STA303 Assignment 1

Haoda Li 1003918335

Solutions

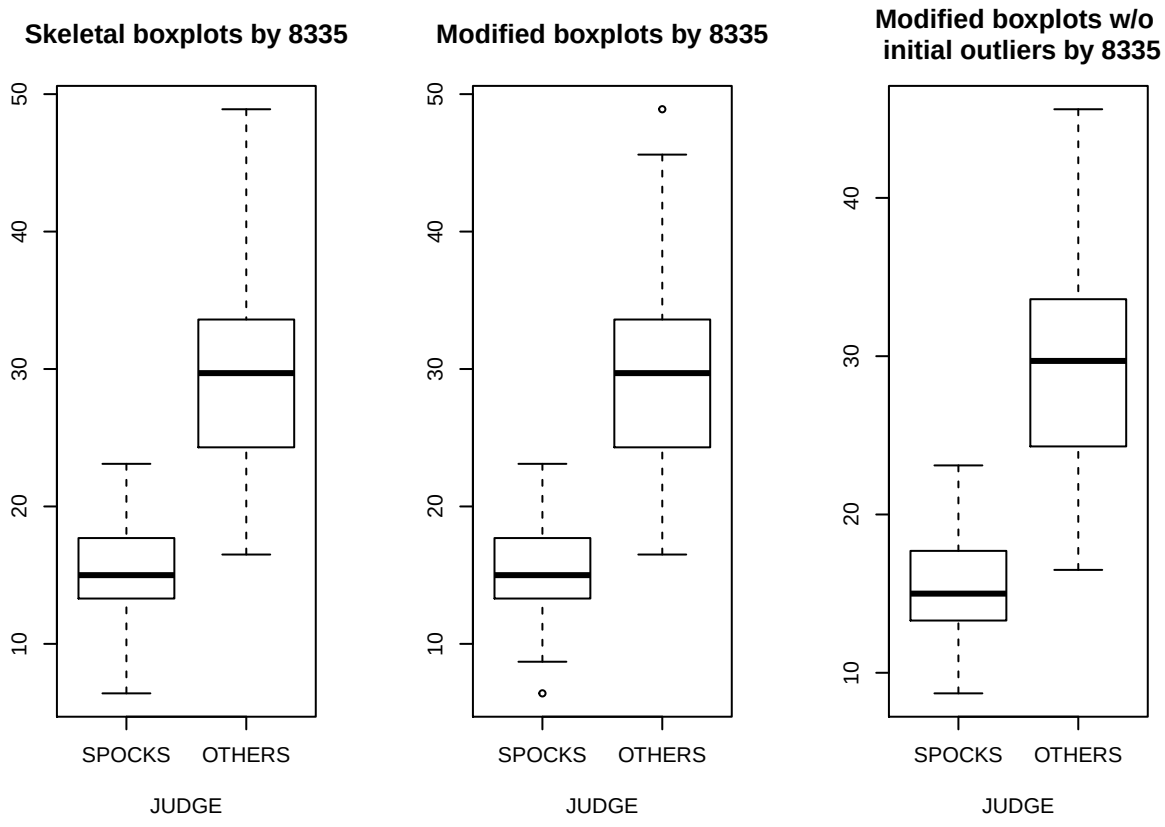
Question 1

(a) The code is the following (see Appendix for importing data and grouping data):

```
par(mfrow=c(1,3))
boxplot(groupS, groupNS, range=0, xlab='JUDGE', names=c('SPOCKS', 'OTHERS'),
        main='Skeletal boxplots by 8335')

boxplot(groupS, groupNS, xlab='JUDGE', names=c("SPOCKS","OTHERS"),
        main="Modified boxplots by 8335")

boxplot(groupS, groupNS, outline=F, xlab='JUDGE', names=c("SPOCKS","OTHERS"),
        main="Modified boxplots w/o \n initial outliers by 8335")
```



(b) The values are:

Group SPOKES in Skeletal boxplots
The first quartile is 13.3
The second quartile is 15.0

The third quartile is 17.7
The end points of the two whiskers are 6.4 (lower) 23.1 (upper)
No outliers is shown on the plot

