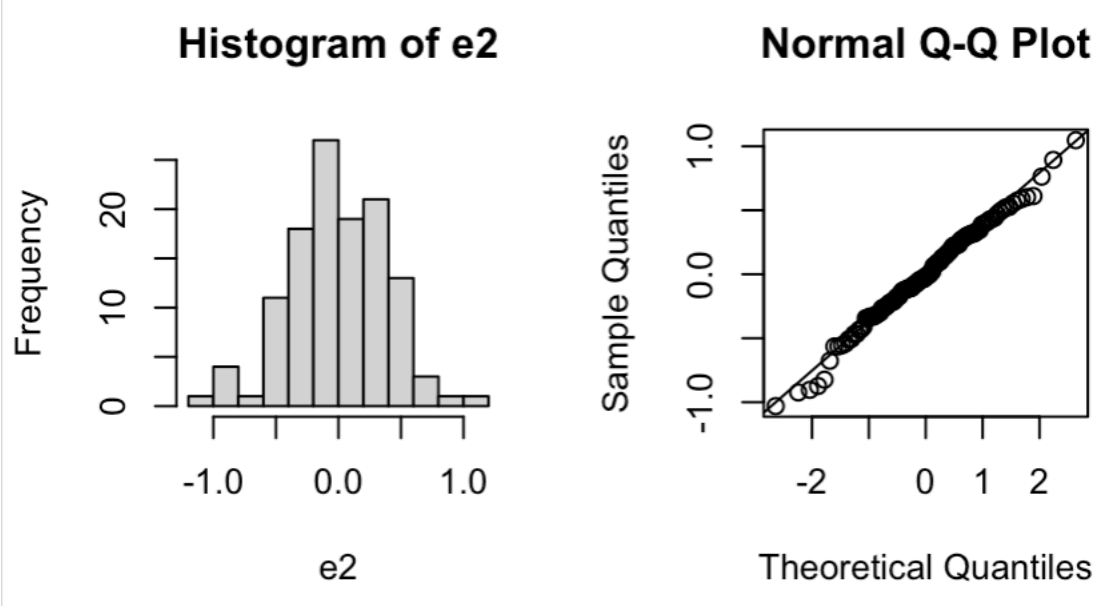
(1 – 0.05 / 4)100% = 98.75%

```
> confint(model2, level = 0.9875)
                 0.625 %     99.375 %
(Intercept) -2.757037887 -0.918054650
act_score   -0.024633541  0.021616253
IQ_score     0.032252724  0.048375346
class_rank  -0.001994164  0.007240791
>
```

