

Homework1

Charles Kekeh

Sunday, January 31, 2016

```
load.data <- function(){  
  load(file.path("./data", "birthweight_w271.rdata"))  
  return (data)  
}
```

Question 1

```
# Load the dataframe  
data <- load.data()
```

Question 2

```
# Display summary information about the dataframe  
print(str(data))
```

```
## 'data.frame':    1388 obs. of  14 variables:  
## $ faminc   : num  13.5 7.5 0.5 15.5 27.5 7.5 65 27.5 27.5 37.5 ...  
## $ cigtax   : num  16.5 16.5 16.5 16.5 16.5 16.5 16.5 16.5 16.5 16.5 ...  
## $ cigprice: num  122 122 122 122 122 ...  
## $ bwght    : num  109 133 129 126 134 118 140 86 121 129 ...  
## $ fatheduc: int   12 6 NA 12 14 12 16 12 12 16 ...  
## $ motheduc: int   12 12 12 12 12 14 14 14 17 18 ...  
## $ parity   : int   1 2 2 2 2 6 2 2 2 2 ...  
## $ male     : int   1 1 0 1 1 1 0 0 0 0 ...  
## $ white    : int   1 0 0 0 1 0 1 0 1 1 ...  
## $ cigs     : int   0 0 0 0 0 0 0 0 0 0 ...  
## $ lbwght   : num   4.69 4.89 4.86 4.84 4.9 ...  
## $ bwghtlbs: num   6.81 8.31 8.06 7.88 8.38 ...  
## $ packs    : num   0 0 0 0 0 0 0 0 0 0 ...  
## $ lfaminc  : num   2.603 2.015 -0.693 2.741 3.314 ...  
## - attr(*, "datalabel")= chr ""  
## - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"  
## - attr(*, "formats")= chr  "%9.0g" "%9.0g" "%9.0g" "%8.0g" ...  
## - attr(*, "types")= int   254 254 254 252 251 251 251 251 251 251 ...  
## - attr(*, "val.labels")= chr  "" "" "" "" ...  
## - attr(*, "var.labels")= chr  "1988 family income, $1000s" "cig. tax in home state, 1988" "cig. pri  
## - attr(*, "version")= int 10  
## NULL
```

```
print(summary(data))
```

```
##          faminc          cigtax          cigprice          bwght
## Min.   : 0.50   Min.   : 2.00   Min.   :103.8   Min.   : 0.0
## 1st Qu.:14.50   1st Qu.:15.00   1st Qu.:122.8   1st Qu.:106.0
## Median :27.50   Median :20.00   Median :130.8   Median :119.0
## Mean   :29.03   Mean   :19.55   Mean   :130.6   Mean   :117.9
## 3rd Qu.:37.50   3rd Qu.:26.00   3rd Qu.:137.0   3rd Qu.:132.0
## Max.   :65.00   Max.   :38.00   Max.   :152.5   Max.   :271.0
##
##          fatheduc          motheduc          parity          male
## Min.   : 1.00   Min.   : 2.00   Min.   :1.000   Min.   :0.0000
## 1st Qu.:12.00   1st Qu.:12.00   1st Qu.:1.000   1st Qu.:0.0000
## Median :12.00   Median :12.00   Median :1.000   Median :1.0000
## Mean   :13.19   Mean   :12.94   Mean   :1.633   Mean   :0.5209
## 3rd Qu.:16.00   3rd Qu.:14.00   3rd Qu.:2.000   3rd Qu.:1.0000
## Max.   :18.00   Max.   :18.00   Max.   :6.000   Max.   :1.0000
## NA's   :196    NA's   :1
##          white          cigs          lbwght          bwghtlbs
## Min.   :0.0000   Min.   : 0.000   Min.   :0.000   Min.   : 0.000
## 1st Qu.:1.0000   1st Qu.: 0.000   1st Qu.:4.663   1st Qu.: 6.625
## Median :1.0000   Median : 0.000   Median :4.779   Median : 7.438
## Mean   :0.7846   Mean   : 2.087   Mean   :4.726   Mean   : 7.366
## 3rd Qu.:1.0000   3rd Qu.: 0.000   3rd Qu.:4.883   3rd Qu.: 8.250
## Max.   :1.0000   Max.   :50.000   Max.   :5.602   Max.   :16.938
##
##          packs          lfaminc
## Min.   :0.0000   Min.   : -0.6931
## 1st Qu.:0.0000   1st Qu.: 2.6741
## Median :0.0000   Median : 3.3142
## Mean   :0.1044   Mean   : 3.0713
## 3rd Qu.:0.0000   3rd Qu.: 3.6243
## Max.   :2.5000   Max.   : 4.1744
##
```

```
print(stat.desc(data))
```

```
##          faminc          cigtax          cigprice          bwght
## nbr.val      1.388000e+03 1.388000e+03 1.388000e+03 1.388000e+03
## nbr.null      0.000000e+00 0.000000e+00 0.000000e+00 1.000000e+01
## nbr.na        0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## min          5.000000e-01 2.000000e+00 1.038000e+02 0.000000e+00
## max          6.500000e+01 3.800000e+01 1.525000e+02 2.710000e+02
## range        6.450000e+01 3.600000e+01 4.870000e+01 2.710000e+02
## sum          4.028900e+04 2.713950e+04 1.812159e+05 1.635790e+05
## median       2.750000e+01 2.000000e+01 1.308000e+02 1.190000e+02
## mean         2.902666e+01 1.955295e+01 1.305590e+02 1.178523e+02
## SE.mean      5.029888e-01 2.092448e-01 2.749764e-01 6.085627e-01
## CI.mean.0.95 9.867009e-01 4.104705e-01 5.394145e-01 1.193803e+00
## var          3.511608e+02 6.077135e+01 1.049495e+02 5.140438e+02
## std.dev      1.873928e+01 7.795598e+00 1.024448e+01 2.267253e+01
## coef.var     6.455888e-01 3.986916e-01 7.846632e-02 1.923809e-01
```

```
##          fatheduc      motheduc      parity      male
## nbr.val      1.192000e+03 1.387000e+03 1.388000e+03 1.388000e+03
## nbr.null      0.000000e+00 0.000000e+00 0.000000e+00 6.650000e+02
## nbr.na        1.960000e+02 1.000000e+00 0.000000e+00 0.000000e+00
## min           1.000000e+00 2.000000e+00 1.000000e+00 0.000000e+00
## max           1.800000e+01 1.800000e+01 6.000000e+00 1.000000e+00
## range         1.700000e+01 1.600000e+01 5.000000e+00 1.000000e+00
## sum           1.571800e+04 1.794200e+04 2.266000e+03 7.230000e+02
## median        1.200000e+01 1.200000e+01 1.000000e+00 1.000000e+00
## mean          1.318624e+01 1.293583e+01 1.632565e+00 5.208934e-01
## SE.mean       7.953531e-02 6.381773e-02 2.399695e-02 1.341381e-02
## CI.mean.0.95  1.560449e-01 1.251898e-01 4.707424e-02 2.631355e-02
## var           7.540432e+00 5.648838e+00 7.992848e-01 2.497434e-01
## std.dev       2.745985e+00 2.376728e+00 8.940273e-01 4.997433e-01
## coef.var      2.082462e-01 1.837322e-01 5.476213e-01 9.593966e-01
##          white      cigs      lbwght      bwghtlbs
## nbr.val      1.388000e+03 1388.0000000 1.388000e+03 1.388000e+03
## nbr.null      2.990000e+02 1176.0000000 1.000000e+01 1.000000e+01
## nbr.na        0.000000e+00 0.0000000 0.000000e+00 0.000000e+00
## min           0.000000e+00 0.0000000 0.000000e+00 0.000000e+00
## max           1.000000e+00 50.0000000 5.602119e+00 1.693750e+01
## range         1.000000e+00 50.0000000 5.602119e+00 1.693750e+01
## sum           1.089000e+03 2897.0000000 6.559307e+03 1.022369e+04
## median        1.000000e+00 0.0000000 4.779123e+00 7.437500e+00
## mean          7.845821e-01 2.0871758 4.725725e+00 7.365769e+00
## SE.mean       1.103880e-02 0.1603153 1.195727e-02 3.803517e-02
## CI.mean.0.95  2.165455e-02 0.3144867 2.345628e-02 7.461267e-02
## var           1.691349e-01 35.6730005 1.984510e-01 2.007984e+00
## std.dev       4.112601e-01 5.9726879 4.454784e-01 1.417033e+00
## coef.var      5.241772e-01 2.8616123 9.426668e-02 1.923809e-01
##          packs      lfaminc
## nbr.val      1.388000e+03 1388.00000000
## nbr.null      1.176000e+03 0.00000000
## nbr.na        0.000000e+00 0.00000000
## min           0.000000e+00 -0.69314718
## max           2.500000e+00 4.17438745
## range         2.500000e+00 4.86753464
## sum           1.448500e+02 4262.92435274
## median        0.000000e+00 3.31418610
## mean          1.043588e-01 3.07127115
## SE.mean       8.015767e-03 0.02464214
## CI.mean.0.95  1.572434e-02 0.04833990
## var           8.918250e-02 0.84284246
## std.dev       2.986344e-01 0.91806452
## coef.var      2.861612e+00 0.29892005
```