Group OTHERS in Skeletal boxplots
The first quartile is 24.3
The second quartile is 29.7
The third quartile is 33.6
The end points of the two whiskers are 16.5 (lower) 48.9 (upper)
No outliers is shown on the plot

Group SPOKES in Modified boxplots
The first quartile is 13.3
The second quartile is 15.0
The third quartile is 17.7
The end points of the two whiskers are 8.7 (lower) 23.1 (upper)
There exists one outlier at 6.4

Group OTHERS in Modified boxplots
The first quartile is 24.3
The second quartile is 29.7
The third quartile is 33.6
The end points of the two whiskers are 16.5 (lower) 45.6 (upper)
There exists one outlier at 48.9

Group SPOKES in Modified boxplots w/o initial outliers
The first quartile is 13.3
The second quartile is 15.0
The third quartile is 17.7
The end points of the two whiskers are 8.7 (lower) 23.1 (upper)
No outliers is shown on the plot

Group OTHERS in Modified boxplots w/o initial outliers
The first quartile is 24.3
The second quartile is 29.7
The third quartile is 33.6
The end points of the two whiskers are 16.5 (lower) 45.6 (upper)
No outliers is shown on the plot

(c) The Modified boxplots (Pair #2) best represents the data because it gives the information about the potential outliers and it indicates that the two sample groups potentially have different means.

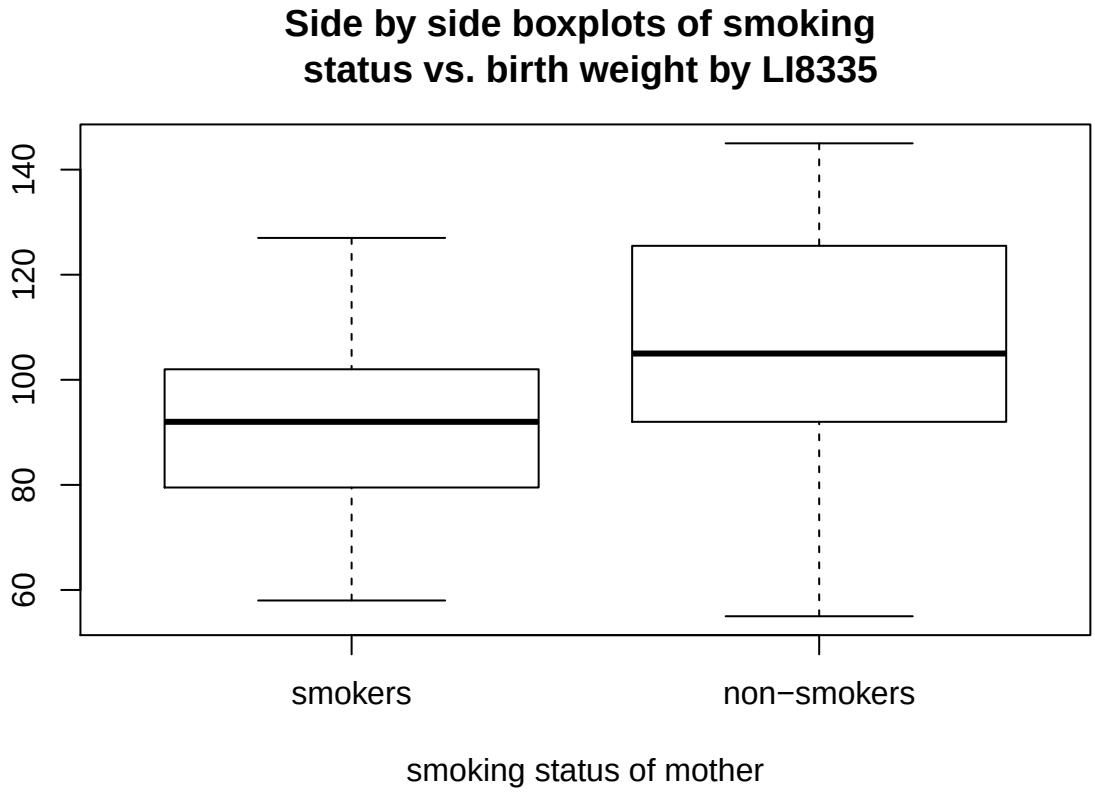
Question 2

(see the source code and outputs in the Appendix)

(a) The categorical value is smoke (smoking status of mother). The levels are 0 and 1, where 0 is coded for non-smoking mother and 1 is coded for smoking mother.