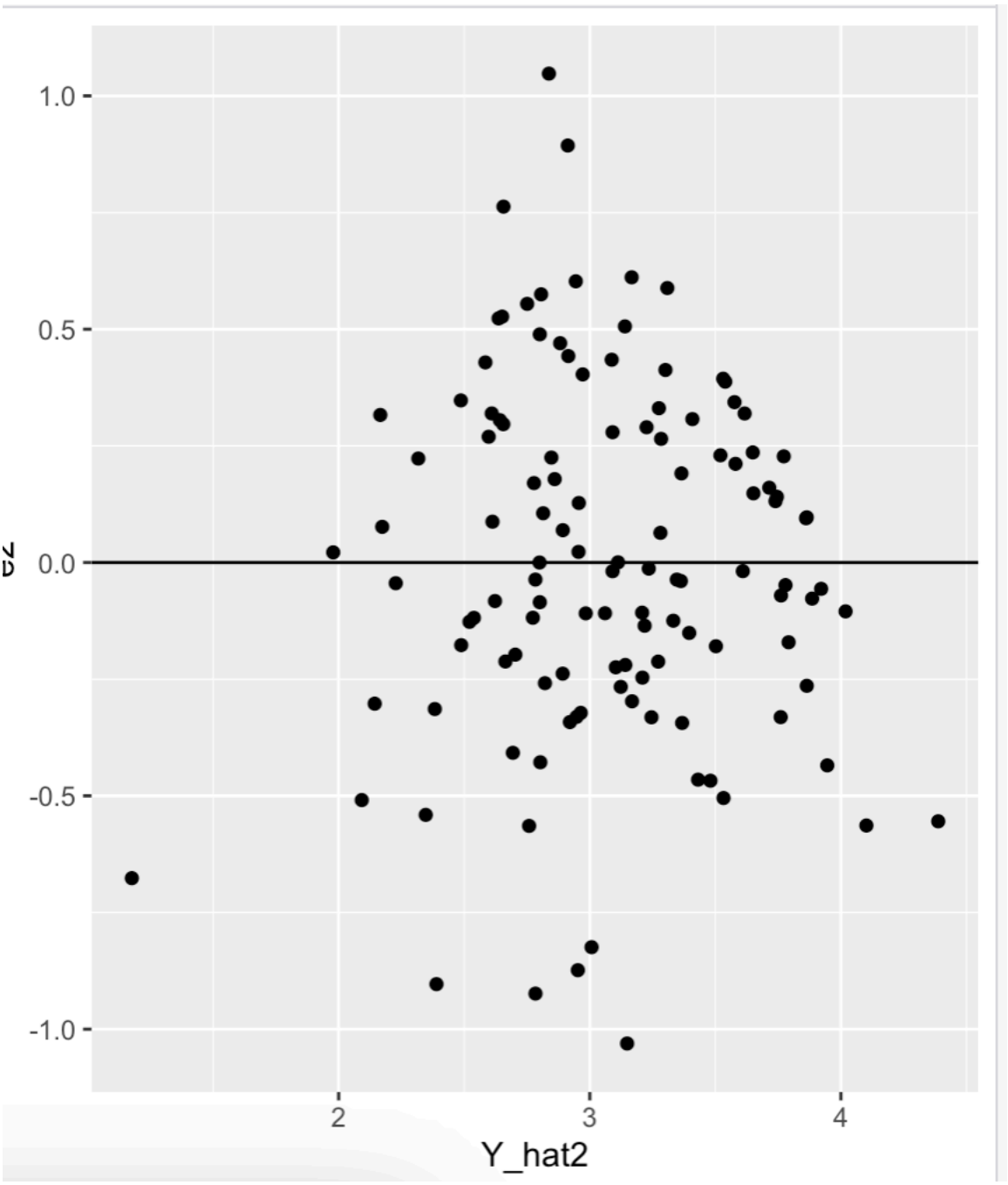
f) Analyze the residuals from the full model, and assess the same assumptions as in the previous problem.

Shapiro-Wilk normality test

data: e2
W = 0.99257, p-value = 0.7741
With p-value = 0.7741 > 0.05, we fail to reject Ho. There is significant evidence that the residuals have a normal distribution.



Looks bell shaped, which could suggest normality. Most data points are along the 45 degree reference line, which suggest normality.



There is linearity because the points seem to be randomly scattered. There may be non-constant variance because Y-hat values 2-4 seem to contain the most points.

g) Report the R_a^2 statistic.

```
> summary(model2)$adj.r.squared
[1] 0.6347072
```

h) Test whether the full model is any better than the reduced model fit in problem 4.

```
> anova(model, model2)
Analysis of Variance Table

Model 1: gpa ~ IQ_score
Model 2: gpa ~ act_score + IQ_score + class_rank
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    118 17.936
2    116 17.593  2   0.34316 1.1313 0.3261
>
```

Since p-value = 0.3261 > 0.05, this means that there is no significant evidence that adding ACT scores and class rank improves the model.

