STA303_A2

Solution

1. (15 marks) Create two new variables: (1) maturity- by converting gestational age to a factor with 3 levels; 1 if the baby was preterm and spent less than 259 days in the womb, 3 if gestational age was beyond 293 and 2 otherwise, and (2) MatSmoke- a variable that combines maturity level and maternal smoking status.Construct three sets of side-by-side boxplots: 1. to compare birth weight between mothers who smoked and those who did not smoke during pregnancy, 2. to compare birth weight among the three maturity levels, and 3. to compare birth weight among the 6 categories of babies grouped by the combination of their maturity level and maternal smoking status. Do there appear to be any differences?

Two new variables

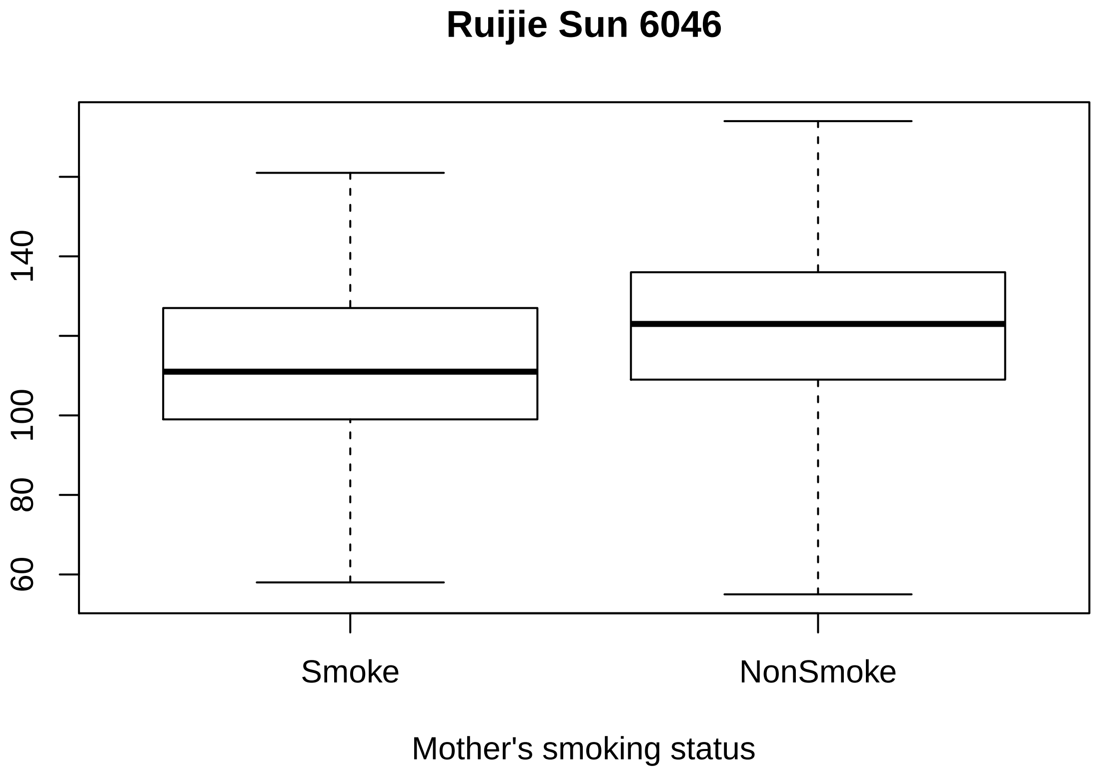
```
head(maturity)
```

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

```
head(MatSmoke)
```

```
## [1] "PreSmoke" "PreSmoke" "PreSmoke" "PreSmoke" "PreSmoke" "PreSmoke"
```

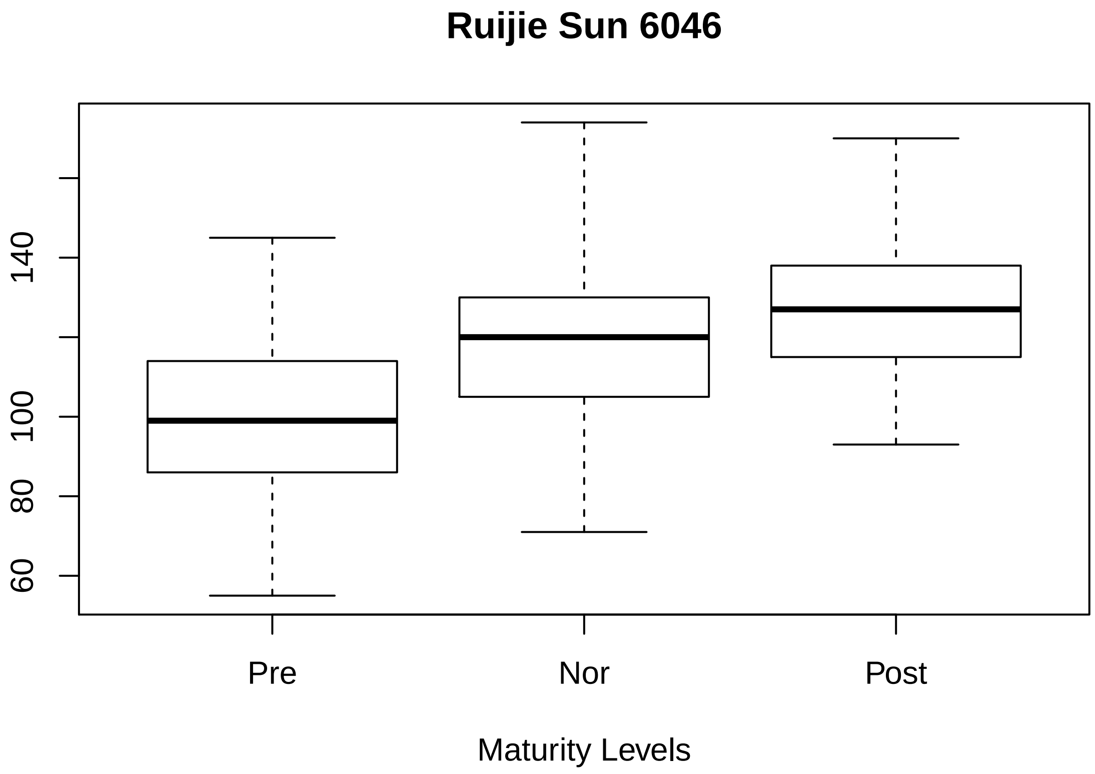
i.side-by-side boxplot of birth weight between mothers who smoked and those who did not smoke during pregnancy.