(b)

i. side by side boxplots



From the side by side boxplots, we can observe some evidence that there is a difference of the mean birth weight between two groups.

ii. Null and alternative hypotheses
Null hypothesis: There is no difference of the mean birth weight between smokers and non-smokers.
Alternative hypothesis: There is a difference of the mean birth weight between smokers and non-smokers.

iii. Test statistic and distribution
I will fit the data into the simple linear model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $X_i = 1$ if ith observation if from "smokers" and $x = 0$ otherwise.
Then, the hypothesis test becomes $H_0 : \beta_1 = 0 \, vs \, H_a : \beta_1 \neq 0$.
The distribution is $t = \frac{b_1}{se(b_1)} \sim t_{n-2}$ where n is the total number of observations, b_1 is the MSE for β_1.
In this case, the distribution is $t \sim t_{97}$.
The test statistic is

$$t = \frac{-13.451}{4.073} = -3.302$$

(see Appendix for model summary).

iv. Test assumptions
The linear model is appropriate.
$E(\epsilon_i) = 0$.
$var(\epsilon) = \sigma^2$, the random errors are uncorrelated.
The errors follow a normal distribution.

v. Test diagnostics
(see Appendix for plots referred below)
The linear model is appropriate and we are doing a dummy variable test.
According to the histogram of residuals, Residuals plot, and residuals vs fitted plots, we can observe that

the errors are spread evenly across 0, hence the assumption that $E(\epsilon) = 0$ is not violated. According to the Residuals plot, the residuals spread evenly and has no particular pattern. The assumption that errors are uncorrelated is not violated. According to the Normal Q-Q plot, only a few points are off the qqline. The assumptions that the errors follows the normal distribution is not violated.

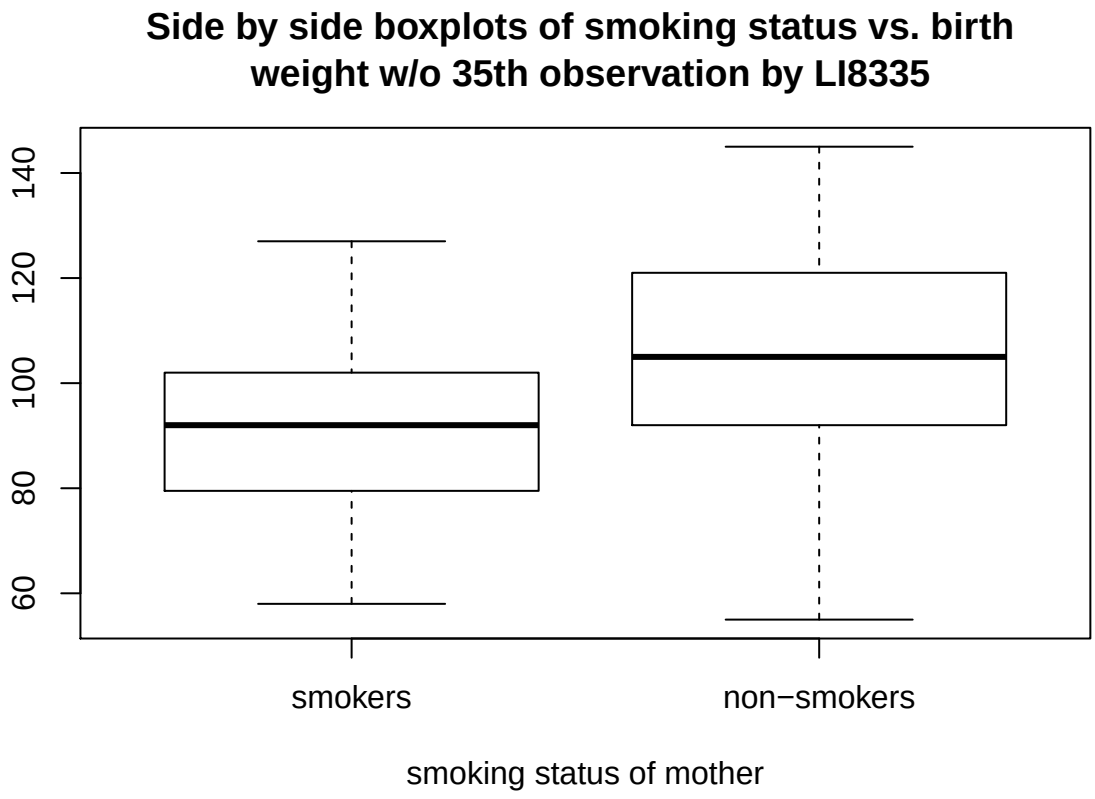
vi. P-value
Given in the summary of the model (see Appendix), the P-value is 0.00134