Code for problems 1-2

```
#Homework 4

"""
1.	The data set apnea.txt (UBLearns) contains data on a facial measurement thought to
be related to a disruptive sleep condition. Thirty-one subjects with sleep apnea
(group 1) were measured, along with thirty-nine subjects without sleep apnea (group 2). (18 pts)
"""

#a)	Produce side-by-side boxplots to compare the two samples.  Comment on any visual differences.

apnea <- read.csv("~/Downloads/apnea.txt", sep="")
View(apnea)

boxplot(length ~ group, data = apnea,
        col = (length(unique(apnea$group))),
        main = "Side-by-Side Box Plot",
        xlab = "Group", ylab = "Length")
#Most of Group 1 data seems be around 100, while most of Group 2 data is around 110.

#b)	Investigators wish to test whether this facial measurement differs on average between healthy and apnea patients. Write the
appropriate null and alternative hypotheses.

#H0:  μ_1= μ_2(Facial measurement is same between healthy and apnea patients)
#Ha: μ_1≠ μ_2(Facial measurement is different between healthy and apnea patients)
```

```
#c)  Carry out the test (α=0.05) assuming that the groups have equal variance.
#Give the test statistic, p-value, and conclusion in context.

t.test(length ~ group, data = apnea, var.equal = TRUE)
#t = -8.5514, df = 68, p-value = 2.187e-12
#With P-value = 2.187e-12 < 0.05, we reject H0. There is significant evidence that facial measurement is different between healthy and
apnea patients


#d)  Check for equal variance using Bartlett's test - see the help file for bartlett.test(.). Give the hypotheses, p-value, and conclusion.

#H0: var1 = var2(there's equal variance between the healthy and apnea group)
#Ha: Var1 != var2(there's unequal variance between the two groups)
bartlett.test(length ~ group, data = apnea)

# Bartlett's K-squared = 0.65871, df = 1, p-value = 0.417
#With p-value = 0.417 > 0.05, we fail to reject H0. There is significant evidience that there's equal variance between the healthy and apnea
group.


#e)  Test for normality in both groups.
apnea_pateients <- subset(apnea, group == 1)$length
healthy_patients <- subset(apnea, group == 2)$length
apnea_pateients
healthy_patients

#qqplot
qqnorm(apnea_pateients)
qqline(apnea_pateients, col = "purple")
#The points closely follow the 45 degree reference line, which could indicate normality. Lets check Shapiro-Wilk Test

qqnorm(healthy_patients)
qqline(healthy_patients, col = "blue")
#The points closely follow the 45 degree reference line aside from a couple of outlier points, which could indicate normality. Lets check
Shapiro-Wilk Test

shapiro.test(apnea_pateients)
shapiro.test(healthy_patients)

"""
data:  apnea_pateients
W = 0.98584, p-value = 0.9456

data:  healthy_patients
W = 0.97145, p-value = 0.4157

with p-values both >0.05, we fail to reject H0. There is evidence that the data is normal.

"""

#f)  Give a 95% confidence interval for the difference between the two population means.
t.test(length ~ group, data = apnea, var.equal = TRUE)



"""
2.  A rookie statistician views the output from Bartlett's test and believes the group variances are different.
Re-analyze the sleep apnea data using a nonparametric alternative to the two-sample t-test. Give the hypotheses,
p-value, and conclusion in context. (4 pts)
"""

wilcox.test(length ~ group, data = apnea, exact = FALSE)
#H0: The distributions of facial measurements for apnea and healthy groups are the same.
#Ha: The distribution is not the same.
#W = 81, p-value = 6.261e-10

#with p-value = 6.261e-10 < 0.05, we reject H0. There is significant difference in facial measurements for apnea and healthy groups.
```

