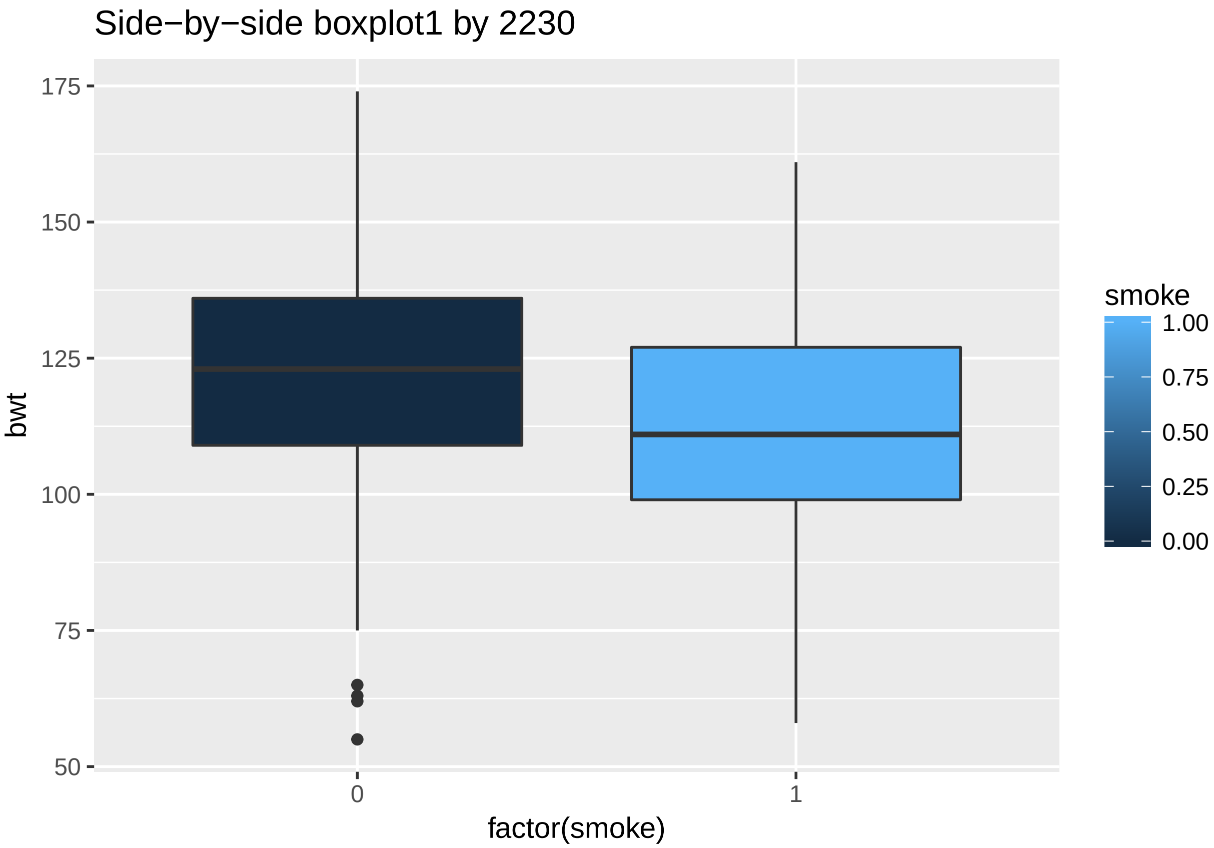
Assignment 2

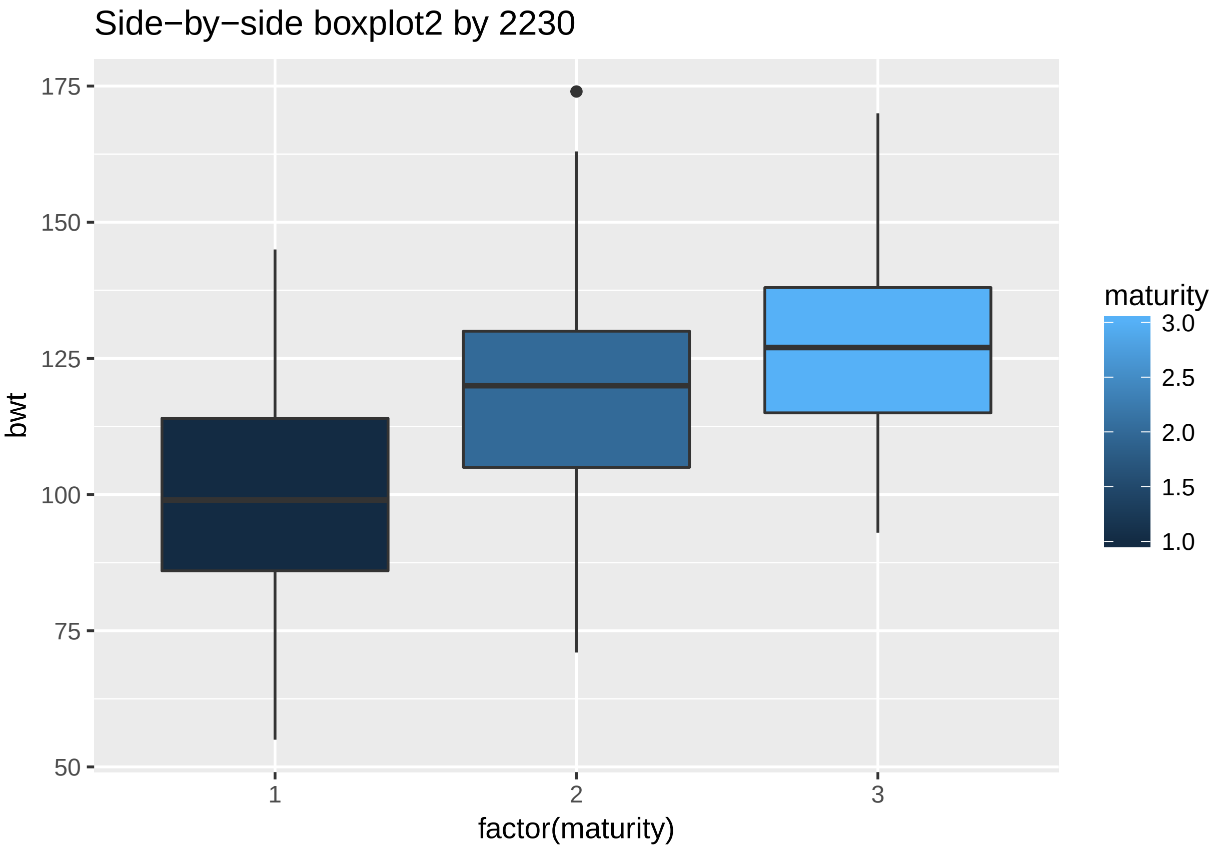
Last name: Du
First name: Min
Student ID: 1002602230

