title: ST303 - Assignment 2

| Last name: Wu

| First name: Lunwei

| Student ID: 1001303957

| Course section: STA303H-winter 2017

data1 = read.csv("/Users/jesse/Desktop/bbw.csv")

attach(data1)

maturity=array(0,length(gestation))

MatSmoke=array(0,length(smoke))

for (i in 1:length(gestation))

{

if (gestation[i]<259)

{maturity[i]=1}

else if (gestation[i]>293)

{maturity[i]=3}

else {maturity[i]=2}

}

for (i in 1:length(smoke))

{

if (maturity[i]==1 & smoke[i]==1)

{MatSmoke[i]="PreSmoke"}

else if (maturity[i]==1 & smoke[i]==0)

{MatSmoke[i]="PreNoSmoke"}

else if (maturity[i]==2 & smoke[i]==1)

{MatSmoke[i]="NorSmoke"}

else if (maturity[i]==2 & smoke[i]==0)

{MatSmoke[i]="NorNoSmoke"}

else if (maturity[i]==3 & smoke[i]==1)

{MatSmoke[i]="PostSmoke"}

else {MatSmoke[i]="PostNoSmoke"}

}

#1.

S = bwt[smoke == 1]

NS = bwt[smoke ==0]

boxplot(S,NS,xlab ="smoke" , main = "s3957", names =c("Smoked mother", "non-smoked mother"),range = 0)

#The graph "s3957" shows that the centre of baby weight with smoked mother slightly exceeds than the centre of the baby weight with non-smoked mother during the pregnancy.Also, both boxplols are reasonably symmetric.

m1 = bwt[maturity==1]

m3 = bwt[maturity==3]

m2 = bwt[maturity==2]

boxplot(m1,m2,m3,xlab ="smoke" , main = "m3957", names =c("m1", "m2","m3"),range = 0)

#The boxplot "m3957" shows that sample m3(the variation for the baby weight with the gestation greater than 293 days) is comparatively small, and it has the highest centre among others. Also, there is a increase trend for the centres with the increase of the maturity. It is also worth to note that sample m1, m2 and m3 are reasonably symmetric.

presmoke = bwt[MatSmoke=="PreSmoke"]

prenosmoke=bwt[MatSmoke=="PreNoSmoke"]

norsmoke = bwt[MatSmoke=="NorSmoke"]

nornosmoke = bwt[MatSmoke=="NorNoSmoke"]

postsmoke = bwt[MatSmoke=="PostSmoke"]

postnosmoke = bwt[MatSmoke=="PostNoSmoke"]

boxplot(presmoke,prenosmoke,norsmoke,nornosmoke, postsmoke, postnosmoke, xlab ="smoke" , main = "p3957", names =c("presmoke", "prenosmoke","norsmoke","nornosmoke","postsmoke","postnosmoke"),range = 0)

#Overall we see a increase trend of baby weight with the increasing of the maturity, which Presmoke has the smallest centre while postsmoke has the largest. Sample prenosmoke is slight skewed to the right while the rest of sample are reasonably symmetric. Also, sample nornosmoke shares similar centres with sample postsmoke.

#2

#check nomarity:

qqnorm(S)

qqline(S)

qqnorm(NS)

qqline(NS)

#From both grapsh shown from the normal Q-Q plot, we see minimal deviation from the sraight line. This indicates normal distribution.

#check independent:

#Since the data we collected were all independent, so we don't need to conduct a test to check if the assumption of IID is hold.

#check whether the variance are same:

var.test(S, NS)

#The p-value is 0.43. which is bigger than benchmark significant level of 0.05. Therefor, we fail to reject our null hypothesis of their variances are not equal, which means the variance between the mean birth weight between babies born to mothers who were smokes and babies born to mothers who were nonsmokers are same.

t.test(S,NS,var.equal = T)

#We see the p-value for the test is much less than benchmark significant level of 0.05. So we have a strong evidence to reject our null hypothesis. Therefore the mean birth weight between babies born to mothers who were smokers and babies born to mothers who were nonsmokers is not equal.