vii. Results
Since the P-value $= 0.00134 < 0.05$, we can reject our null hypothesis. Based on our diagnostics, the assumptions are not violated.
Therefore, we can make the conclusion that there are some evidence suggesting that 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.

(c) We can also do a **two-sample t-test** or an **one-way ANOVA**

(d) I'll remove the observation with id. 35

i. side by side boxplots



From the side by side boxplots, we can observe some evidence that there is a difference of the mean birth weight between two groups.

ii. Null and alternative hypotheses
Null hypothesis: There is no difference of the mean birth weight between smokers and non-smokers.
Alternative hypothesis: There is a difference of the mean birth weight between smokers and non-smokers.

iii. Test statistic and distribution
I will fit the data into the simple linear model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $X_i = 1$ if ith observation if

from "smokers" and $x = 0$ otherwise.
Then, the hypothesis test becomes $H_0 : \beta_1 = 0\ vs\ H_a : \beta_1 \neq 0$.
The distribution is $t = \frac{b_1}{se(b_1)} \sim t_{n-2}$ where n is the total number of observations, b_1 is the MSE for β_1.
In this case, the distribution is $t \sim t_{96}$.
The test statistic is

$$t = \frac{-12.976}{4.074} = -3.185$$

(see Appendix for model summary).

iv. Test assumptions
The linear model is appropriate.
$E(\epsilon_i) = 0$.
$var(\epsilon) = \sigma^2$, the random errors are uncorrelated.
The errors follow a normal distribution.

v. Test diagnostics
(see Appendix for plots referred below)
The linear model is appropriate and we are doing a dummy variable test.
According to the histogram of residuals, Residuals plot, and residuals vs fitted plots, we can observe that the errors are spread evenly across 0, hence the assumption that $E(\epsilon) = 0$ is not violated. According to the Residuals plot, the residuals spread evenly and has no particular pattern. The assumption that errors are uncorrelated is not violated. According to the Normal Q-Q plot, only a few points are off the qqline. The assumptions that the errors follows the normal distribution is not violated.

vi. P-value
Given in the summary of the model (see Appendix), the P-value is 0.00195

vii. Results
Since the P-value = 0.00195 < 0.05, we can reject our null hypothesis. Based on our diagnostics, the assumptions are not violated.
Therefore, we can make the conclusion that there are some evidence suggesting that 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.

(e) According to the test statistics (-3.302 vs. -3.185) and P-values (0.00134 vs. 0.00195) of the two tests, the difference is not significant. Therefore, The removed observation (id. 35) is not influential.

Appendix

Question 1

```
# import data
juries <- read.csv('juries.csv')
attach(juries)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     JUDGE, PERCENT
```

```
# put data into two group
groupS<-PERCENT[JUDGE=="SPOCKS"]
groupNS<-PERCENT[JUDGE!="SPOCKS"]
```