(1)

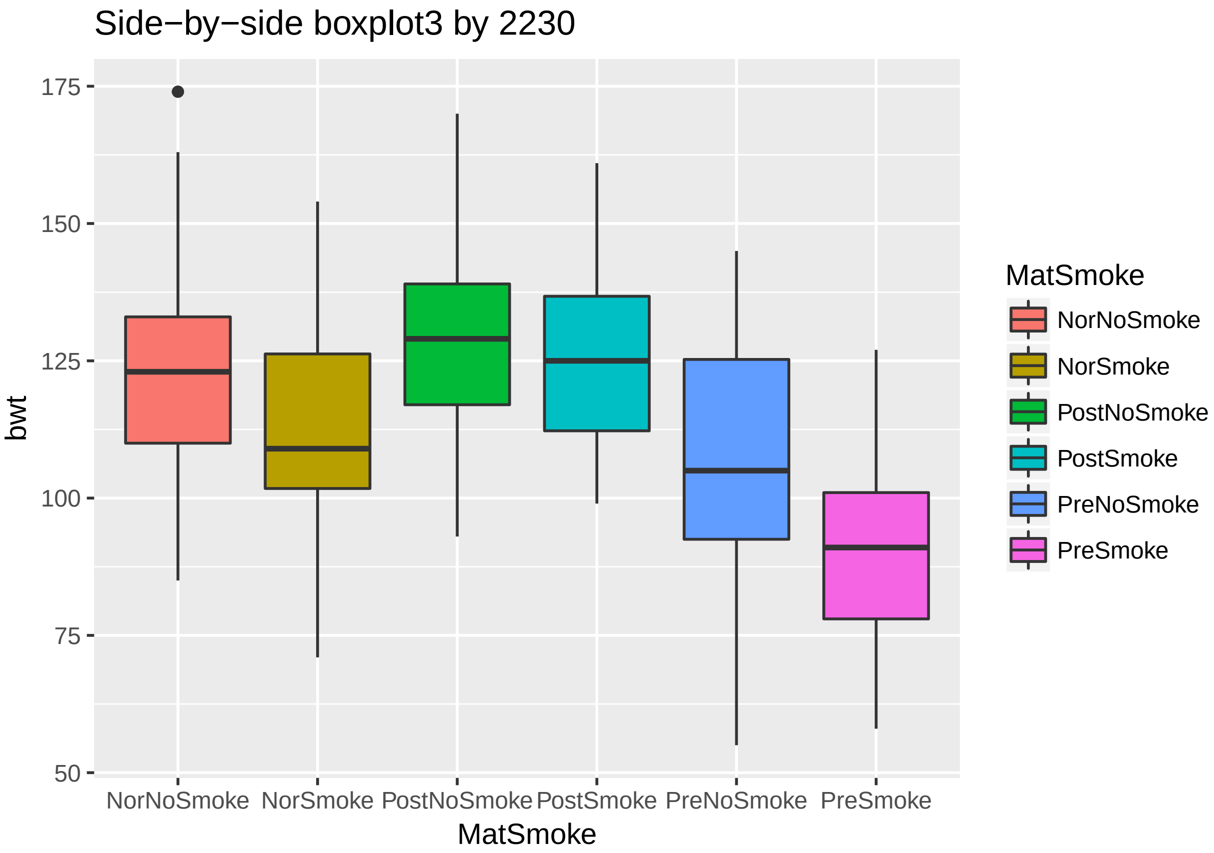
1. Side-by-side boxplots of birth weight between mothers who smoked and those who did not smoke during pregnancy.



2. Side-by-side boxplot of birth weight among the three maturity levels.



3. Side-by-side boxplot of birth weight among the 6 categories of babies grouped by the combination of their maturity level and maternal smoking status.



From the first side-by-side boxplot between mothers who smoked and who did not smoke, the birthweight of babies whose mother did not smoke is around 124 ounces, the other is 113 ounces, birthweight of baby's mother did not smoke is 11 ounces larger than smoked.

From the second side-by-side boxplot among three maturity levels, the median of birthweight of maturity1 is about 99 ounces, which is the smallest among three maturity levels; the median of birthweight of maturity 2 is about 122 ounces; the median of birthweight of maturity 3 is around 126 ounces, which is the largest among three maturity levels.