Conclusion:

Based on the box plot above, there is difference in birth weight between mothers who smoked and those who did not smoke during pregnancy.

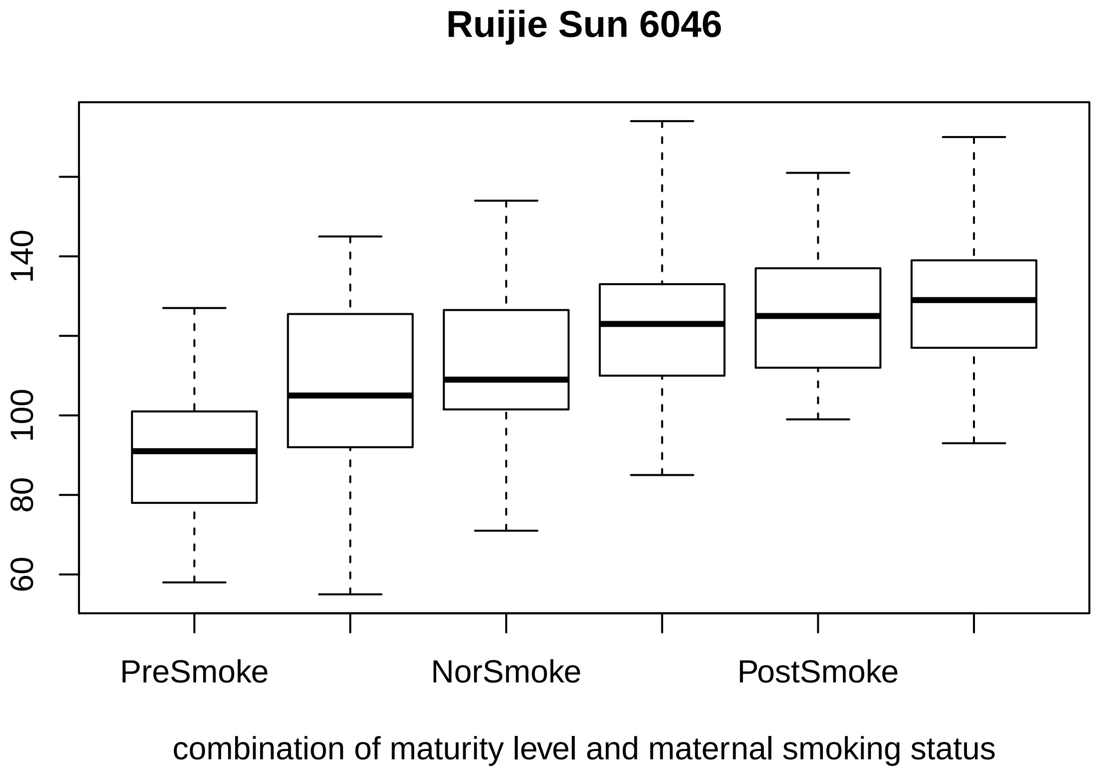
ii.side-by-side boxplot to compare birth weight among the three maturity levels.



Conclusion:

Based on the box plot above, there is difference in birth weight among the three maturity levels.

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



Conclusion:

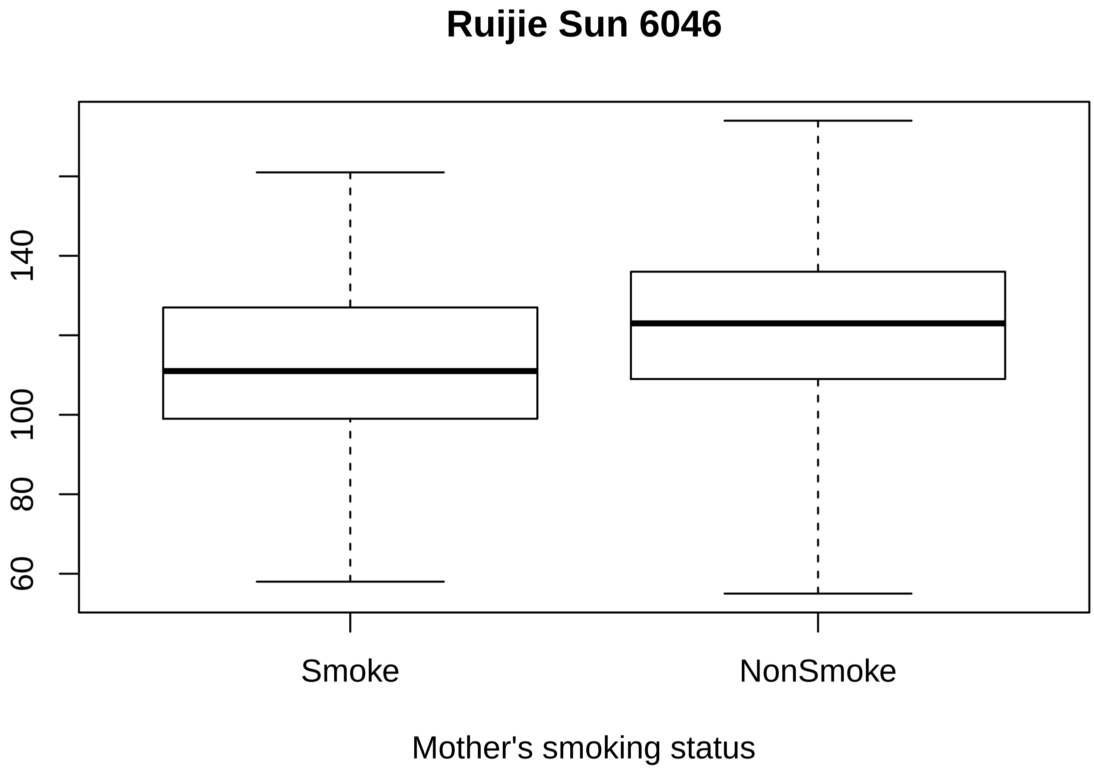
1. In term of both median and variance,compared to the baby's weight of PreSmoke group, there is obvious difference with PreNonSmoke group's baby's weight.
2. In term of median,compared to the baby's weight of NorSmoke group, there is obvious difference with NorNonSmoke group's baby's weight. But their variances are close.
3. In term of median ,there is no big difference between PostSmoke and PostNonSmoke groups.But the variance of PostNonSmoke group is bigger than PostSmoke group.