#3.

results=aov(bwt~maturity, data = data1)

summary(results)

#Studying the output of the ANOVA table above we see that the F-statistic is 113.4 with a p-value equal much less than 0.05. We clearly reject the null hypothesis of equal means of birth weight for all three maturity groups.

pairwise.t.test(bwt,maturity,p.adjust.method = "bonf")

#The results show that all the 1-2,1-2 and 2-3 diferences are significant. Hence, we can conclude that the mean of birth weight is significant different among the all 3 groups in maturity.

#4

result=aov(bwt~MatSmoke, data = data1)

summary(result)

#Studying the output of the ANOVA table above we see that the F-statistic is 36.09 with a p-value equal much less than 0.05. We clearly reject the null hypothesis of equal means of birth weight for all 6 categories in MatSmoke.

pairwise.t.test(bwt,MatSmoke,p.adjust.method = "bonf")

#From the output, we see the p-value for Postnosmoke-Nornosmoke is 0.1625, Prenosmoke-Norsmoke is 0.2824, Postsmoke-Nornosmoke is 1.0, Postsmoke-Postnosmoke is 1.0, which are all greater than our significance level of 0.05. This means we fail to reject the null hypothesis that the mean of birth weight among them are equal. The rest pair, which are Norsmoke-Nornosmoke, Presmoke-Norsmoke, Prenosmoke-Nornosmoke, Prenosmoke-Postnosmoke, Presmoke-Postsmoke... shares equal mean of birth weight.

#5

#check nomaliry:

plot(lm(bwt~MatSmoke), which = 1)

##The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.Also note that Points 165 164 and 409 are detected as outliers, which can severely affect normality and homogeneity of variance.It can be useful to remove outliers to meet the test assumptions.

qqnorm(presmoke)

qqline(presmoke)

qqnorm(prenosmoke)

qqline(prenosmoke)

qqnorm(norsmoke)

qqline(norsmoke)

qqnorm(nornosmoke)

qqline(nornosmoke)

qqnorm(postsmoke)

qqline(postsmoke)

qqnorm(postnosmoke)

qqline(postnosmoke)

#From graph shown from the their normal Q-Q plots, we see minimal deviation from the sraight line. This indicates normal distribution.

#check IID:

#Since the data we collected were all independent, so we don't need to conduct a test to check if the assumption of IID is hold.

bartlett.test(bwt~MatSmoke, data = data1)

#From the output we can see the p-value is 0.09627, which is not less than the significance level of 0.05. This means we cannot reject the null hypothesis that the variance is the same for all 6 categories in MatSmoke. This means the there is no evidence to suggest that the variance in birth weight among 6 categories in MatSmoke is different.

#6

#a)

#The number of predictor variables will be same as in the model used in question 4. This is because two-way ANOVA in question 6, it has smoking status(yes/no) and the the maturity status(maturity[1],maturity[2] and maturity[3]) and (3-1)(2-1) interaction term, which in total there are 5 (2+1+2X1=5) variables. So, it has same amount of variabls as question 4.

#b)The F-test for the presence of interatctin between maturity level and smoking status would be statistically significant, because from the output of the pairwise.t.test for question 4, we see some significant differences, which means that there are significant interaction effects.

#7

#We don't need to concern. There would be no effects for the average bwt of each maturity level with different numbers of babies in each level since one-way anova allow to use unequal number.

#8

#The question asks us to compare the differnce of the "gestitation" as a dummy variable(level 1 with gestitation<259, level 2 with 259 <gestitation <293 and level 3 with gestitation >293) in an additive linear model for the mean of birth weight,or a numerical variable (which consider gestitation as any numerical number). It would make no difference for the output for both consideration.