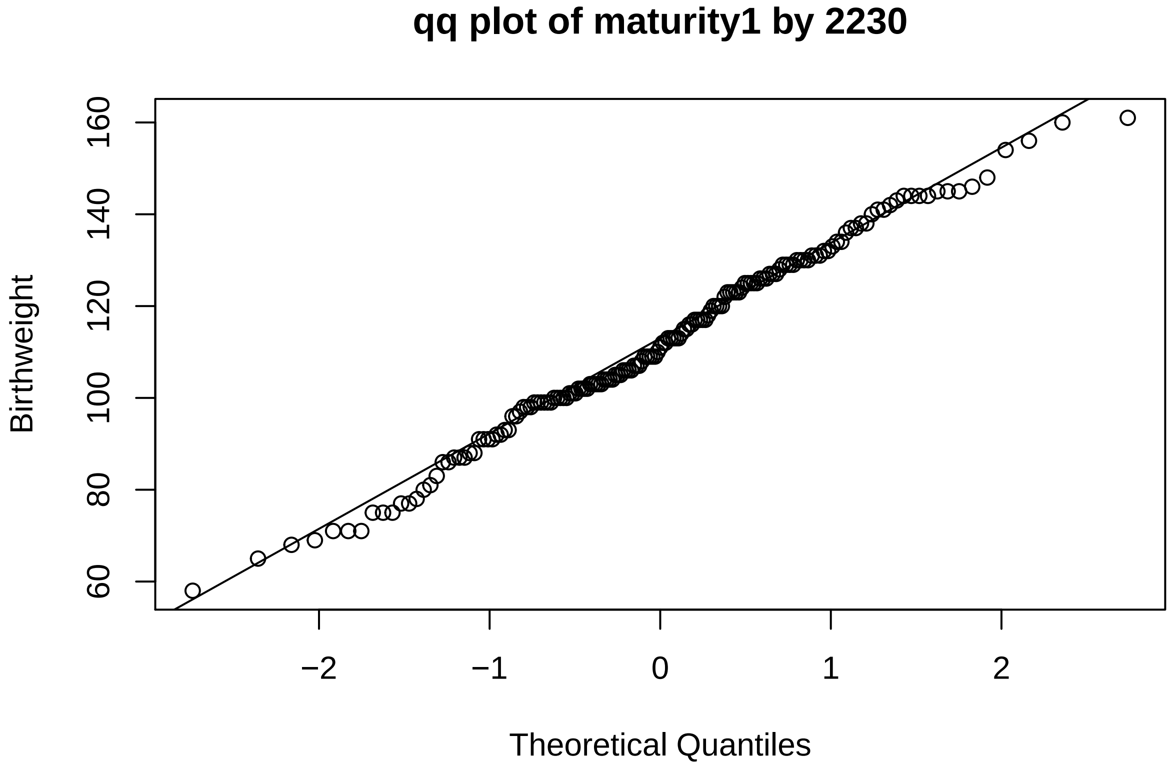
From the third side-by-side boxplot among the 6 categories of babies grouped by the combination of mother's maturity level and maternal smoking status. The median birthweight of babies from smallest to largest is PreSmoke < PreNoSmnoke < NorSmoke < NorNoSmoke < PostSmoke < PostNoSmoke. Under the same gestational age, the birthweight of baby whose mother did not smoke is heavier than smoked. Also, for those mothers who smoked, the larger gestation age, the heaveir baby's birthweight.

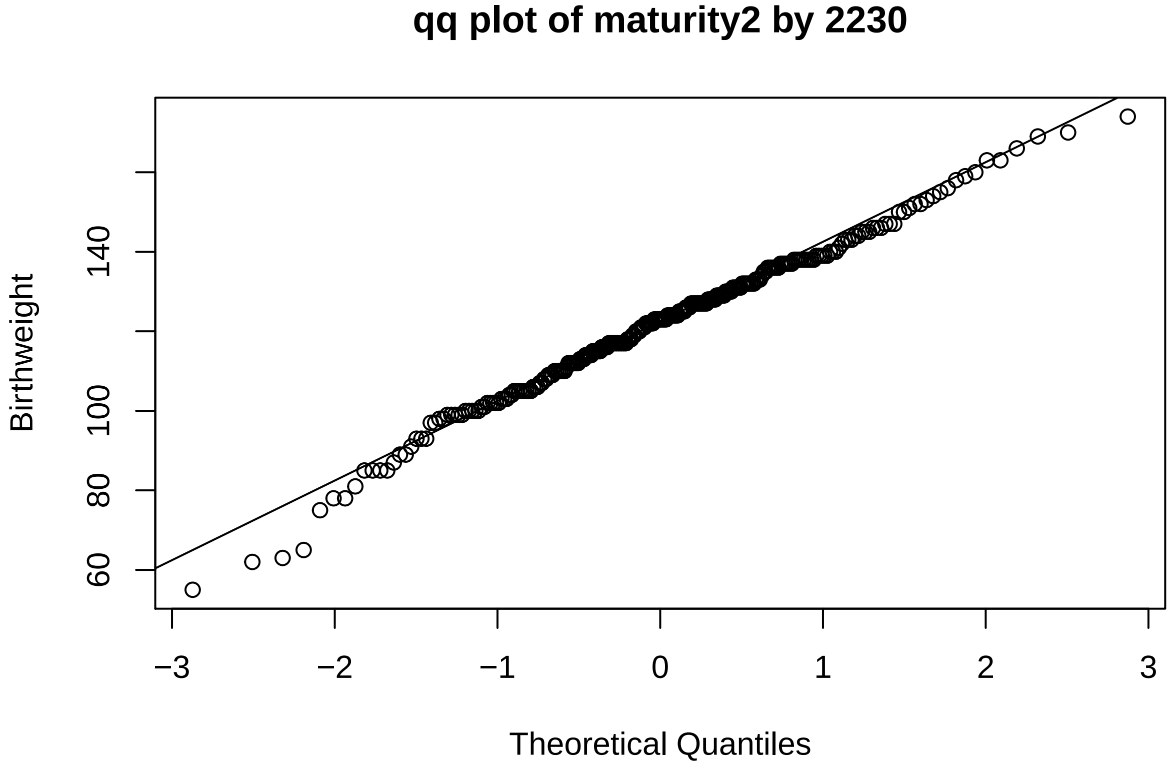
(2)

Assumptions: 1.QQ plots below show that samples follows normal distribution. 2.Equal variances assumption satisfied by F-test.