Summary:

According to the 3 box-plots above, given the different babies mother's smoking status, time in womb, and corresponding intersection, the mean weights are very likely different.

2. (10 marks) Using the R t.test procedure, investigate whether or not there is a difference in the mean birth weight between babies born to mothers who were smokers and babies born to mothers who were nonsmokers.

i. Side-by-side boxplots



ii.Null and Alternative Hypothesis

Null Hypothesis: the average weight of babies born to mother who were smokers is equal to the the average weight of babies born to mother who were nonsmokers.

Alternative Hypothesis:the average weight of babies born to mother who were smokers is not equal to the the average weight of babies born to mother who were nonsmokers.

iii. A test statistic and it's distribution

To test if the two sample have same variance

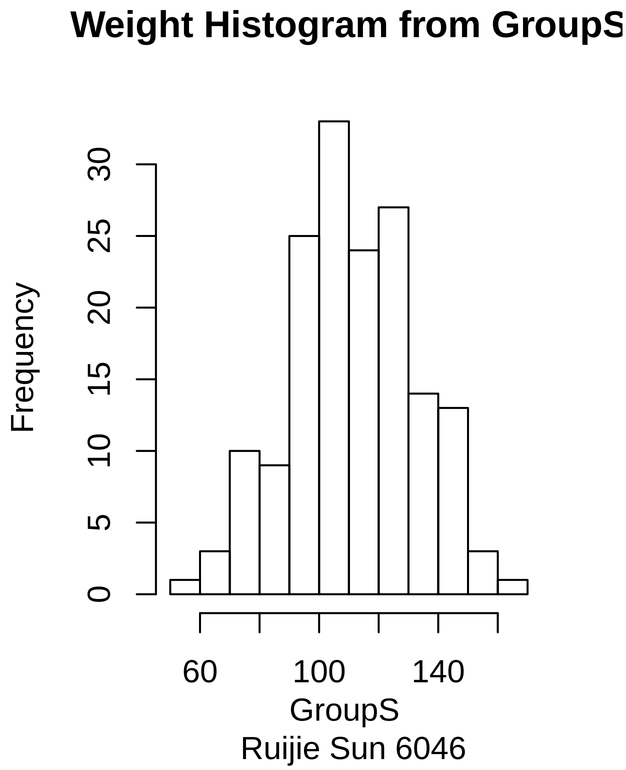
Based on the result from var.test(), p-value $= 0.43 > 0.05$. So the two group have same variance.

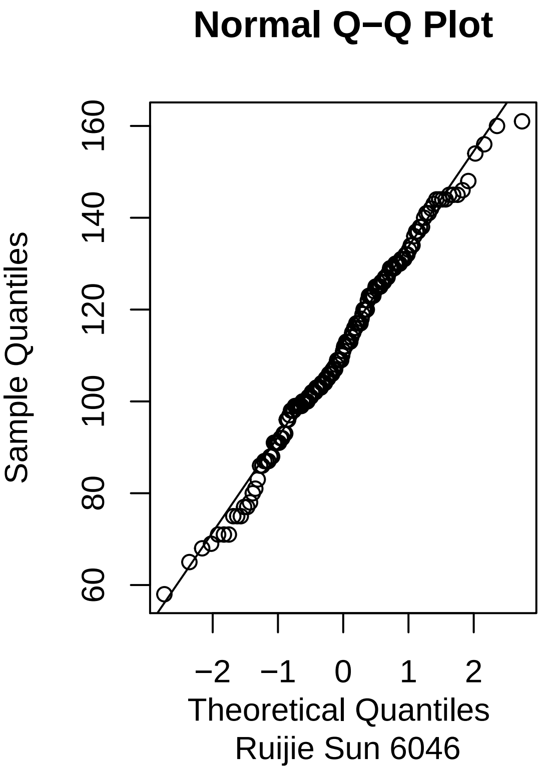
test-statistic $= \frac{\overline{x}-\overline{y}}{{s_p}^2(\frac{1}{n_x}+\frac{1}{n_y})} = -4.6937$ follows t distribution with degree freddom 407, where ${s_p}^2 = \frac{(n_x-1){S_x}^2+(n_y-1){S_y}^2}{n_x+n_y-2}$