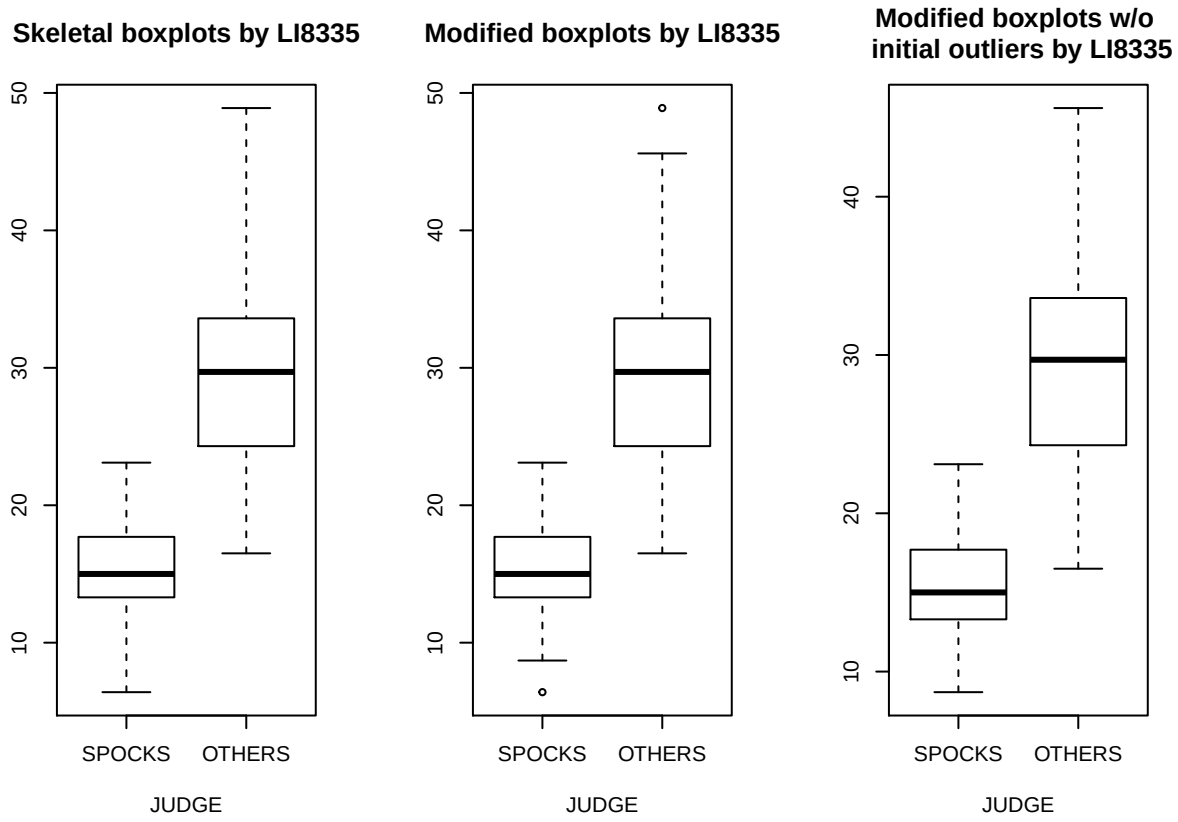
Code and output for (a)

```
par(mfrow=c(1,3))
```

```
# skeletal
splot <- boxplot(groupS, groupNS, range=0, xlab='JUDGE', names=c('SPOCKS', 'OTHERS'),
        main='Skeletal boxplots by LI8335')

# modified
mplot <- boxplot(groupS, groupNS, xlab='JUDGE', names=c("SPOCKS","OTHERS"),
        main="Modified boxplots by LI8335")

# modified without outliers
nplot <- boxplot(groupS, groupNS, outline=F, xlab='JUDGE', names=c("SPOCKS","OTHERS"),
        main="Modified boxplots w/o \n initial outliers by LI8335")
```



Code and output for (b)

```
# for quartiles, min/max, and outliers
splot$stats
```

```
##      [,1] [,2]
## [1,]  6.4 16.5
## [2,] 13.3 24.3
## [3,] 15.0 29.7
## [4,] 17.7 33.6
## [5,] 23.1 48.9
```

```
splot$out
```

```
## numeric(0)
```

```
mplot$stats
```

```
##      [,1] [,2]
## [1,]  8.7 16.5
## [2,] 13.3 24.3
## [3,] 15.0 29.7
## [4,] 17.7 33.6
## [5,] 23.1 45.6
```

```
mplot$out
```

```
## [1]  6.4 48.9
```

```
nplot$stats
```

```
##      [,1] [,2]
## [1,]  8.7 16.5
## [2,] 13.3 24.3
## [3,] 15.0 29.7
## [4,] 17.7 33.6
## [5,] 23.1 45.6
```