1.According to the below normal qq plots for both groups, most observations lie along the

45-degree line in the QQ-plots, so we may assume that normality holds here.





2.Check whether the variance of group of smoker and nonsmoker is equal. Using F-test, $H_0 : \sigma_1^2/\sigma_2^2 = 1, H_a : \sigma_1^2/\sigma_2^2 \neq 1$,we get p-value=0.43, which is greater than 0.05, then fail to reject H_0, the variance of the group of smoker and nonsmoker are equal.

```
##
##  F test to compare two variances
##
## data:  smoker and non_smoker
## 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
```

Since the above assumptions are satisfied, then we can carry two sample pooled t-test.H_0 : $\mu_1 = \mu_2$, $H_a : \mu_1 \neq \mu_2$, p-value < 3.672e-06, which is smaller than 0.05, Then we reject H_0, there is a difference in the mean birth weight between babies whose mother did not smoke and smnoked.

```
##
##  Two Sample t-test
##
## data:  smoker and non_smoker
## 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
```

(3)

One-way analysis of variance test, whether or not there is a difference in mean birth weight among babies classified by gestational maturity.

```
##         1         2         3
##  99.90722 118.83230 127.88742
```

From the ouput, there is significant difference between the mean of three levels of maturity.

```
##                   Df Sum Sq Mean Sq F value Pr(>F)    
## factor(maturity)   2  46586   23293   71.28 <2e-16 ***
## Residuals        406 132680     327                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to one-way analysis of variance. P-value of F-test is <2e-16, which is smaller than

0.05, then reject H_0, the mean birth weight among babies classified by gestational maturity is different.

In order to see which levels of maturity differ, using bonferroni's method.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  bbw2230$bwt and maturity
##
##   1       2
## 2 1.4e-14 -
## 3 < 2e-16 3.8e-05
##
## P value adjustment method: bonferroni
```

From the output above, we can see all of the three p-values are smaller than significance level of 0.05, which means there are significant differences between the mean value of all three levels.

(4)

One-way analysis of variance test, Whether or not there is a difference in mean birth weight among the six categories of babies.

```
##  NorNoSmoke    NorSmoke PostNoSmoke   PostSmoke  PreNoSmoke    PreSmoke
##   122.83871   113.35294   129.35052   125.25926   105.89286    91.73171
```

From the output, there are difference in the mean value among 6 categories.

```
##              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
```

According to the one-way Anova test, p-value of F-test is <2e-16, which is smaller than 0.05, we reject H_0, the mean birth weight among 6 categories is different.

In order to see which categories differ, using bonferroni's method.