#The mathematical equation: bwt= beta0+beta1I(smoke==1)+beta2Xgestitation+e

#9

#eg1: mother's alchonic status: yes/no

#eg2: father's smoking status: yes/no

Appendix:

> data1 = read.csv("/Users/jesse/Desktop/bbw.csv")

> attach(data1)

> maturity=array(0,length(gestation))

> MatSmoke=array(0,length(smoke))

> for (i in 1:length(gestation))

+ {

+ if (gestation[i]<259)

+ {maturity[i]=1}

+ else if (gestation[i]>293)

+ {maturity[i]=3}

+ else {maturity[i]=2}

+ }

> for (i in 1:length(smoke))

+ {

+ if (maturity[i]==1 & smoke[i]==1)

+ {MatSmoke[i]="PreSmoke"}

+ else if (maturity[i]==1 & smoke[i]==0)

+ {MatSmoke[i]="PreNoSmoke"}

+ else if (maturity[i]==2 & smoke[i]==1)

+ {MatSmoke[i]="NorSmoke"}

+ else if (maturity[i]==2 & smoke[i]==0)

+ {MatSmoke[i]="NorNoSmoke"}

+ else if (maturity[i]==3 & smoke[i]==1)

+ {MatSmoke[i]="PostSmoke"}

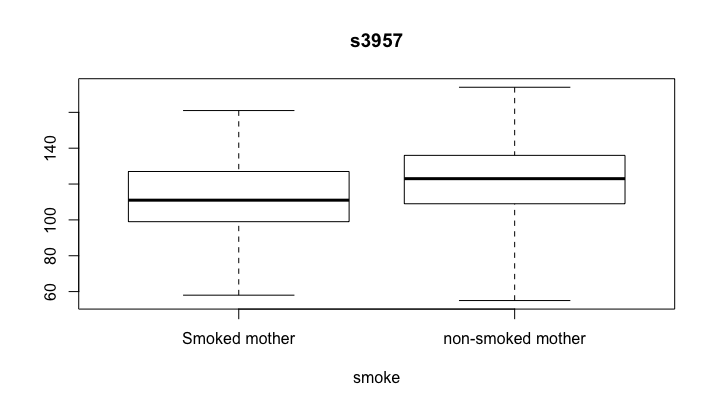
+ else {MatSmoke[i]="PostNoSmoke"}

+ }

> S = bwt[smoke == 1]

> NS = bwt[smoke ==0]

> boxplot(S,NS,xlab ="smoke" , main = "s3957", names =c("Smoked mother", "non-smoked mother"),range = 0)

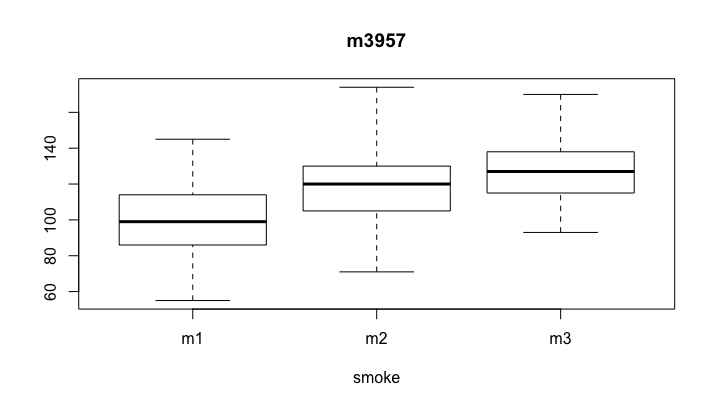


> m1 = bwt[maturity==1]

> m3 = bwt[maturity==3]

> m2 = bwt[maturity==2]

> boxplot(m1,m2,m3,xlab ="smoke" , main = "m3957", names =c("m1", "m2","m3"),range = 0)



> presmoke = bwt[MatSmoke=="PreSmoke"]

> prenosmoke=bwt[MatSmoke=="PreNoSmoke"]

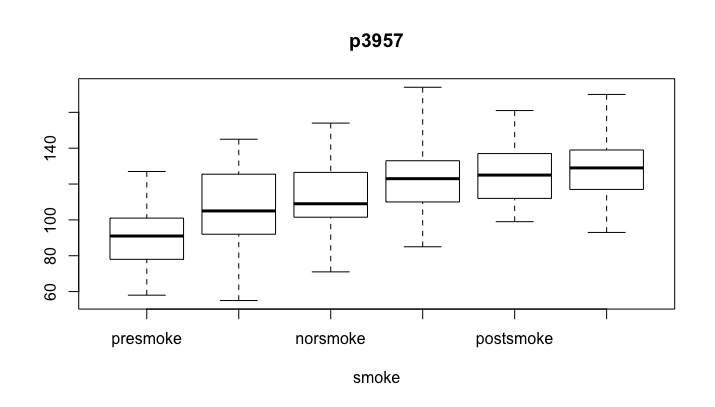
> norsmoke = bwt[MatSmoke=="NorSmoke"]