Question 2

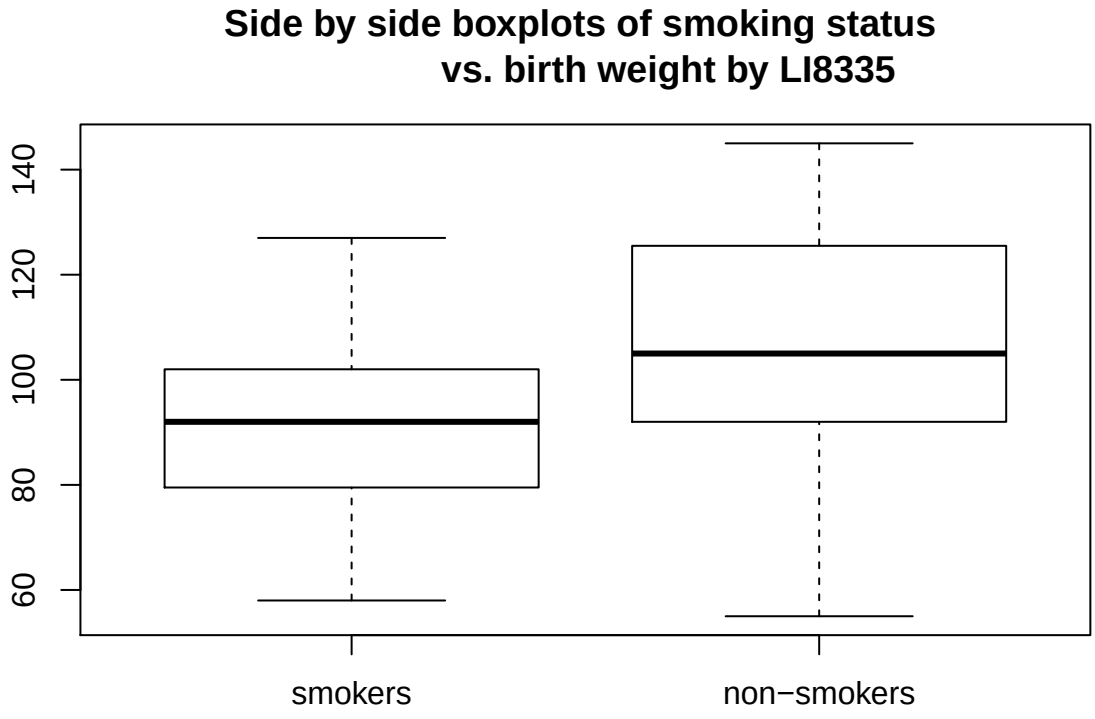
```
# import data
birth = read.csv("bbw99.csv")
attach(birth)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     bwt, smoke
```

```
# group data
smoker = bwt[smoke == 1]
nonsmoker = bwt[smoke == 0]
```

Code for (b) i.

```
# side by side boxplots
splot = boxplot(smoker, nonsmoker, xlab='smoking status of mother',
                names=c('smokers', 'non-smokers'),
                main='Side by side boxplots of smoking status
                vs. birth weight by LI8335')
```




```
summary(smoker)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   58.00   79.50   92.00   92.44  102.00  127.00
```

```
length(smoker)
```

```
## [1] 43
```

```
summary(nonsmoker)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0    92.5   105.0   105.9   125.2   145.0
```

```
length(nonsmoker)
```

```
## [1] 56
```

Code for (b) iii. vi.