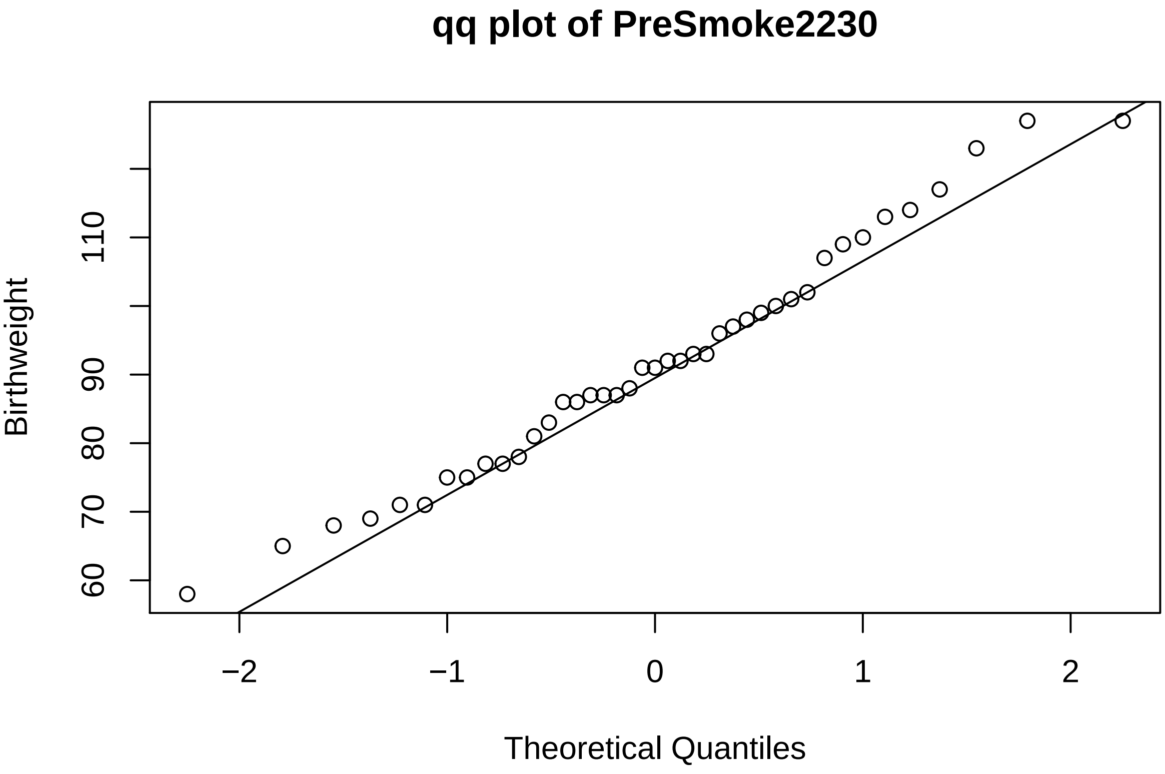
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  bbw2230$bwt and MatSmoke
##
##             NorNoSmoke NorSmoke PostNoSmoke PostSmoke PreNoSmoke
## NorSmoke    0.0114     -         -           -         -
## PostNoSmoke 0.1625     2.4e-07   -           -         -
## PostSmoke   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
```

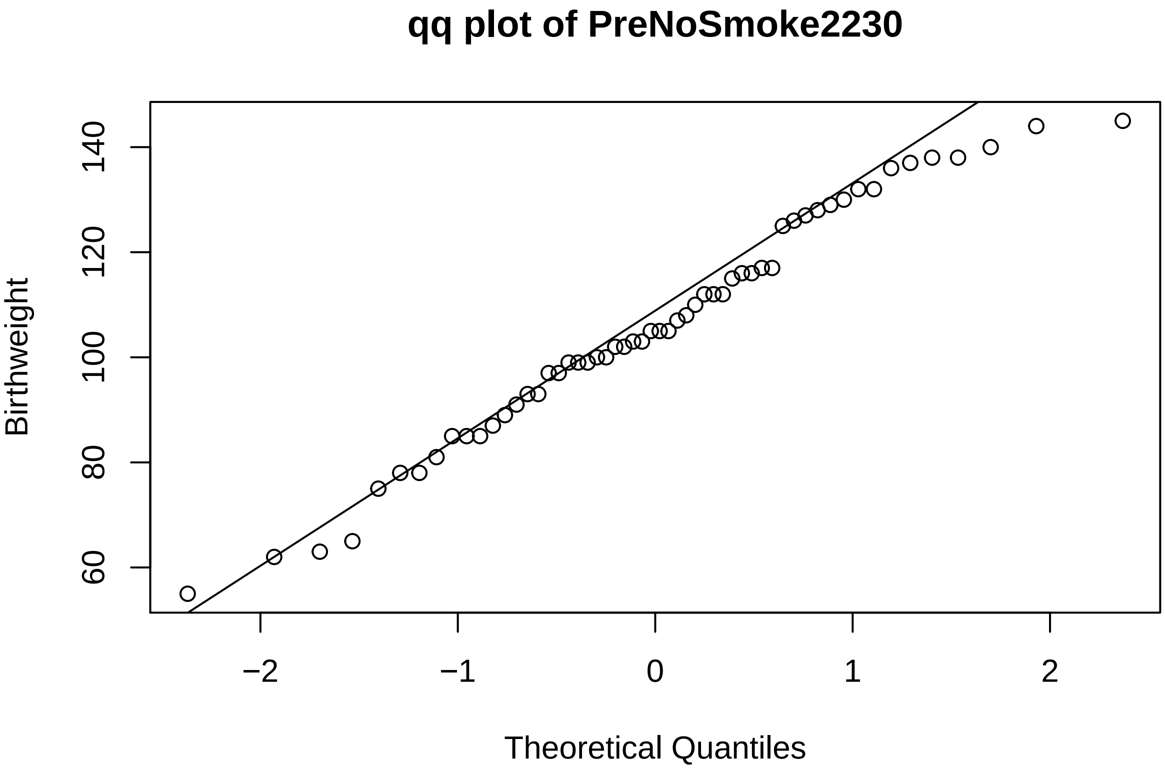
From the output, we observe that the p-value of NorNoSmoke and NorSmoke, NorNoSmoke and PreNoSmoke, NorNoSmoke and PreSmoke, NorSmoke and PostNoSmoke,NorSmoke and PostSmoke, NorSmoke and PreSmoke,PostNoSmoke and PreNoSmoke, PostNoSmoke and PreSmoke, PostSmoke and PreNoSmoke,PostSmoke