Question 3

```
# Get summary statistics about the birthweight variable.
print(summary(data$bwght))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##      0.0   106.0   119.0   117.9   132.0   271.0
```

```
print(sum(is.nan(data$bwght)))
```

```
## [1] 0
```

```
print(quantile(data$bwght, probs = c(.1, .5, .10, .25, .50, .75, .90, .95, .99, 1)))
```

```
##      10%      50%      10%      25%      50%      75%      90%      95%      99%     100%
##  93.00 119.00  93.00 106.00 119.00 132.00 143.00 149.00 160.13 271.00
```

```
# Plot the histogram of bwght at 15 bins
```

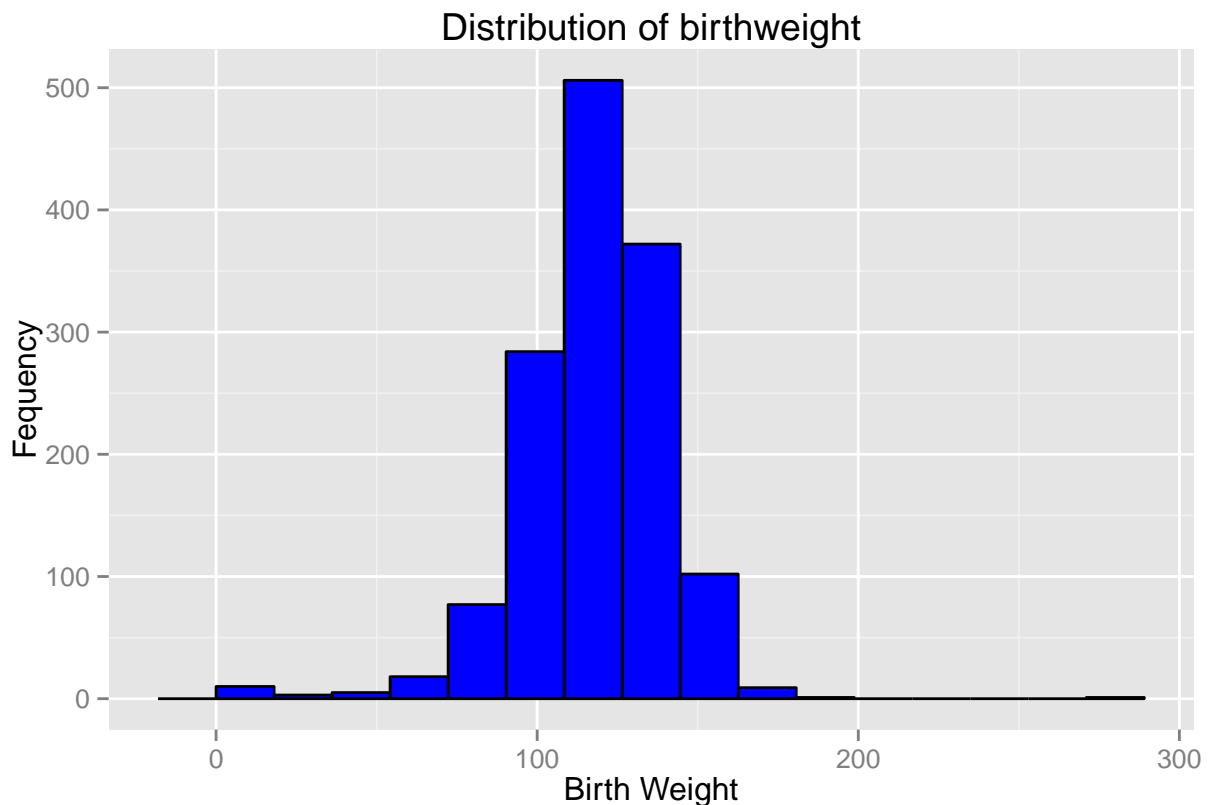
```
bwght.hist <- ggplot(data, aes(bwght)) +
```

```
  theme(legend.position = "none") +
```

```
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$bwght)[2] - range(data$bwght)[1]) / 15)
```

```
  labs(title = "Distribution of birthweight", x = "Birth Weight", y = "Fequency")
```

```
plot(bwght.hist)
```



```
# Plot the histogram of bwght at 30 bins
```

```
bwght.hist <- ggplot(data, aes(bwght)) +
```