Code for problem 3:

```
"""
3. Businesses love to sell gift cards because many of the cards sold end up being lost, discarded, or used incompletely.
The data set gift_card.txt (UBLearns) contains results from a year-long study of gift card usage in four different
industries (labeled g1-g4). Each observation is the percentage of total gift card dollars spent at a business within a
year of the gift card's initial sale. Take care that the data are read in correctly during the initial import phase. (16 pts)
"""
library(ggplot2)

data <- read.table("~/Downloads/giftcard.txt", header = FALSE, skip = 1, sep = "", stringsAsFactors = FALSE)
View(data)



colnames(data) <- c("g1", "g2", "g3", "g4")
data_stacked <- stack(data[, -1])

#a)    Summarize the data using side-by-side boxplots. Also give a table containing group-specific sample means and standard deviations.
#This may require you to reorganize the data into a stacked form; one option is to explore the stack(.) function.

data <- data.frame(lapply(data, as.numeric))
View(data)

data_stacked <- stack(data)

colnames(data_stacked) <- c("Percentage_Spent", "Industry")

ggplot(data_stacked, aes(x = Industry, y = Percentage_Spent, fill = Industry)) +
  geom_boxplot() +
  labs(title = "Gift Card Usage by Industry", x = "Industry", y = "Percentage Spent")



means <- tapply(data_stacked$Percentage_Spent, data_stacked$Industry, mean)
sds <- tapply(data_stacked$Percentage_Spent, data_stacked$Industry, sd)
summary_table <- data.frame(Industry = names(means), Mean = means, SD = sds)
summary_table


#b)    Fit an ANOVA model to compare the group means. Give the hypotheses, test statistic, p-value, and conclusion in context.
aov(Percentage_Spent ~ Industry, data = data_stacked)
summary(aov(Percentage_Spent ~ Industry, data = data_stacked))


#H0: u1 = u2 = u3 = u4(the means of all the industries are the same)
#Ha: not all the same

#F-value = 8.639, P-Value = 6.76e-05
#With P-value = 6.76e-05 < 0.05, we reject H0. There is significant evidence that atleast one of the indutries means is different.


#c)    If the previous null hypothesis was rejected, test for pairwise differences between group means using a Bonferroni correction.
#Summarize which pairs are found to be statistically different.

pairwise.t.test(data_stacked$Percentage_Spent, data_stacked$Industry, p.adjust.method = "bonferroni")

#Pairs (g1 and g2), (g1 and g4), (g2 and g3) are statistically different because their p-value < 0.05.


#d)    Apply Levene's test. Give the hypotheses, p-value, and conclusion.
library(car)

leveneTest(Percentage_Spent ~ Industry, data = data_stacked)
#H0: var of the industries is equal (homogeneous).
#Ha: atleast one of the industry variance is different(heterogeneous).
```

```
#F-value - 0.5063, P-Value = 0.6793
#With P-Value = 0.6793 > 0.05, we fail to reject Ho. There is significant evidence that the variance across all of the industries is equal and there's homogeneous.
```