and PreSmoke, PreNoSmoke and PreSmoke are smaller than 0.05, which means there are significant differences between these MatSmoke means.

(5)

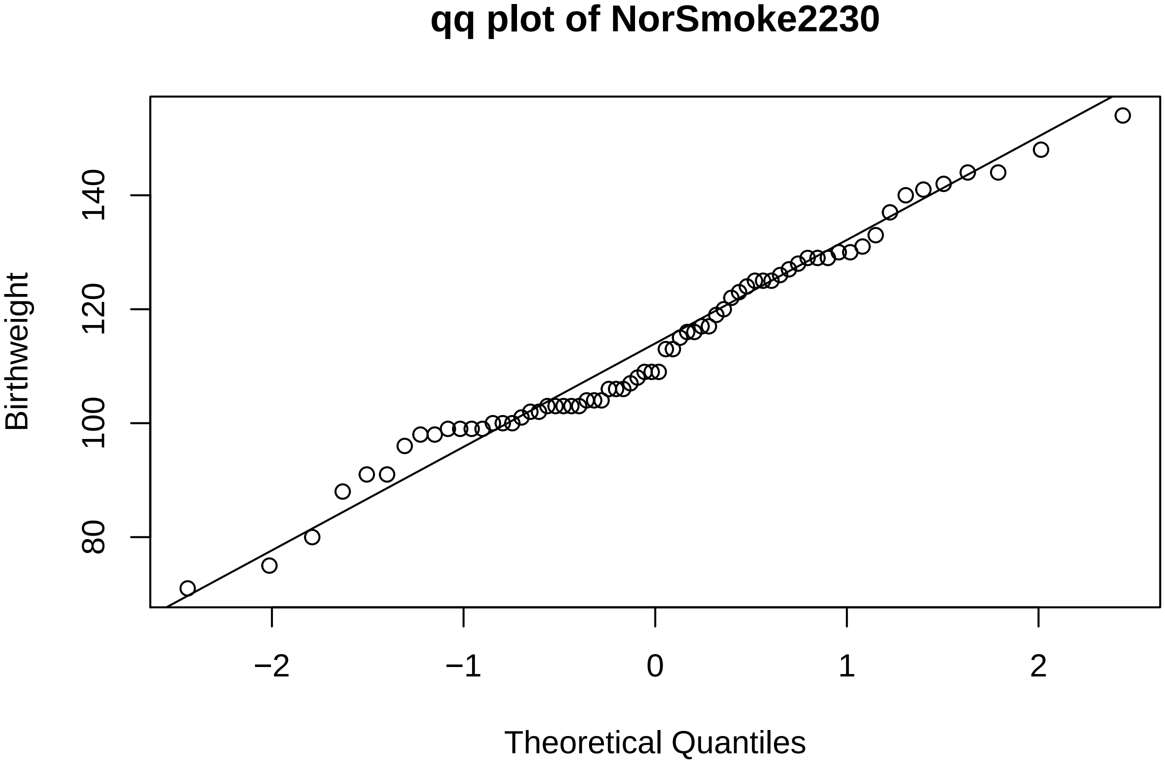
Yes, I trust the result of the statistical tests carried out in question 4.

Check assumptions: 1.whether all levels form Normal populations. 2.Variances are equal.

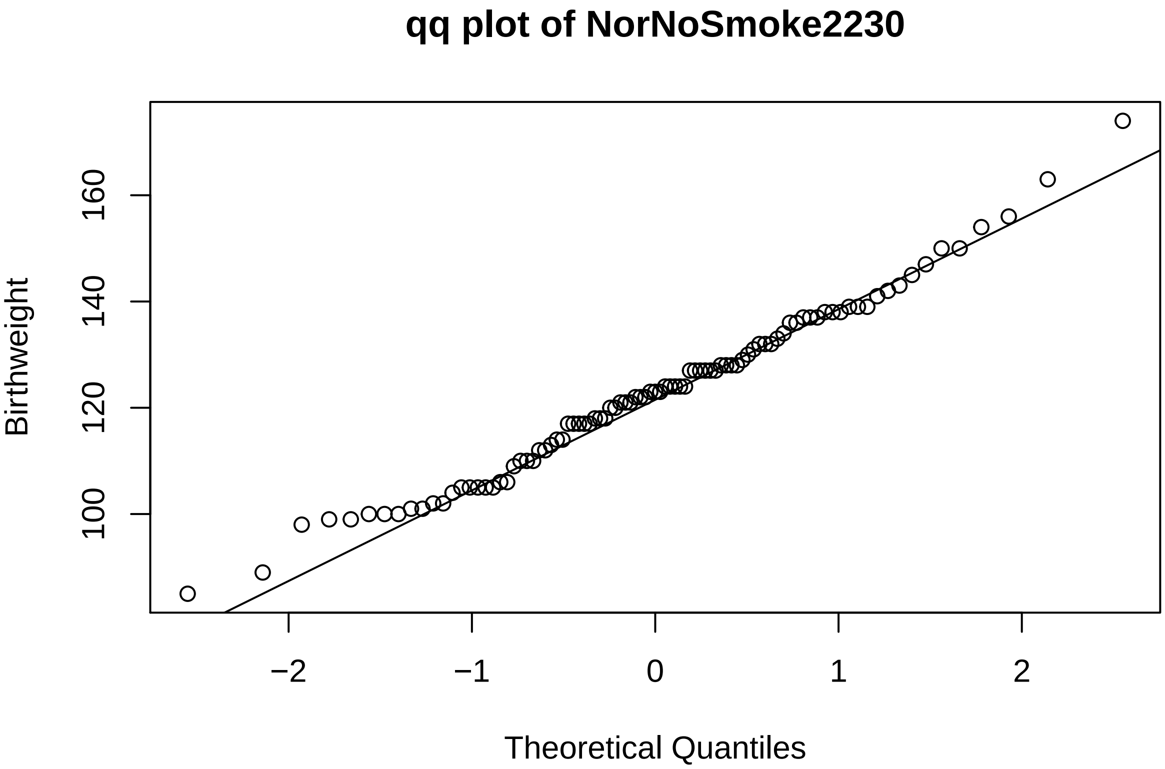


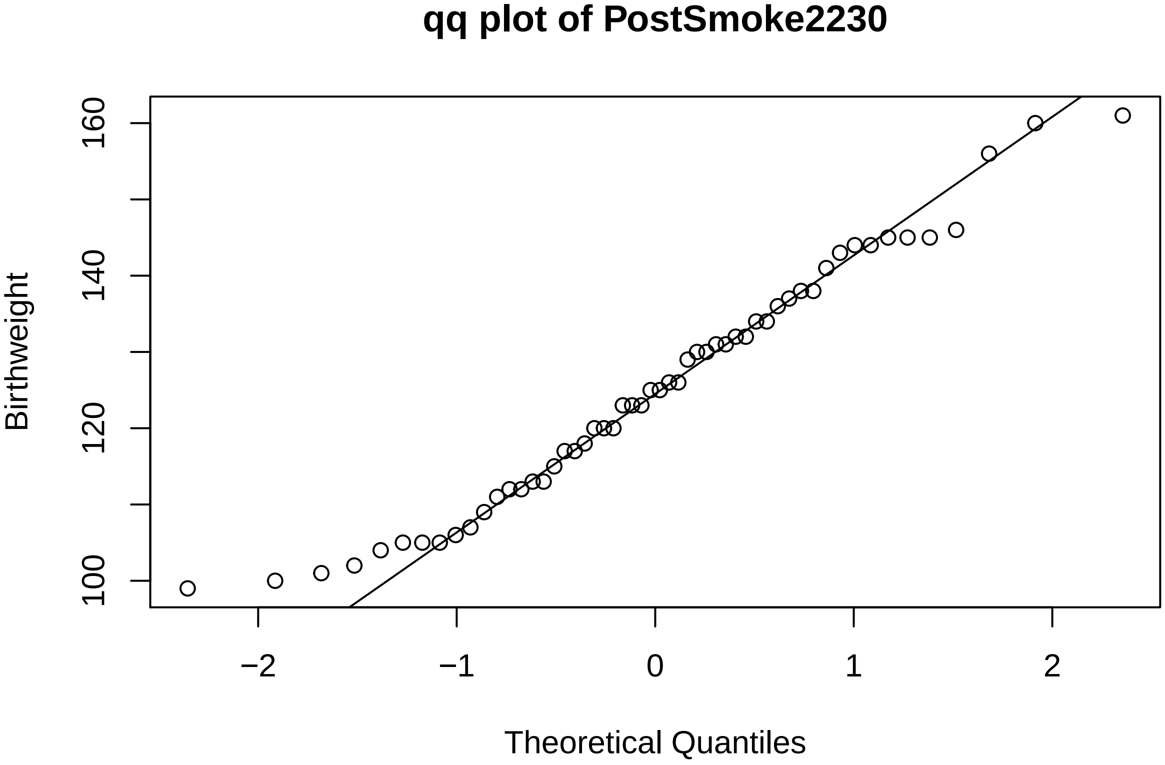


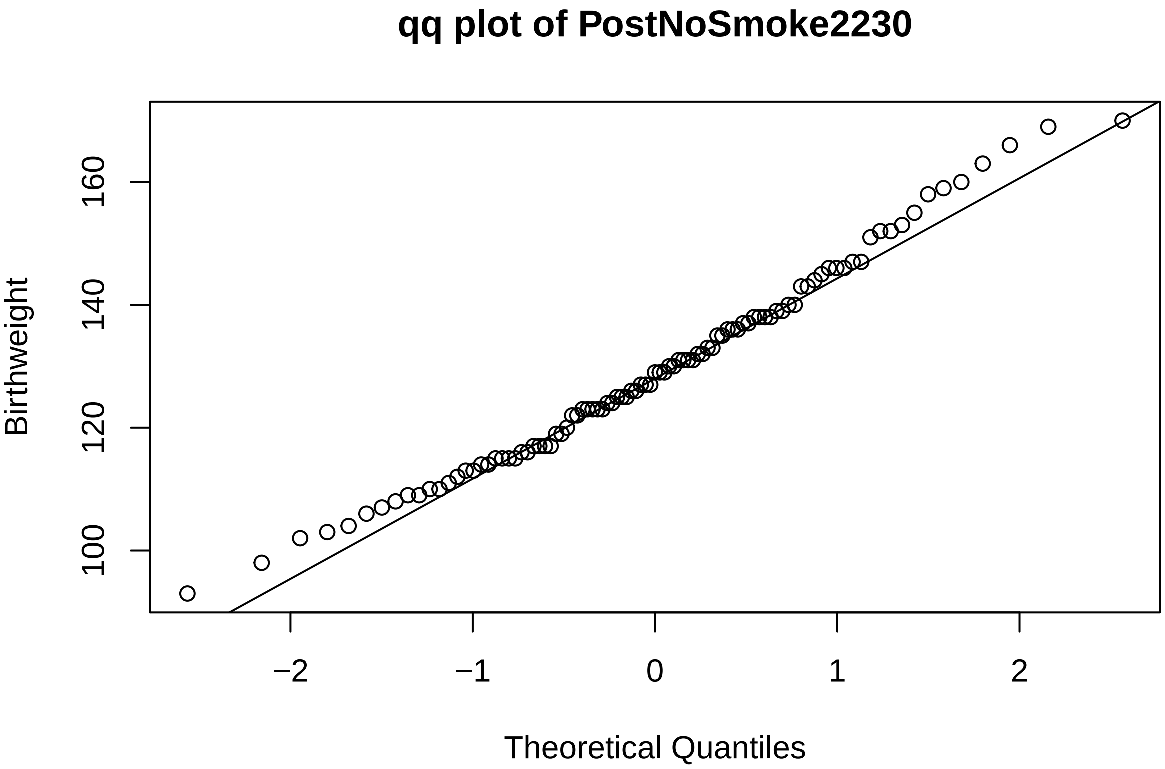
qq plot of PreNoSmoke2230



qq plot of NorSmoke2230


qq plot of NorNoSmoke2230
Birthweight
100
120
140
160
−2
−1
0
1
2
Theoretical Quantiles


qq plot of PostSmoke2230
Birthweight
100
120
140
160
−2
−1
0
1
2
Theoretical Quantiles