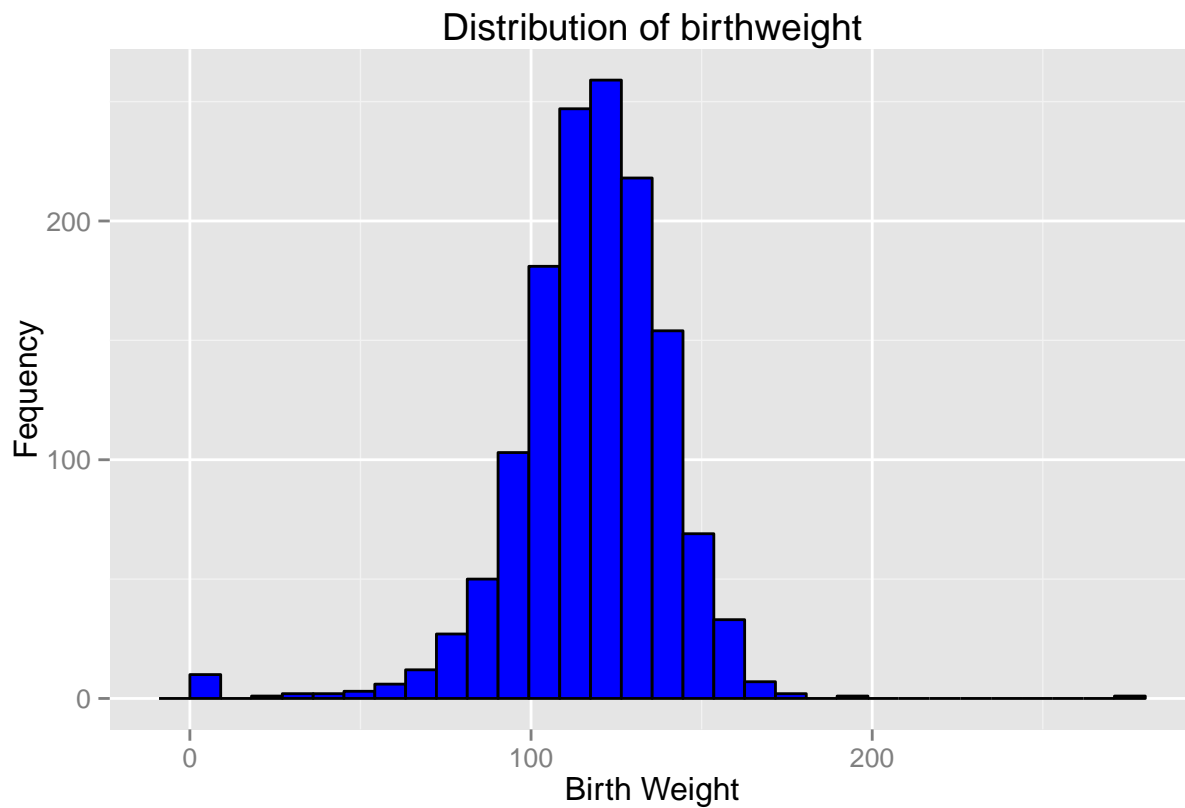
```
  theme(legend.position = "none") +
```

```
  geom_histogram(fill = "Blue", colour = "Black") +
```

```
  labs(title = "Distribution of birthweight", x = "Birth Weight", y = "Fequency")
```

```
plot(bwght.hist)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
# Plot the histogram of bwght at 60 bins
```

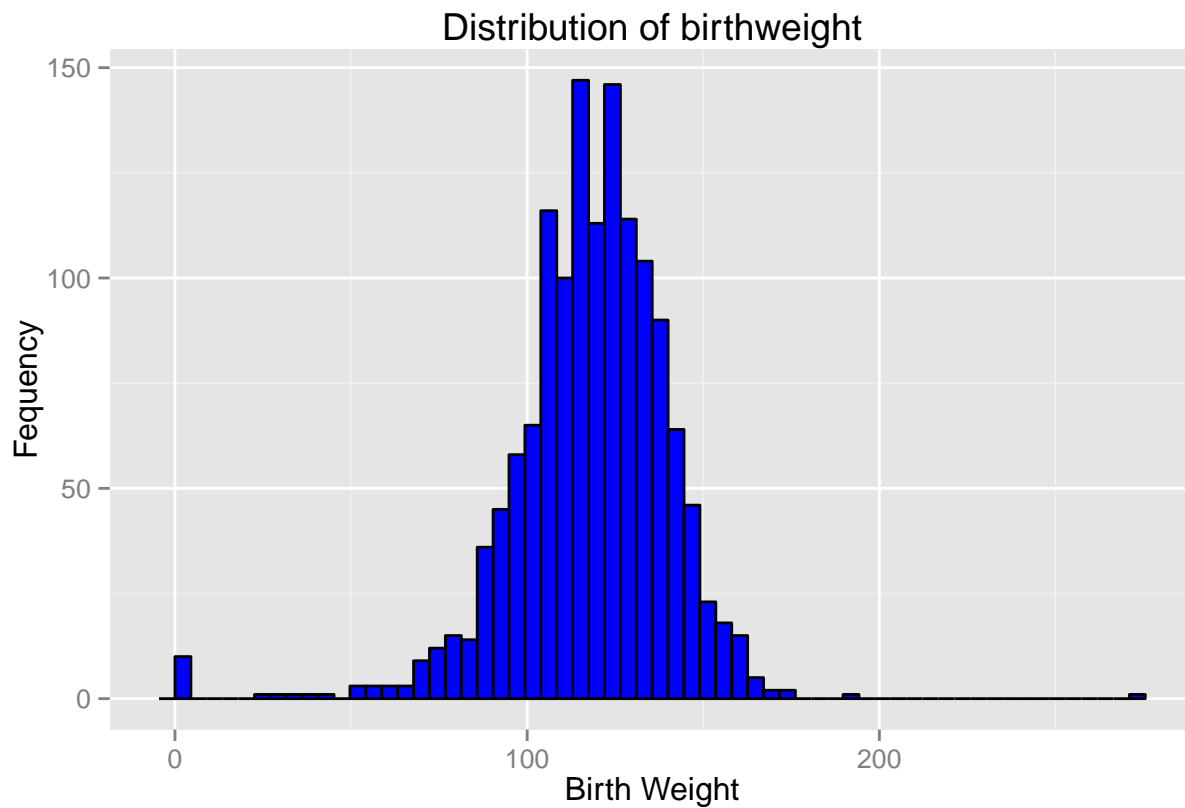
```
bwght.hist <- ggplot(data, aes(bwght)) +
```

```
  theme(legend.position = "none") +
```

```
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$bwght)[2] - range(data$bwght)[1]) / 60)
```

```
  labs(title = "Distribution of birthweight", x = "Birth Weight", y = "Frequency")
```

```
plot(bwght.hist)
```



Data observations:

These are the outlier observations. There are babies with birthweights of 0 and over 200 once
`print(sum(data$bwght == 0))`
`print(data$bwght[data$bwght > 200])`
 We should remove the zero baby weights from the data. They probably correspond to data entry issues. We can probably keep the single 271 pounds baby observation in the data set.