> nornosmoke = bwt[MatSmoke=="NorNoSmoke"]

> postsmoke = bwt[MatSmoke=="PostSmoke"]

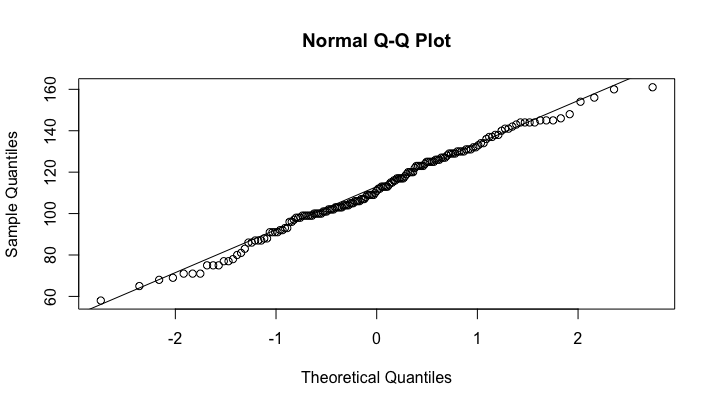
> postnosmoke = bwt[MatSmoke=="PostNoSmoke"]

> boxplot(presmoke,prenosmoke,norsmoke,nornosmoke, postsmoke, postnosmoke, xlab ="smoke" , main = "p3957", names =c("presmoke", "prenosmoke","norsmoke","nornosmoke","postsmoke","postnosmoke"),range = 0)



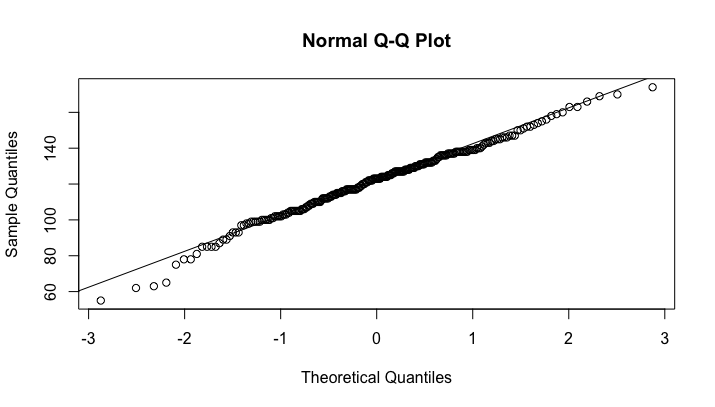
> qqnorm(S)

> qqline(S)



> qqnorm(NS)

> qqline(NS)



> var.test(S, NS)

F test to compare two variances

data: S and NS

F = 1.1178, num df = 162, denom df = 245, p-value = 0.43

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.8469957 1.4873633

sample estimates:

ratio of variances

1.11782

> t.test(S,NS,var.equal = T)

Two Sample t-test

data: S and NS

t = -4.6937, df = 407, p-value = 3.672e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-13.748207 -5.631563

sample estimates:

mean of x mean of y

111.8589 121.5488

> results=aov(bwt~maturity, data = data1)

> summary(results)

Df Sum Sq Mean Sq F value Pr(>F)

maturity 1 44254 44254 133.4 <2e-16 \*\*\*

Residuals 407 135012 332

---

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

> pairwise.t.test(bwt,maturity,p.adjust.method = "bonf")

Pairwise comparisons using t tests with pooled SD

data: bwt and maturity

1 2

2 1.4e-14 -

3 < 2e-16 3.8e-05

P value adjustment method: bonferroni

> result=aov(bwt~MatSmoke, data = data1)

> summary(result)

Df Sum Sq Mean Sq F value Pr(>F)

MatSmoke 5 55448 11090 36.09 <2e-16 \*\*\*

Residuals 403 123818 307

---

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

> pairwise.t.test(bwt,MatSmoke,p.adjust.method = "bonf")