```
##
##  Bartlett test of homogeneity of variances
##
## data:  bbw2230$bwt by MatSmoke
## Bartlett's K-squared = 9.3393, df = 5, p-value = 0.09627
```

According to the normal qq plot of each levels, all of the birthweight of the MatSmoke levels follow normality.

Using barlett test, It can be seen that the p-value of 0.09627 is bigger than the significance level of 0.05, we do not reject H_0, the variance are equal.

Therefore, the necessary assumptions of the model hold.

(6)

a) Yes, they are the same.

There are 2 smoking status, 3 maturity levels. Then there are 1+2+2=5 predictor variables in Two-way analysis of variance.

Also since there are 6 categories of babies classfied by the combination of maturity level and mother's. Then there are 5 predictor variables.
b) Yes, the presence of interaction between maturity level and smoking status would be statistically significant.

In question4, some MatSmoke terms are significant, which means there are significant interaction effects, then the presence of interaction between maturity level and smoking status should also be statistically significant.

(7)

No, we do not need to concern the data contained different numbers of babies in the three maturity levels.

Check whether variance is equal.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  bbw2230$bwt by maturity
## Bartlett's K-squared = 8.6637, df = 2, p-value = 0.01314
```

From the output, p-value = 0.01314, which is smaller than significance level 0.05, so we reject H_0, variance is not equal.