Question 4

```
# Get summary statistics about the cigarettes smoked variable.
print(summary(data$cigs))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.000   0.000   0.000   2.087   0.000   50.000
```

```
print(sum(is.nan(data$cigs)))
```

```
## [1] 0
```

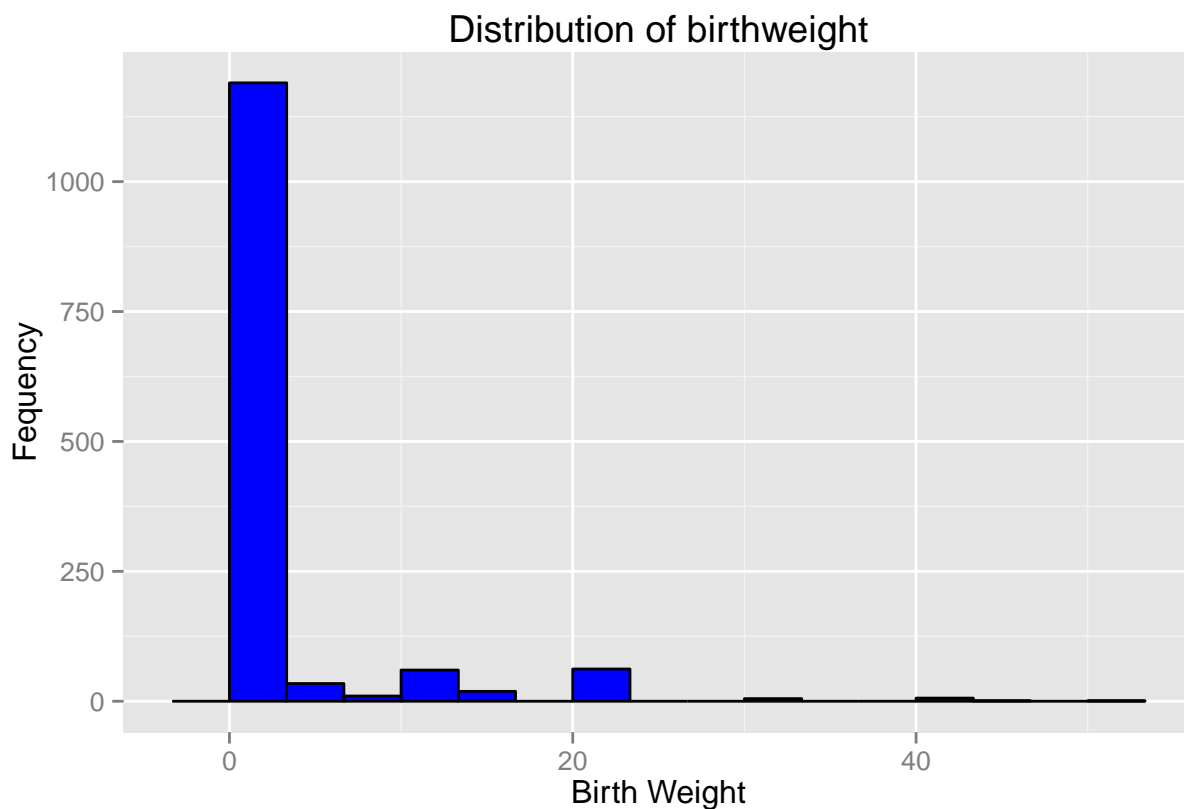
```
print(quantile(data$cigs, probs = c(.1, .5, .10, .25, .50, .75, .90, .95, .99, 1)))
```

```
## 10% 50% 10% 25% 50% 75% 90% 95% 99% 100%
## 0 0 0 0 0 0 10 20 20 50
```

```
# Plot the histogram of cigs at 15 bins
```

```
cigs.hist <- ggplot(data, aes(cigs)) +
  theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$cigs)[2] - range(data$cigs)[1]) / 15,
  labs(title = "Distribution of birthweight", x = "Birth Weight", y = "Frequency")
```

```
plot(cigs.hist)
```

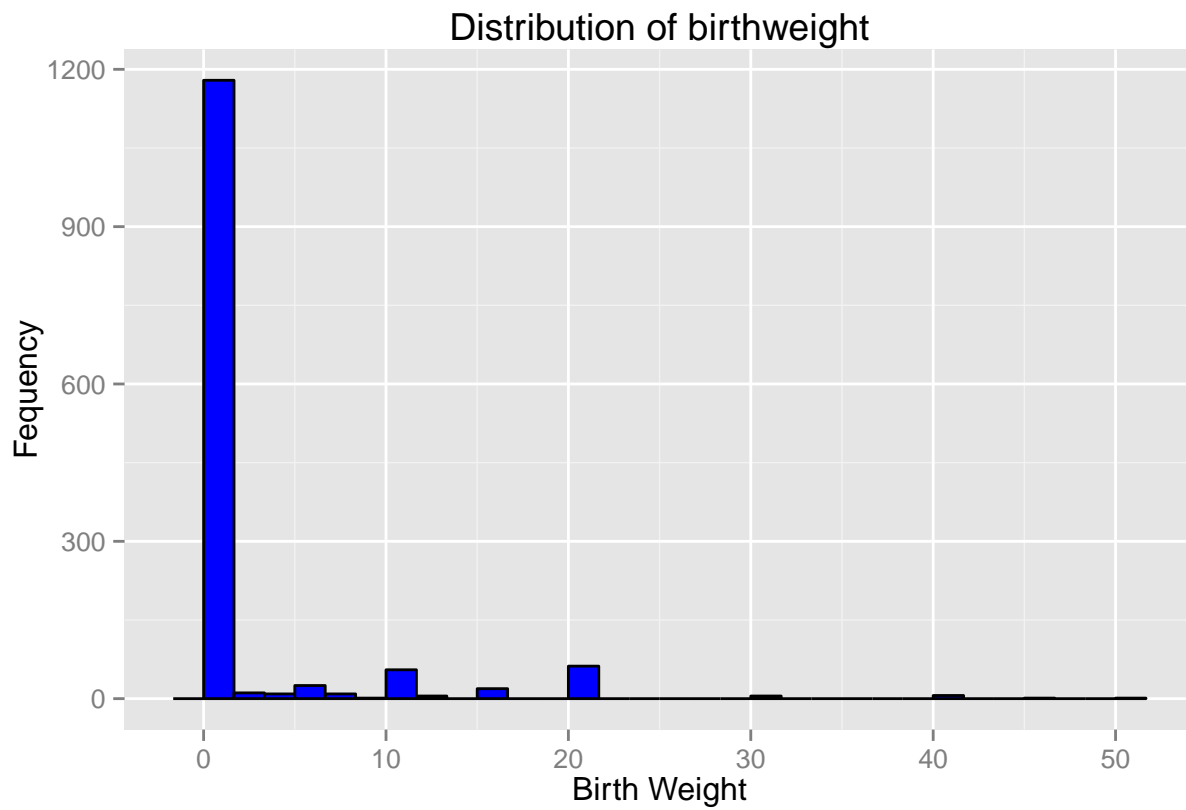


```
# Plot the histogram of cigs at 30 bins
```

```
cigs.hist <- ggplot(data, aes(cigs)) +
  theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black") +
  labs(title = "Distribution of birthweight", x = "Birth Weight", y = "Frequency")
```