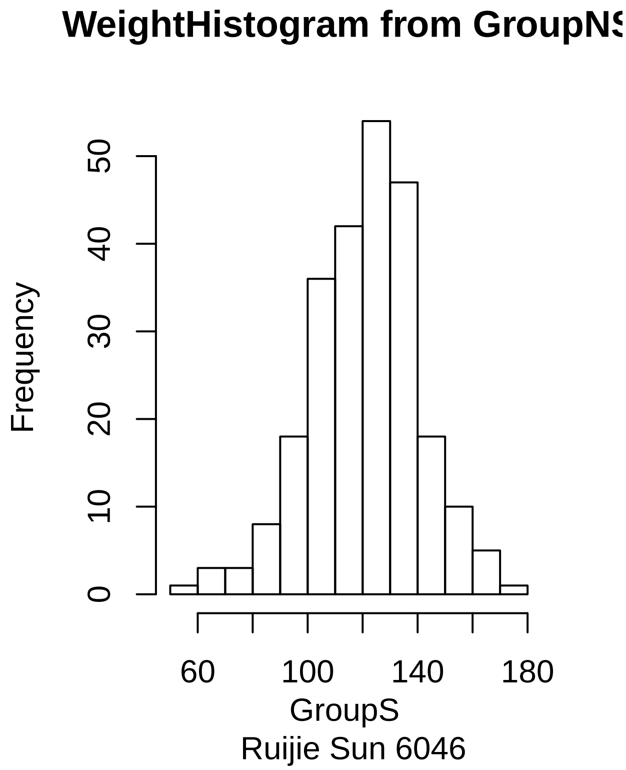
iv. Test assumptions

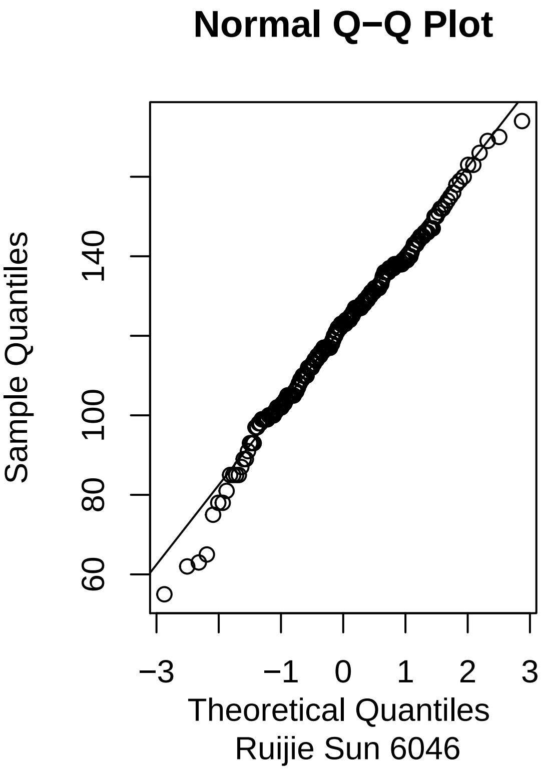
1. The two samples are iid from approximateluy Normal population.
2. The two samples are independent of each other.

v. Test diagnostics









Compared to GroupNS, data from GroupS is more approximately Normal. But sample of GroupNS is still acceptable for normality.

vi P-value

```
##
##  Two Sample t-test
##
## data:  GroupS and GroupNS
## 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
```

p-value = 3.672e-06 < 0.05

vii Results:

So there is sufficient evidence to reject Null Hypothesis. Thus, there is a difference in the mean birth weight between babies born to mothers who were smokers and babies born to mothers who were nonsmokers.

3. (15 marks) Investigate whether or not there is a difference in mean birth weight among babies classified by gestational maturity, using a one-way analysis of variance. If there is a difference among the levels of maturity, carry out an appropriate analysis to see which levels of maturity differ.

i.

Result of One-Way Anova:

```
##                   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
```

P < 2e-16. So there is sufficient evidence to reject Null Hypothesis. There is a difference in mean birth weight among babies classified by gestational maturity

ii.

Bonferroni's method

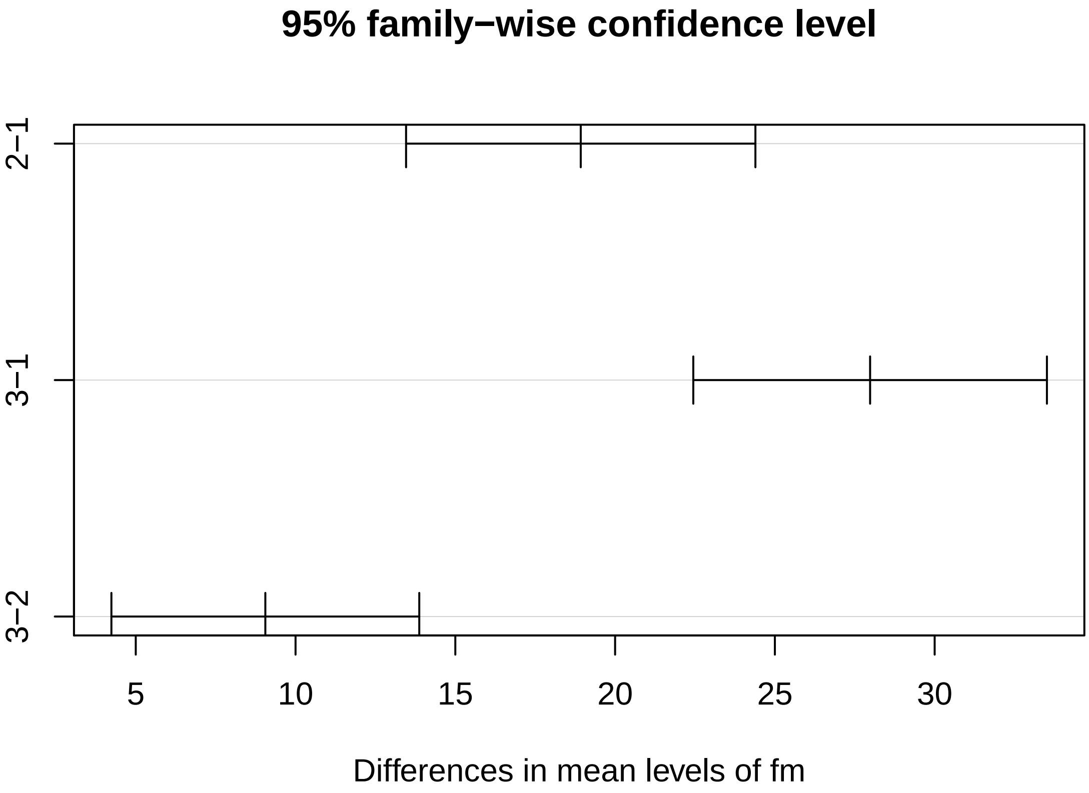
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  bwt and factor(maturity)
##
```