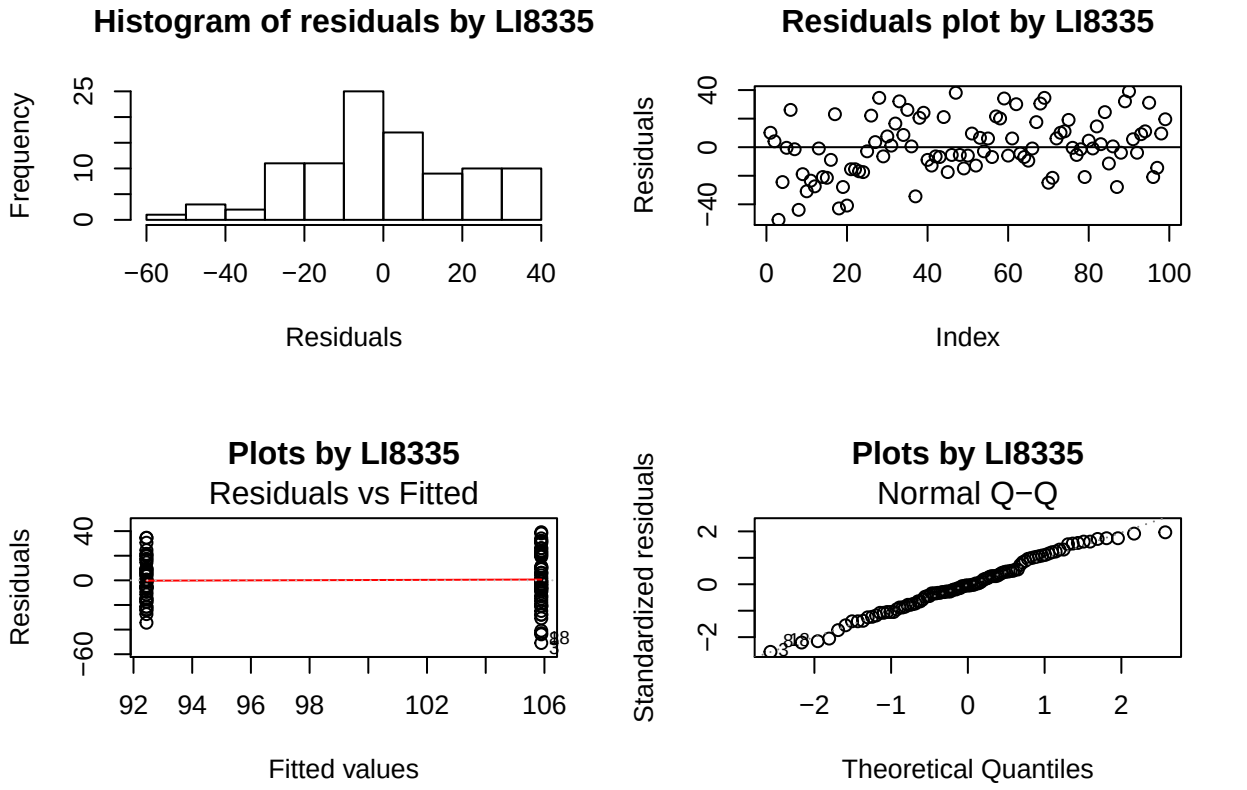
```
model <- lm(bwt~smoke)
summary(model)
```

```
##
## Call:
## lm(formula = bwt ~ smoke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.893 -13.667  -0.893  12.833  39.107
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  105.893      2.684  39.448  < 2e-16 ***
## smoke        -13.451      4.073  -3.302  0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.09 on 97 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.0918
## F-statistic: 10.91 on 1 and 97 DF,  p-value: 0.001343
```

Code for (b) v.