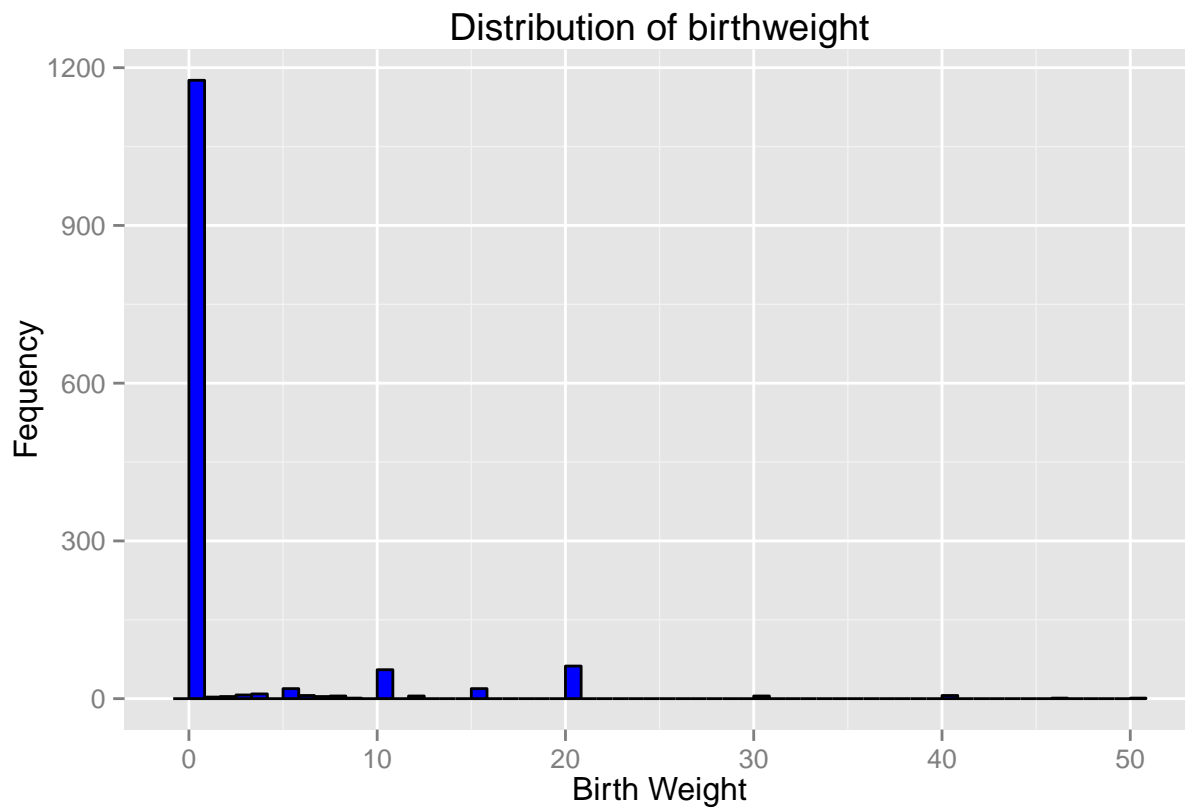
```
plot(cigs.hist)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
# Plot the histogram of cigs at 60 bins
cigs.hist <- ggplot(data, aes(cigs)) +
  theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$cigs)[2] - range(data$cigs)[1]) / 60) +
  labs(title = "Distribution of birthweight", x = "Birth Weight", y = "Frequency")

plot(cigs.hist)
```

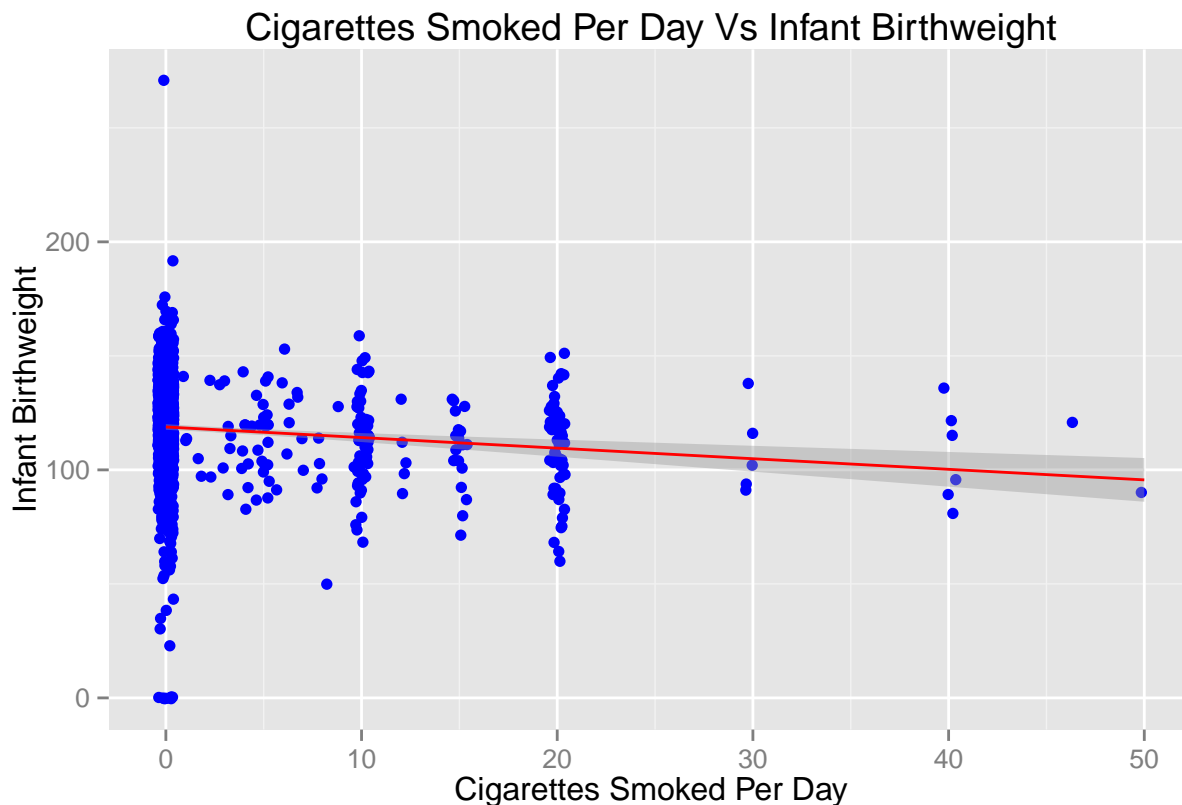



Data observations ————— The histogram for the number of cigarettes smoked is negatively skewed with a very high proportion of individuals smoking zero cigarettes per day during heir pregnancy. There are no visible signs of other anomalies in the data.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Question 5:

```
# Create a scatterplot of cigarettes smoked per day vs baby birthweight
scatter.bwght.cigs <- ggplot(data, aes(cigs, bwght)) +
  geom_point(colour = "Blue", position = "jitter") +
  geom_smooth(method = "lm", colour = "Red") +
  labs(x = "Cigarettes Smoked Per Day",
       y = "Infant Birthweight",
       title = "Cigarettes Smoked Per Day Vs Infant Birthweight")
plot(scatter.bwght.cigs)
```