Code for 4-5:

```
"""
4. The data set gpa.txt contains data on 120 college students who had just completed their first year.
We wish to investigate whether performance measures collected while these students were still in high school
can be used to predict their first-year GPA (grade point average). (20 pts)
"""
library(ggplot2)
"""
a)	One of the measures recorded while the students were in high school was an IQ test.
Produce a scatter plot involving the variables “gpa” and “IQ_score.”
Be sure that you can properly identify which of these should be treated as the predictor variable and which should be treated as the response variable.
"""
data <- read.csv("~/Downloads/gpa.txt", sep="")
View(data)

ggplot(data, aes(x = IQ_score, y = gpa)) +
  geom_point(color = "blue") +  # Scatter plot points
  labs(title = "Scatter Plot of GPA vs IQ Score",
       x = "IQ Score",
       y = "GPA")


#b)	Fit the simple linear regression model. Provide the table of regression estimates.
model <- lm(gpa ~ IQ_score, data = data)
summary(model)

#c)	A previous study suggests that a 1-unit increase in IQ score is associated with a 0.05-unit increase in college GPA.
#Test whether the regression slope is different than 0.05.

"""
looking back at model summary.

F-statistic:   207 on 1 and 118 DF,  p-value: < 2.2e-16

With p-value: < 2.2e-16 < 0.05, we reject Ho. There is significant evidence that the regression slope is different than 0.05.
"""


#d)	Interpret the R^2 statistic, and show how it is calculated (i.e. use a formula).
anova(model)
"""
Multiple R-squared:  0.637, Adjusted R-squared:  0.6339
SSE = 0.3899, Therefore SSR = 1 - 0.3899 = 0.6101
SSTO = SSE + SSR = 1

R^2=SSR/SSTO  = 0.6101/1

31.470 / 31.470 + 17.936
"""

#e)	Use the fitted model to estimate the first-year GPA for a student whose high school IQ score was 120.
#Also provide a 95% confidence interval for the mean first-year GPA when IQ score is 120.
smart <- data.frame(IQ_score = 120)

predicted <- predict(model, newdata = smart)
predicted
```

```
predict(model, smart, interval="confidence", level=0.95)


#f)    Use the model residuals to assess the assumptions of normality and constant variance.
e <- model$residuals
e

shapiro.test(e)
#W = 0.99132, p-value = 0.6566
#With p-value = 0.6566 > 0.05, we fail to reject Ho. There is significant evidence that the residuals have a normal distribution.
hist(e)
#looks bell shaped, which could suggest normality.
qqnorm(e)
qqline(e)
#most data points are along the 45 degree reference line, which suggest normality

Y_hat <-  model$fitted.values
ggplot(model, aes(x=Y_hat, y=e)) +
  geom_point() +
  geom_hline(yintercept=0)

#There is linearity because the points seem to be randomly scattered. There may be non-constant variance because Y-hat values 2-4 seem to
contain the most points.

"""
5. We will again use the GPA data. Two additional predictors (ACT score and high school class rank) are available. (22 pts)
"""

#a)    Use the cor(.) function to obtain sample correlations between all pairs of variables in the data set.
#Which variable is most strongly correlated with first-year GPA?

cor(data)
#IQ_score is most strongly correlated with first-year GPA.


#b)    Fit a multiple regression model using all three available predictors.
#This will be known as the full model. Present a table containing parameter estimates.

model2 <- lm(gpa ~ act_score + IQ_score + class_rank, data=data)
summary(model2)

#c)    Which of the predictor variables are found to have significant association with first-year GPA?
#Provide p-values when you identify the significant predictors.

#IQ_score is the significant predictor because its p-value <2e-16 < 0.05, therefore this predictor is significant.


#d)    Use the fitted model to estimate the first-year GPA for a student with ACT score 20, IQ score 100, and class rank 65.
student <- data.frame(act_score = 20, IQ_score = 100, class_rank = 65)
predicted_gpa <- predict(model2, newdata = student)
predicted_gpa


"""
e)      The current model has g=4 regression parameters (β_0,β_1,β_2,β_3), and the confint(.) function produces four 95% confidence
intervals.
The probability that all four intervals contain their target parameters is 〖0.95〗^4=0.8145.
A Bonferroni correction can be applied to a set of g confidence intervals so that their joint probability is at least (1-α)
by computing each interval at level (1-α/g)100%. Provide a set of Bonferroni-adjusted confidence intervals that maintains joint coverage
probability 95%.
"""

#(1 – 0.05 / 4)100% = 98.75%

confint(model2, level = 0.9875)


#f)    Analyze the residuals from the full model, and assess the same assumptions as in the previous problem.
e2 <- model2$residuals
```

```
e2

shapiro.test(e2)
#W= 0.99257, p-value = 0.7741
#With p-value = 0.7741 > 0.05, we fail to reject Ho. There is significant evidence that the residuals have a normal distribution.
hist(e2)
#looks bell shaped, which could suggest normality.
qqnorm(e2)
qqline(e2)
#most data points are along the 45 degree reference line, which suggest normality

Y_hat2 <-  model2$fitted.values
ggplot(model2, aes(x=Y_hat2, y=e2)) +
  geom_point() +
  geom_hline(yintercept=0)


#g)    Report the R_a^2 statistic.
summary(model2)$adj.r.squared


#h)    Test whether the full model is any better than the reduced model fit in problem 4.
anova(model, model2)


#Since p-value = 0.3261 > 0.05, this means that there is no significant evidence that adding ACT scores and class rank improves the model
```