```
##   1       2
## 2 1.4e-14 -
## 3 < 2e-16 3.8e-05
##
## P value adjustment method: bonferroni
```

The p-values are 1.4e-14, 2e-16, 3.8e-05. The three levels of maturity all differ to each other based on Bonferroni's method.

Tukey's method

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = bwt ~ fm)
##
## $fm
##          diff      lwr      upr    p adj
## 2-1 18.925082 13.45932 24.39084 0.00e+00
## 3-1 27.980201 22.44681 33.51359 0.00e+00
## 3-2  9.055119  4.23769 13.87255 3.74e-05
```



Since 0 does not fall into any line, the result is same as Bonferroni's method. They all differ with each other.

4. (15 marks) Use one-way analysis of variance to investigate whether or not there is a difference in mean birth weight among the six categories of babies classified by the combination of their maturity level and mother???s smoking status. If there is evidence of differences among the six categories of babies, carry out an appropriate analysis to see which differ.

i.

Result of One-Way Anova:

```
##              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
```

p-value < 2e-16. So there is sufficient evidence to reject Null Hypothesis. There is evidence of differences among the six categories of babies.

ii.

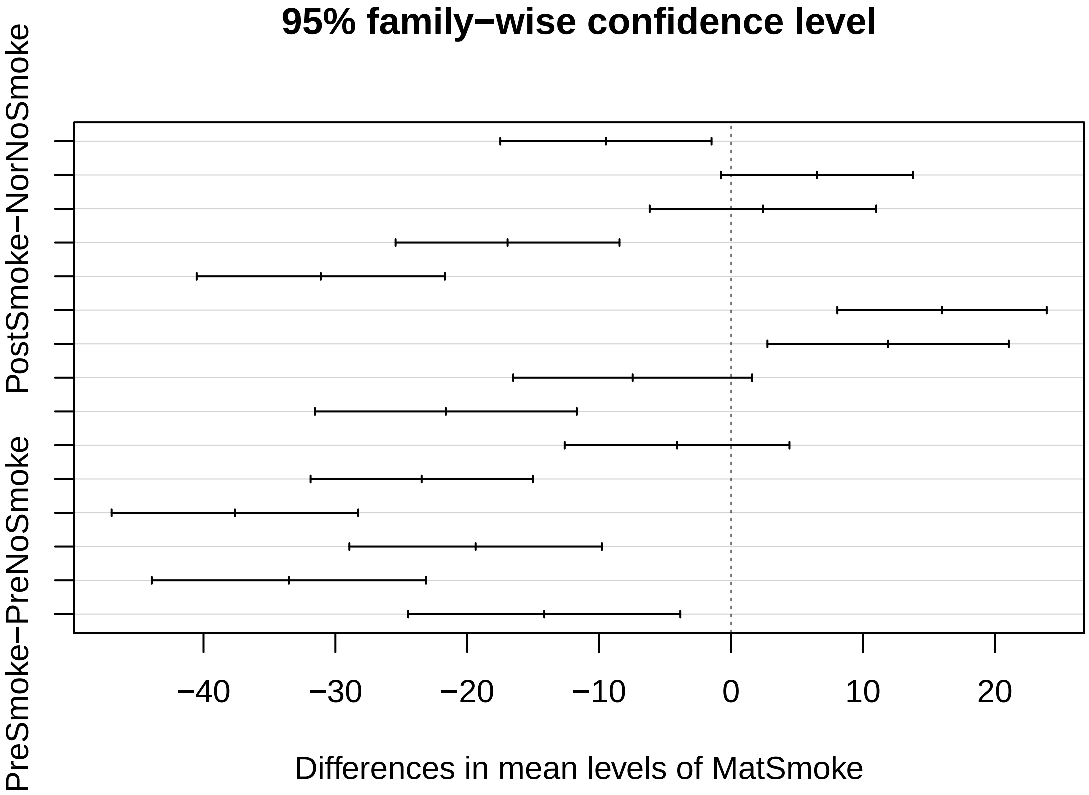
Bonferroni's method

```
## 
##  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
```

Based on Bonferroni's method, mean weight of following paired groups are different: NorNoSmoke and NorSmoke, PostNoSmoke and NorSmoke, PostSmoke and NorSmoke, PreNoSmoke and NorNoSmoke, PreNoSmoke and NorSmoke, PreNoSmoke and PostNoSmoke, PostNoSmoke and PostSmoke, PreNoSmoke and PostSmoke, PreSmoke and others.