Pairwise comparisons using t tests with pooled SD

data: bwt and MatSmoke

NorNoSmoke NorSmoke PostNoSmoke PostSmoke PreNoSmoke

NorSmoke 0.0114 - - - -

PostNoSmoke 0.1625 2.4e-07 - - -

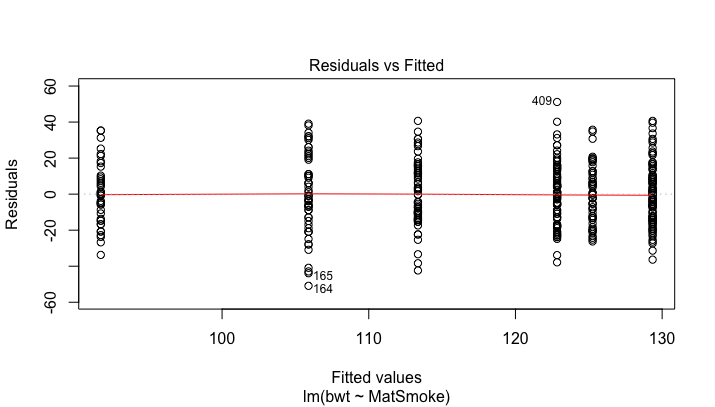
PostSmoke 1.0000 0.0033 1.0000 - -

PreNoSmoke 3.2e-07 0.2824 2.4e-13 2.1e-07 -

PreSmoke < 2e-16 1.7e-08 < 2e-16 < 2e-16 0.0015

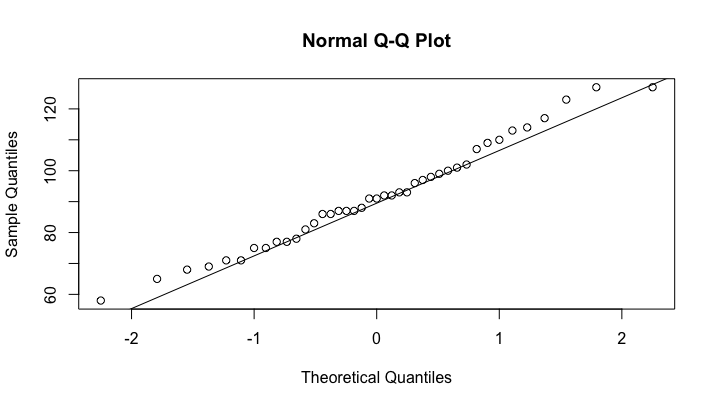
P value adjustment method: bonferroni

> plot(lm(bwt~MatSmoke), which = 1)



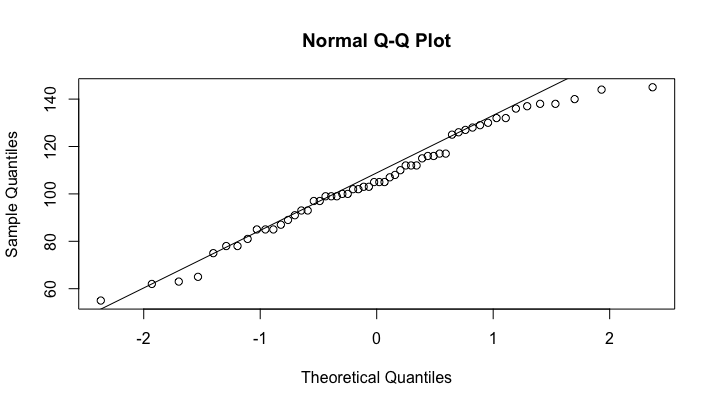
> qqnorm(presmoke)

> qqline(presmoke)