Based on the scatterplot and the fitted lm curve on it, it appears that only a very small amount of the variation of bwght will be explained by cigs. That's because the variation explained in the graph appears to be much lower than the variation of birthweights at any level of daily cigarette consumption

Question 6:

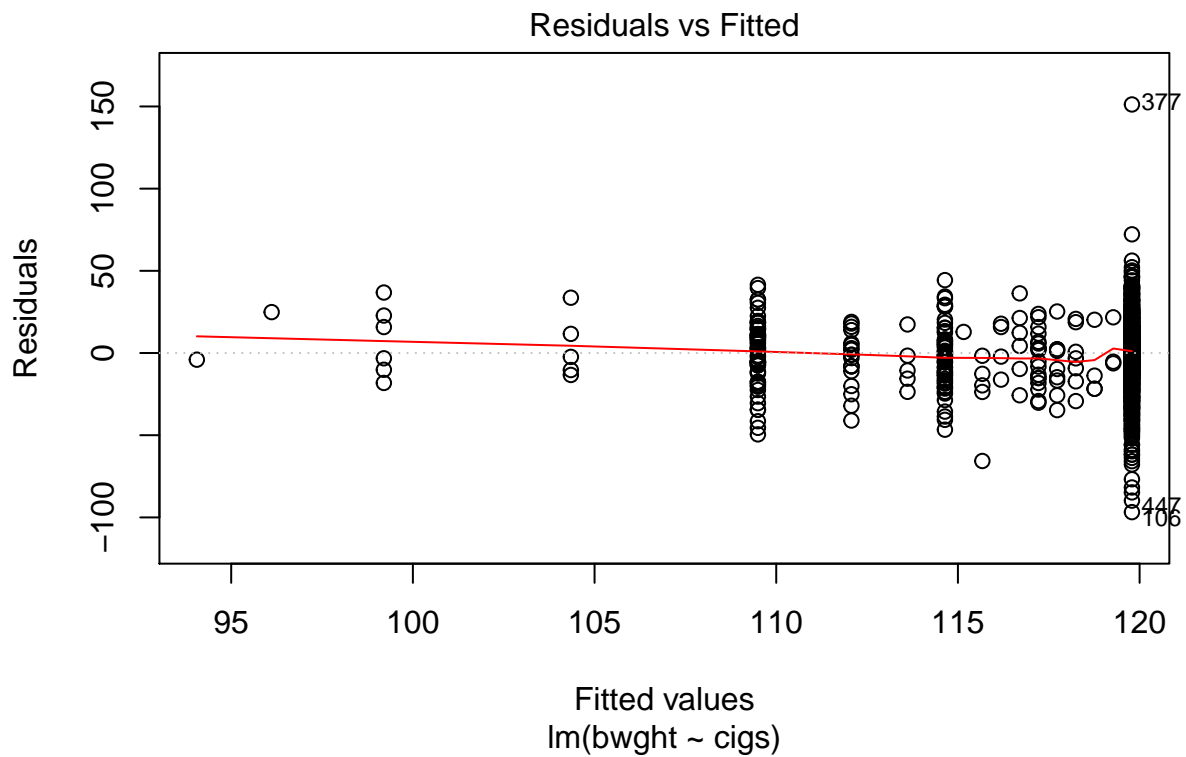
```
# Clean up the data based on previous observations
data <- data[data$bwght != 0, ]

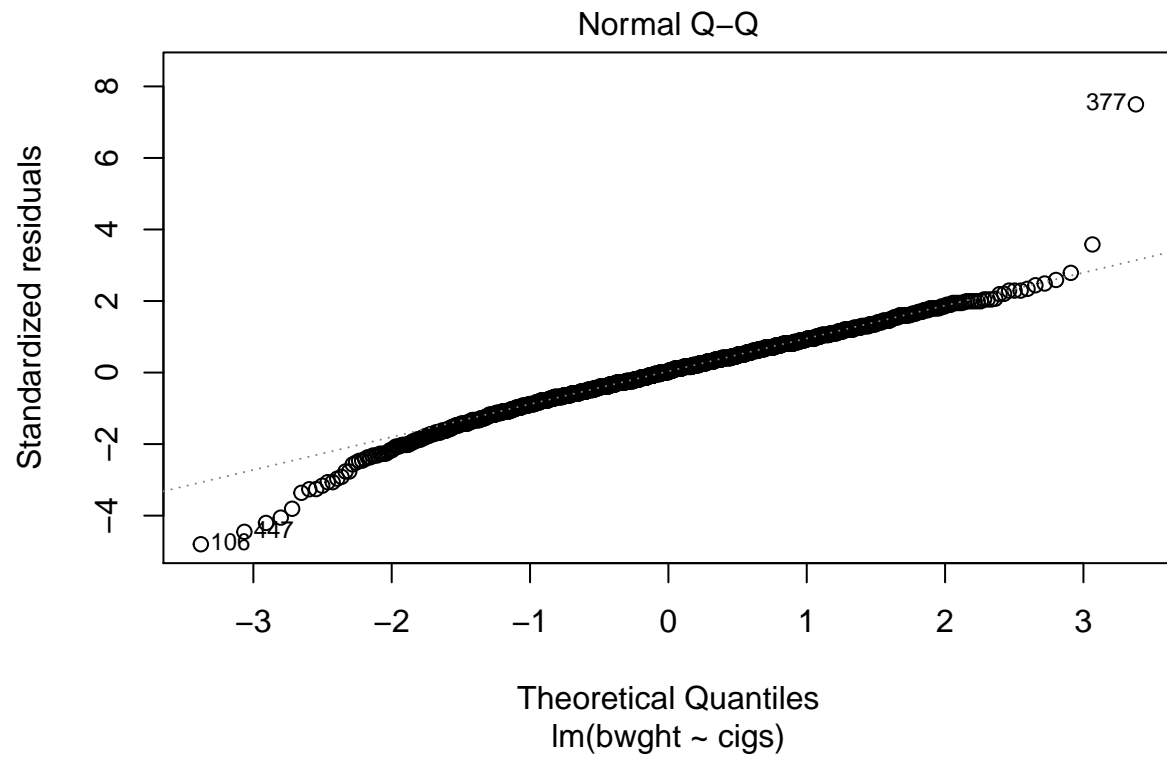
# Now perform the OLS regression
simple.ols.cigs.bwght <- lm(bwght ~ cigs, data = data)
print(summary.lm(simple.ols.cigs.bwght))

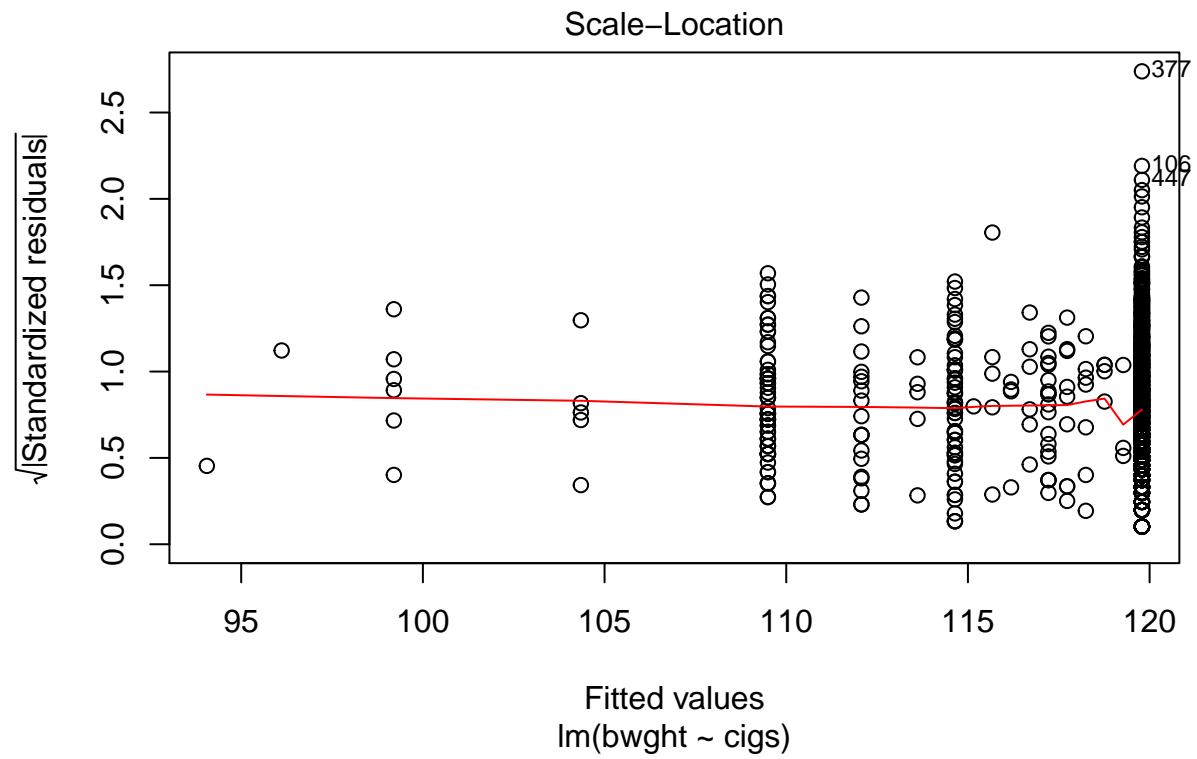
##
## Call:
## lm(formula = bwght ~ cigs, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -96.790 -11.790   0.357  13.210 151.210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119.78960    0.57595  207.987 < 2e-16 ***
```