Tukey's method

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = bwt ~ MatSmoke)
## 
## $MatSmoke
##                              diff         lwr        upr     p adj
## NorSmoke-NorNoSmoke     -9.485769 -17.4942723  -1.477265 0.0098669
## PostNoSmoke-NorNoSmoke   6.511806  -0.7724273  13.796039 0.1097883
## PostSmoke-NorNoSmoke     2.420550  -6.1667120  11.007811 0.9661428
## PreNoSmoke-NorNoSmoke  -16.945853 -25.4355463  -8.456159 0.0000003
## PreSmoke-NorNoSmoke    -31.107002 -40.5162170 -21.697788 0.0000000
## PostNoSmoke-NorSmoke    15.997574   8.0591186  23.936030 0.0000002
## PostSmoke-NorSmoke      11.906318   2.7575429  21.055093 0.0030202
## PreNoSmoke-NorSmoke     -7.460084 -16.5173423   1.597174 0.1736034
## PreSmoke-NorSmoke      -21.621234 -31.5455649 -11.696903 0.0000000
## PostSmoke-PostNoSmoke   -4.091256 -12.6132282   4.430716 0.7422244
## PreNoSmoke-PostNoSmoke -23.457658 -31.8813064 -15.034010 0.0000000
## PreSmoke-PostNoSmoke   -37.618808 -46.9684748 -28.269142 0.0000000
## PreNoSmoke-PostSmoke   -19.366402 -28.9392207  -9.793584 0.0000002
## PreSmoke-PostSmoke     -33.527552 -43.9245359 -23.130568 0.0000000
## PreSmoke-PreNoSmoke    -14.161150 -24.4776954  -3.844604 0.0013893
```



Based on TukeyHSD method, except for PostNoSmoke-NorNoSmoke, PostSmoke-NorNoSmoke, PreNoSmoke-NorSmoke, PostSmoke-PostNoSmoke, all other pairs are all different in term of mean weight.

5. (10 marks) Do you trust the results of the statistical tests carried out in question 4? Assess whether the necessary assumptions of the model hold.

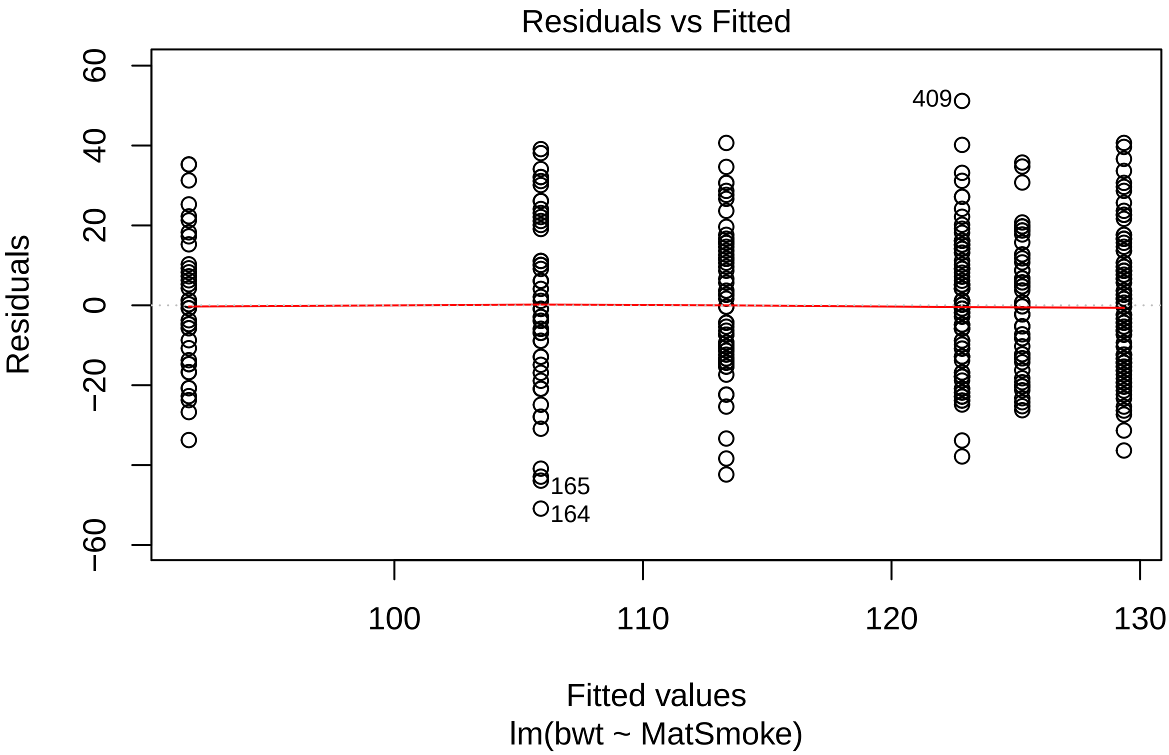
i. Homoscedasticity

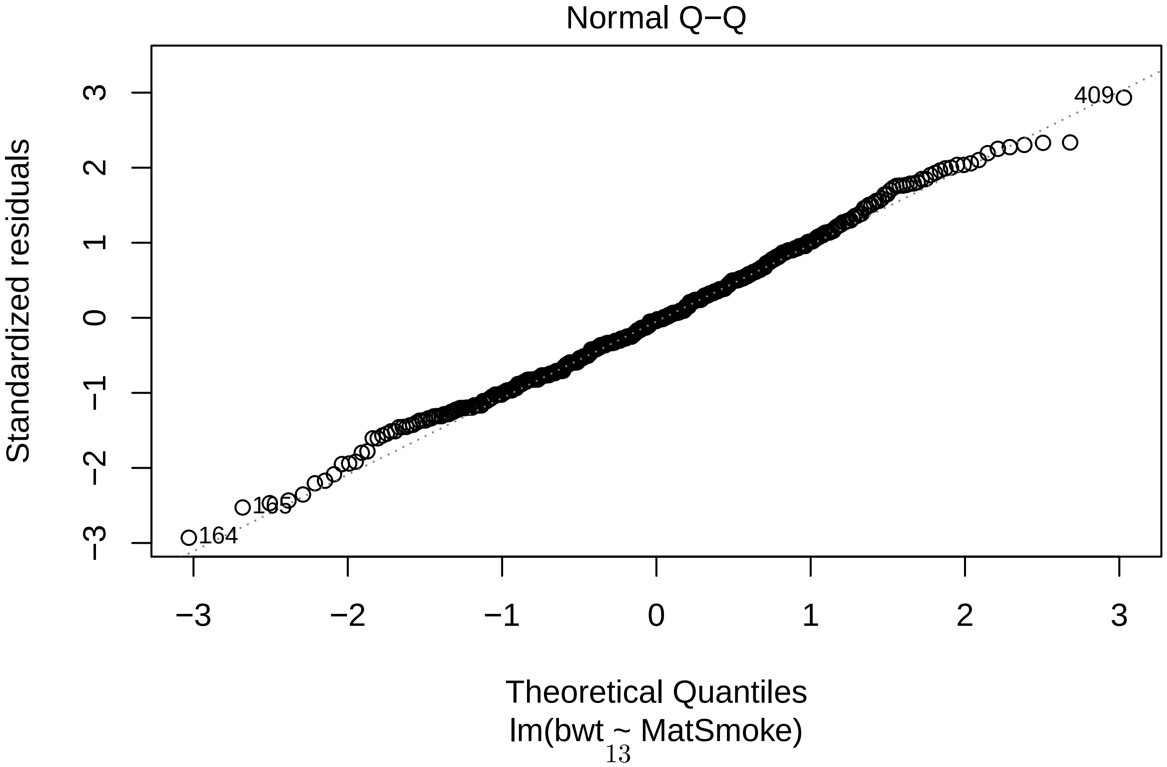
```
##
##  Bartlett test of homogeneity of variances
##
```