Since variance is different, different number of babies will impact variance, then different number of babies in three maturity levels do not need to be concerned.

(8)

$bwt = a_0 + a_1 I_{smoke} + a_2 I_{level1,maturity} + a_3 I_{level2,maturity} + e$,(e is the error term)

$bwt = I(smoke) + gestation$

In the first equation, gestation is a factor. In the second regression, gestation is a quantitative variable.

If gestation is a factor in an additive linear model, we can investigate whether or not there is a difference in mean birth weight among babies classified by gestational maturity by using one-way anova. Furthermore, we can use bonferroni test or tukey HSD test to see which levels of maturity differ.

If gestation is a quantitive variable, we can calculate the difference between birthweight and gestation, while the indicate variable smoke is equal to 0 or 1.

(9)

(a)Sex of baby: 2levels, girl or boy.

(b)Drinking status: 2 levels,drinking = 1, no drinking = 0.

Appendix

(1)

1. Side-by-side boxplots of birth weight between mothers who smoked and those who did not smoke during pregnancy.

```
bbw2230 = read.csv("/Users/mindu/Desktop/STA303/Assignment2/bbw.csv",header=TRUE)
gestation=bbw2230$gestation
smoke=bbw2230$smoke
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"}
  }


library(ggplot2)
ggplot(bbw2230, aes(x=factor(smoke),y=bwt, fill=smoke))+geom_boxplot()+ggtitle("Side-by
```

2. Side-by-side boxplot of birth weight among the three maturity levels.

```
library(ggplot2)
ggplot(bbw2230, aes(x=factor(maturity),y=bwt, fill=maturity))+geom_boxplot()+ggtitle("S
```

3. Side-by-side boxplot of birth weight among the 6 categories of babies grouped by the

combination of their maturity level and maternal smoking status.

```
library(ggplot2)
ggplot(bbw2230, aes(x=MatSmoke,y=bwt, fill=MatSmoke))+geom_boxplot()+ggtitle("Side-by-s
```

(2)

```
qqnorm(bbw2230$bwt[bbw2230$smoke==1],main="qq plot of maturity1 by 2230",ylab="Birthweig
qqline(bbw2230$bwt[bbw2230$smoke==1])
qqnorm(bbw2230$bwt[bbw2230$smoke==0],main="qq plot of maturity2 by 2230",ylab="Birthweig
qqline(bbw2230$bwt[bbw2230$smoke==0])
```

```
smoker=bbw2230$bwt[bbw2230$smoke==1]
non_smoker=bbw2230$bwt[bbw2230$smoke!=1]
var.test(smoker,non_smoker)
```

```
#two sample t-tset
t.test(smoker,non_smoker,var.equal = T)
```

(3)

```
tapply(bbw2230$bwt,maturity,mean)
```

```
summary(aov(bbw2230$bwt~factor(maturity)))
```

```
pairwise.t.test(bbw2230$bwt,maturity,p.adj= "bonf")
```

(4)

```
tapply(bbw2230$bwt,MatSmoke,mean)
```

```
summary(aov(bbw2230$bwt~MatSmoke))
```

```
pairwise.t.test(bbw2230$bwt,MatSmoke,p.adj= "bonf")
```

(5)

```
#check assumptions for one-way anova
#check the data are approximately normal for each level of maturity
qqnorm(bbw2230$bwt[MatSmoke == "PreSmoke"],main="qq plot of PreSmoke2230",ylab="Birthwei
qqline(bbw2230$bwt[MatSmoke == "PreSmoke"])
qqnorm(bbw2230$bwt[MatSmoke == "PreNoSmoke"],main="qq plot of PreNoSmoke2230",ylab="Birt
qqline(bbw2230$bwt[MatSmoke == "PreNoSmoke"])
qqnorm(bbw2230$bwt[MatSmoke == "NorSmoke"],main="qq plot of NorSmoke2230",ylab="Birthwei
qqline(bbw2230$bwt[MatSmoke == "NorSmoke"])
qqnorm(bbw2230$bwt[MatSmoke == "NorNoSmoke"],main="qq plot of NorNoSmoke2230",ylab="Birt
qqline(bbw2230$bwt[MatSmoke == "NorNoSmoke"])
qqnorm(bbw2230$bwt[MatSmoke == "PostSmoke"],main="qq plot of PostSmoke2230",ylab="Birthw
qqline(bbw2230$bwt[MatSmoke == "PostSmoke"])
qqnorm(bbw2230$bwt[MatSmoke == "PostNoSmoke"],main="qq plot of PostNoSmoke2230",ylab="Bi
qqline(bbw2230$bwt[MatSmoke == "PostNoSmoke"])

#check if variance are equal
bartlett.test(bbw2230$bwt~MatSmoke)
```

(7)

```
bartlett.test(bbw2230$bwt~maturity)
```