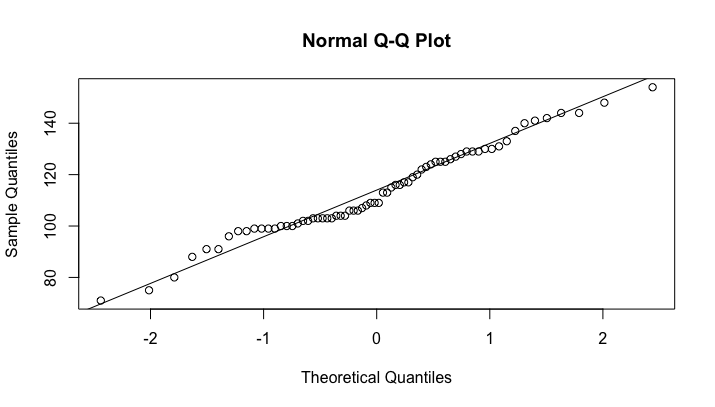
> qqnorm(prenosmoke)

> qqline(prenosmoke)



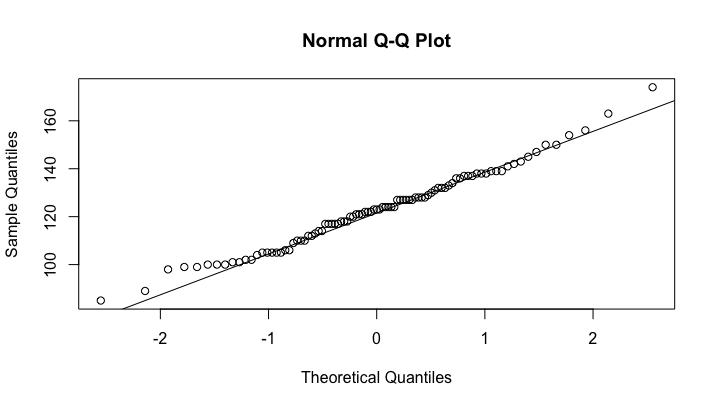
> qqnorm(norsmoke)

> qqline(norsmoke)



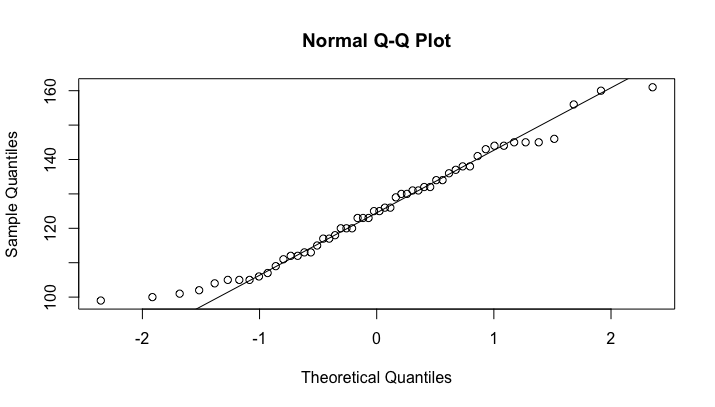
> qqnorm(nornosmoke)

> qqline(nornosmoke)



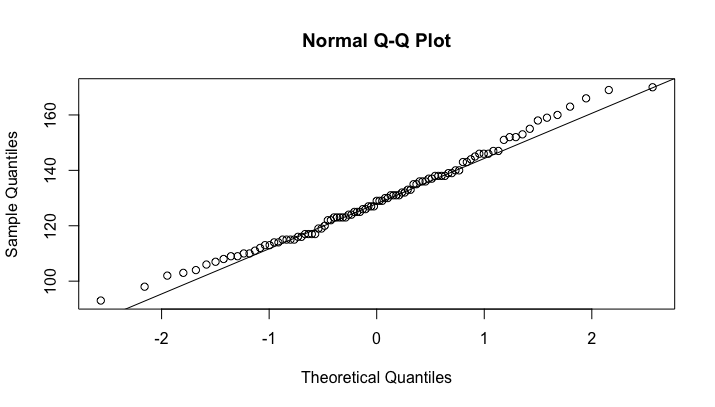
> qqnorm(postsmoke)

> qqline(postsmoke)



> qqnorm(postnosmoke)

> qqline(postnosmoke)



> bartlett.test(bwt~MatSmoke, data = data1)

Bartlett test of homogeneity of variances

data: bwt by MatSmoke

Bartlett's K-squared = 9.3393, df = 5, p-value = 0.09627