```
## data:  bwt by MatSmoke
## Bartlett's K-squared = 9.3393, df = 5, p-value = 0.09627
```

P-value of bartlett test = $0.097 > 0.05$. Do not reject Null Hypothesis. Thus those distributions for each of the groups have the same standard deviation (homogeneity of variances).

ii. Normality





The normality holds. And there are some unusual observation but not influential value.

iii Uncorrelated errors: This is satisfied if sample are chosen independently.

6. (10 marks) Instead of the one-way classification model used in question 4, a two-way analysis of variance model could have been used with maternal smoking status, maturity level and their interaction. WITHOUT fitting this model, answer the following questions.

(a) Would the number of predictor variables be the same as in the model used in question 4? Why or why not?

Yes. In question 4, the number of predictor variables is 5. In two-way Anova model, $thenumberofpredictorvariable = 1 + 2 + 1 \times 2 = 5$

(b) Would the F-test for the presence of interaction between maturity level and smoking status be statistically significant? How do you know from your results of question 4?

Yes, the presence of interaction between maturity level and smoking status would be statistically significant. In question 4, we have already known there is a difference in mean birth weight among the six categories of babies classified by the combination of their maturity level and mother's smoking status.

7. (5 marks) Should we be concerned that the data contained different numbers of babies in the three maturity levels? Why or why not?

Yes. If the numbers of babies in some levels are obvious small, it may influence constant variance test. Thus, we may have a problem. Consider all inferences as only approximate. Besides, if the group size of the three maturity levels are same, the result of bartlett test is more reliable.

8. (5 marks) Discuss the use of gestation as a quantitative explanatory variable rather than as a factor in an additive linear model for mean birth weight. Include mathematical equations to describe the difference in models for mean birth weight.

Additive linear model when gestation as quantitative explanatory variable:

```
##
## Call:
## lm(formula = bwt ~ gestation + factor(smoke))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.605 -12.700  -0.659  12.293  51.278
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.76348    9.83685   0.688    0.492
## gestation       0.40831    0.03476  11.746  < 2e-16 ***
## factor(smoke)1 -8.50620    1.78856  -4.756 2.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 17.68 on 406 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.2885
## F-statistic: 83.71 on 2 and 406 DF,  p-value: < 2.2e-16
```

Additive linear model when gestation as a factor

```
##
## Call:
## lm(formula = bwt ~ factor(maturity) + factor(smoke))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.586 -12.586  -0.805  12.000  51.492
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        103.586      1.938  53.455  < 2e-16 ***
## factor(maturity)2   18.922      2.261   8.370 9.45e-16 ***
## factor(maturity)3   27.414      2.292  11.963  < 2e-16 ***
## factor(smoke)1      -8.703      1.780  -4.889 1.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.59 on 405 degrees of freedom
## Multiple R-squared:  0.3011, Adjusted R-squared:  0.2959
## F-statistic: 58.17 on 3 and 405 DF,  p-value: < 2.2e-16
```

Discussion:

Gestation as quantitative explanatory variable, $bwt = 6.76348 + 0.40831 \times gestation - 8.50620 \times I_{smoke,i}$ where $I_{smoke,i}$ is equal to 1 if the ith baby's mother smoked and 0 if she did not smoke. It means if gestation increses by 1, the mean of birth weight will increase by 0.40831 while others hold constant.

gestation as a factor, $bwt = 103.586 + 18.922 \times I_{maturity2,i} + 27.414 \times I_{maturity3,i} - 8.703 \times I_{smoke,i}$, where $I_{smoke,i}$ is equal to 1 if the ith baby's mother smoked and 0 if she did not smoke, $I_{maturity2,i}$ is equal to 1 if the ith baby spent 259-293 days in the womb and 0 if not, and $I_{maturity3,i}$ is equal to 1 if the ith baby spent more than 293 days in the womb and 0 if not. It means that compared to mean birth weight of maturity level 1 group, mean birth weight from maturity level 2 group is more 18.922 and the mean birth weight from maturity level 3 group is more 27.414 while we hold others constant.

In summary, I prefer that we take gestation as factor(two-way anova model) since it is more intuitive and straightforward.

9. (5 marks) Name two additional potential factors of baby birth weight and briefly describe their levels.

Sex: the baby's gender, male or female

Vegetarian: an indicator variable which is 1 if the baby's mother was vegetarian and 0 if she was not vegetarian.

Appendix

1. (15 marks) Create two new variables: (1) maturity- by converting gestational age to a factor with 3 levels; 1 if the baby was preterm and spent less than 259 days in the womb, 3 if gestational age was beyond 293 and 2 otherwise, and (2) MatSmoke- a variable that combines maturity level and maternal smoking status.Construct three sets of side-by-side boxplots: 1. to compare birth weight between mothers who smoked and those who did not smoke during pregnancy, 2. to compare birth weight among the three maturity levels, and 3. to compare birth weight among the 6 categories of babies grouped by the combination of their maturity level and maternal smoking status. Do there appear to be any differences?