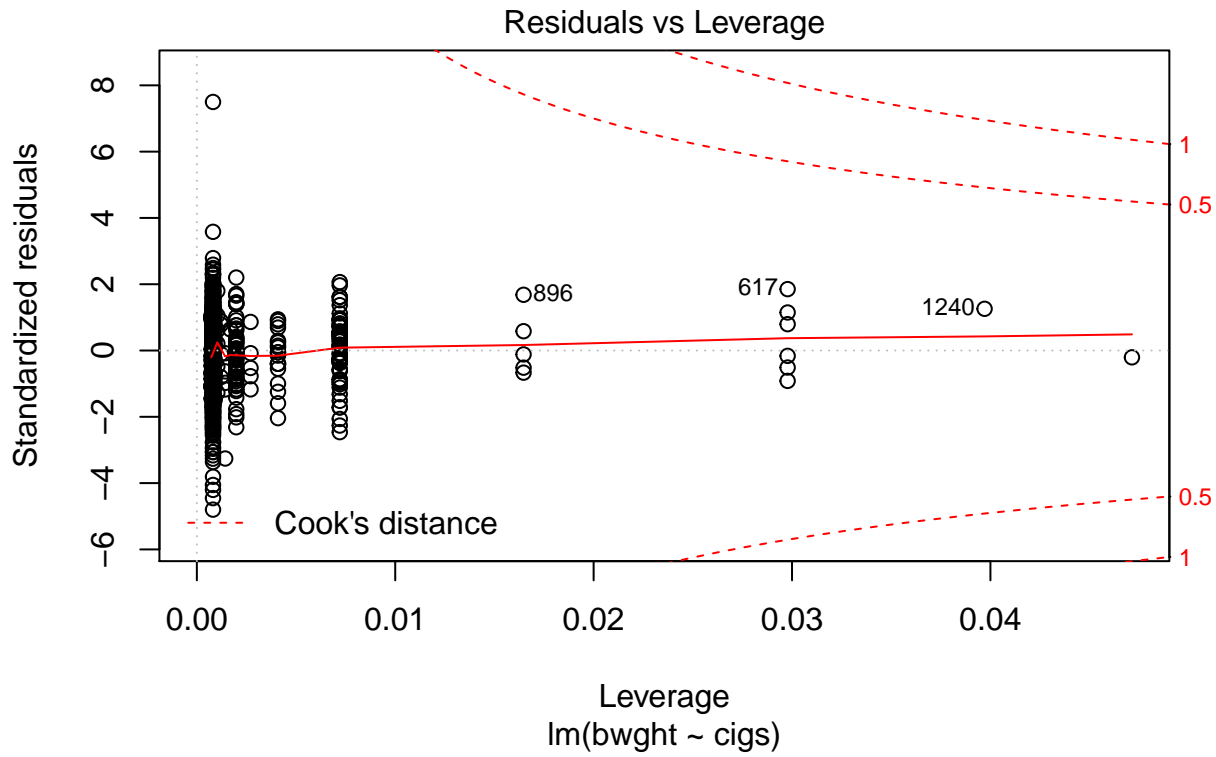
```
## cigs      -0.51470    0.09073   -5.673 1.71e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.17 on 1376 degrees of freedom
## Multiple R-squared:  0.02285,    Adjusted R-squared:  0.02214
## F-statistic: 32.18 on 1 and 1376 DF,  p-value: 1.711e-08
```

```
plot(simple.ols.cigs.bwght)
```









Our null hypothesis is that there is no relationship between the `cigs` variable and the `bwght` variable. The p-value of the f-statistic of the model is significant at 1.711e-08, indicating that we should reject the null hypothesis. The model accounts for 2.28% of the variability in the variable `bwght`. That value is given by a R^2 value of 0.2285 and a R value of 0.151 indicating a relatively small effect size.

The model shows a negative coefficient for the variable `cigs` with a value of -0.5147. The coefficients of the model are statistically significant.