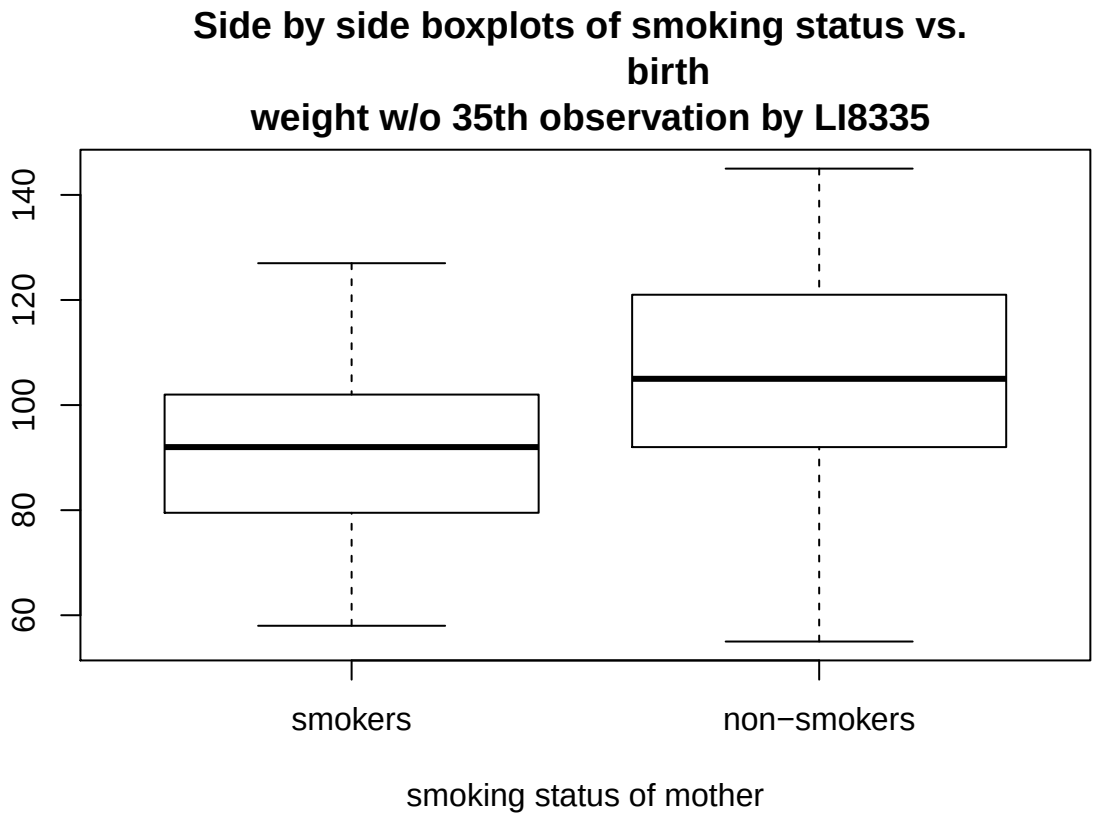
```
par(mfrow=c(2,2))
hist(residuals(model),xlab="Residuals", main="Histogram of residuals by LI8335")
plot(residuals(model),ylab="Residuals", main="Residuals plot by LI8335")
abline(0,0)
plot(model, which=1:2, main = "Plots by LI8335")
```



Code for (d) i.

```
birth2 <- birth[-35,]
smoke2 <- birth2$smoke
bwt2 <- birth2$bwt
smoker2 = bwt2[smoke2 == 1]
nonsmoker2 = bwt2[smoke2 == 0]
splot2 = boxplot(smoker2, nonsmoker2, xlab='smoking status of mother',
```

```
                names=c('smokers', 'non-smokers'),
                main='Side by side boxplots of smoking status vs.
                birth \n weight w/o 35th observation by LI8335')
```




```
summary(smoker2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   58.00   79.50   92.00   92.44  102.00  127.00
```

```
length(smoker2)
```

```
## [1] 43
```

```
summary(nonsmoker2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0    92.0   105.0   105.4   121.0   145.0
```

```
length(nonsmoker2)
```

```
## [1] 55
```

Code for (d) iii. vi.

```
model2 <- lm(bwt2~smoke2)
summary(model2)
```

```
##
## Call:
## lm(formula = bwt2 ~ smoke2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.418 -13.918  -0.442  11.582  39.582
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  105.418      2.698  39.066  < 2e-16 ***
## smoke2       -12.976      4.074  -3.185  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.01 on 96 degrees of freedom
## Multiple R-squared:  0.09559,    Adjusted R-squared:  0.08617
## F-statistic: 10.15 on 1 and 96 DF,  p-value: 0.001951
```

Code for (d) v.

```
par(mfrow=c(2,2))
hist(residuals(model2),xlab="Residuals",
     main="Histogram of residuals \n w/o 35th obeservation by LI8335")
plot(residuals(model2),ylab="Residuals",
     main="Residuals plot w/o \n 35th obeservation by LI8335")
abline(0,0)
plot(model2, which=1:2, main = "Plots w/o 35th \n obeservation by LI8335")
```