```
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"}
  }
```

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

```
GroupS <- bwt[smoke == 1]
GroupNS <- bwt[smoke == 0]

 boxplot(GroupS, GroupNS,xlab="Mother's smoking status", names=c("Smoke","NonSmoke"),range=0)
 title("Ruijie Sun 6046")
```

side-by-side boxplot to compare birth weight among the three maturity levels.

```
GroupPre <- bwt[maturity == 1]
GroupNor <- bwt[maturity == 2]
GroupPost <- bwt[maturity == 3]

 boxplot(GroupPre, GroupNor,GroupPost, xlab="Maturity Levels", names=c("Pre","Nor","Post"),range=0)
 title("Ruijie Sun 6046")
```

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

```
GroupPreSmoke <- bwt[MatSmoke == "PreSmoke" ]
GroupPreNonSmoke <- bwt[MatSmoke == "PreNoSmoke"]
GroupNorSmoke <- bwt[MatSmoke == "NorSmoke"]
GroupNorNonSmoke <-bwt[MatSmoke == "NorNoSmoke"]
GroupPostSmoke <-bwt[MatSmoke == "PostSmoke"]
GroupPostNonSmoke <-bwt[MatSmoke == "PostNoSmoke"]

 boxplot(GroupPreSmoke, GroupPreNonSmoke, GroupNorSmoke,  GroupNorNonSmoke, GroupPostSmoke, GroupPostNo
 title("Ruijie Sun 6046")
```

2. (10 marks) Using the R t.test procedure, investigate whether or not there is a difference in the mean birth weight between babies born to mothers who were smokers and babies born to mothers who were nonsmokers.

i. Side-by-side boxplots

```
GroupS <- bwt[smoke == 1]
GroupNS <- bwt[smoke == 0]

 boxplot(GroupS, GroupNS,xlab="Mother's smoking status", names=c("Smoke","NonSmoke"),range=0)
 title("Ruijie Sun 6046")
```

ii.Null and Alternative Hypothesis

iii. A test statistic and it's distribution

To test if the two sample have same variance

```
var.test(GroupS,GroupNS)
```

```
##
##  F test to compare two variances
##
## data:  GroupS and GroupNS
## 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
```

iv. Test assumptions

v. Test diagnostics

```
par(mfrow=c(1,2))
hist(GroupS,main="Weight Histogram from GroupS",xlab="GroupS \n Ruijie Sun 6046")
qqnorm(GroupS, xlab="Theoretical Quantiles \n Ruijie Sun 6046")
qqline(GroupS)
```

```
par(mfrow=c(1,2))
hist(GroupNS,main="WeightHistogram from GroupNS",xlab="GroupS \n Ruijie Sun 6046")
qqnorm(GroupNS, xlab="Theoretical Quantiles \n Ruijie Sun 6046")
qqline(GroupNS)
```

vi P-value

```
t.test(GroupS,GroupNS,var.equal = T)
```

vii Results:

3. (15 marks) Investigate whether or not there is a difference in mean birth weight among babies classified by gestational maturity, using a one-way analysis of variance. If there is a difference among the levels of maturity, carry out an appropriate analysis to see which levels of maturity differ.

i.

Result of One-Way Anova:

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

ii.

Bonferroni's method

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

Tukey's method

```
fm <- factor(maturity)
amod=aov(bwt~fm)
TukeyHSD(amod,"fm")
```

```r
plot(TukeyHSD(amod,"fm"))
```

4. (15 marks) Use one-way analysis of variance to investigate whether or not there is a difference in mean birth weight among the six categories of babies classified by the combination of their maturity level and mother???s smoking status. If there is evidence of differences among the six categories of babies, carry out an appropriate analysis to see which differ.

i.

Result of One-Way Anova:

```r
summary(aov(bwt~MatSmoke))
```

ii.

Bonferroni's method

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

Tukey's method

```r
amod=aov(bwt~MatSmoke)
TukeyHSD(amod,"MatSmoke")
```

```r
plot(TukeyHSD(amod,"MatSmoke"))
```

5. (10 marks) Do you trust the results of the statistical tests carried out in question 4? Assess whether the necessary assumptions of the model hold.

i. Homoscedasticity

```r
bartlett.test(bwt~MatSmoke)
```

ii. Normality

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

```r
plot(lm(bwt~MatSmoke),which=2,xlab="Ruijie Sun 6046")
```

iii Uncorrelated errors:

6. (10 marks) Instead of the one-way classification model used in question 4, a two-way analysis of variance model could have been used with maternal smoking status, maturity level and their interaction. WITHOUT fitting this model, answer the following questions.

(a) Would the number of predictor variables be the same as in the model used in question 4? Why or why not?

(b) Would the F-test for the presence of interaction between maturity level and smoking status be statistically significant? How do you know from your results of question 4?

7. (5 marks) Should we be concerned that the data contained different numbers of babies in the three maturity levels? Why or why not?

8. (5 marks) Discuss the use of gestation as a quantitative explanatory variable rather than as a factor in an additive linear model for mean birth weight. Include mathematical equations to describe the difference in models for mean birth weight.

Additive linear model when gestation as quantitative explanatory variable:

```
mod1 <- lm(formula = bwt~ gestation + factor(smoke))
summary(mod1)
```

Additive linear model when gestation as a factor

```
mod2 <- lm(formula = bwt ~ factor(maturity) + factor(smoke))
summary(mod2)
```

9. (5 marks) Name two additional potential factors of baby birth weight and briefly describe their levels.