Question 7

```
# Obtain descriptive statistics for the new variable
print(summary(data$faminc))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.50  14.50   27.50   29.03  37.50   65.00
```

```
print(sum(is.nan(data$faminc)))
```

```
## [1] 0
```

```
print(quantile(data$faminc, probs = c(.1, .5, .10, .25, .50, .75, .90, .95, .99, 1)))
```

```
## 10% 50% 10% 25% 50% 75% 90% 95% 99% 100%
## 6.5 27.5 6.5 14.5 27.5 37.5 65.0 65.0 65.0 65.0
```

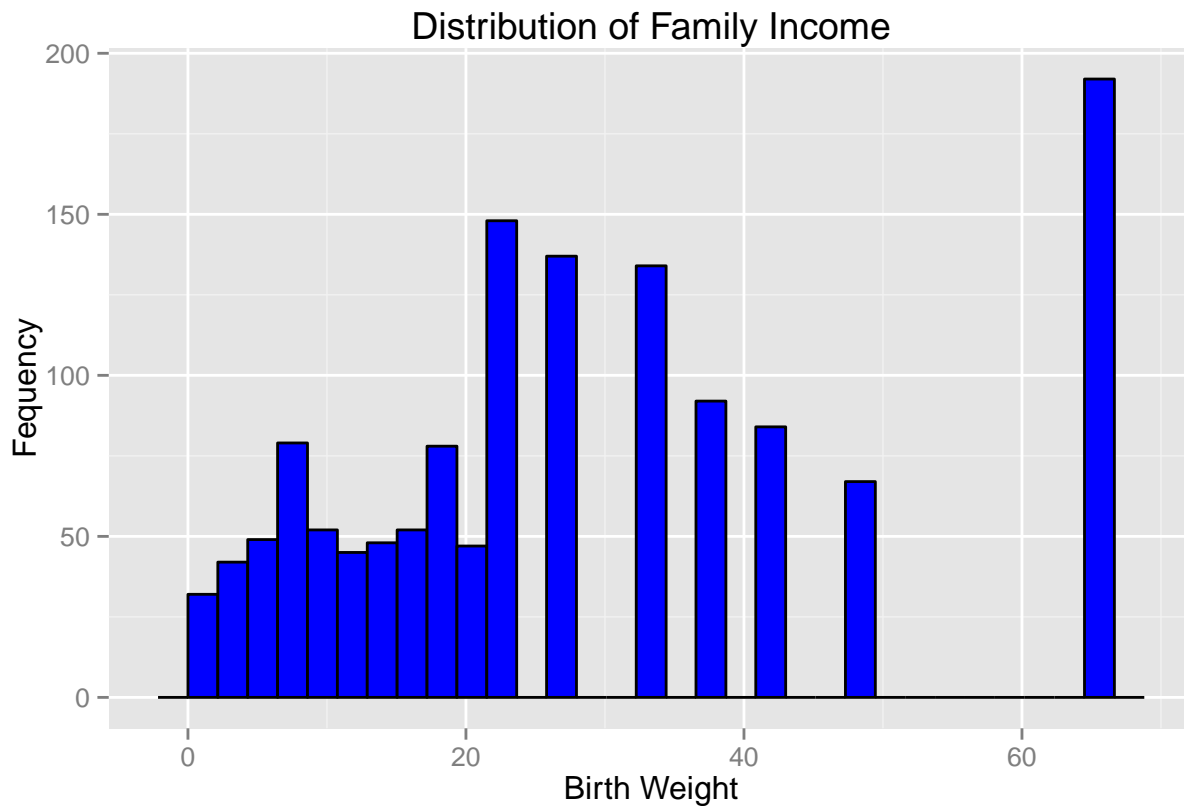
```
print(stat.desc(data$faminc, basic=FALSE, norm = TRUE))
```

```
##          median          mean      SE.mean  CI.mean.0.95          var
## 2.750000e+01 2.902903e+01 5.061299e-01 9.928691e-01 3.529988e+02
##      std.dev      coef.var      skewness      skew.2SE      kurtosis
## 1.878826e+01 6.472233e-01 6.159670e-01 4.672491e+00 -5.395270e-01
##      kurt.2SE      normtest.W      normtest.p
## -2.047800e+00 9.191810e-01 1.595212e-26
```

```
# Plot the histogram of faminc at 30 bins
faminc.hist <- ggplot(data, aes(faminc)) +
  theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black") +
  labs(title = "Distribution of Family Income", x = "Birth Weight", y = "Fequency")

plot(faminc.hist)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



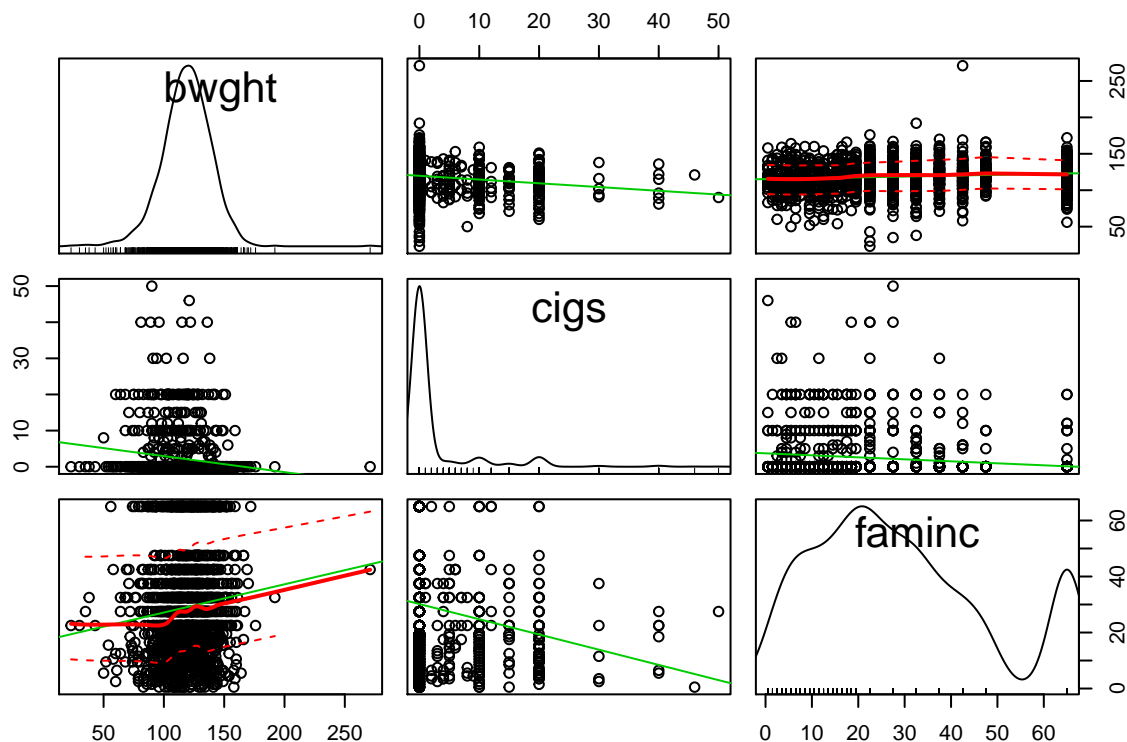
```
# Produce a scatterplot of bwght, cigs and faminc
scatterplotMatrix(~ bwght + cigs + faminc, data = data)
```

```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```

```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```

```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```

```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```



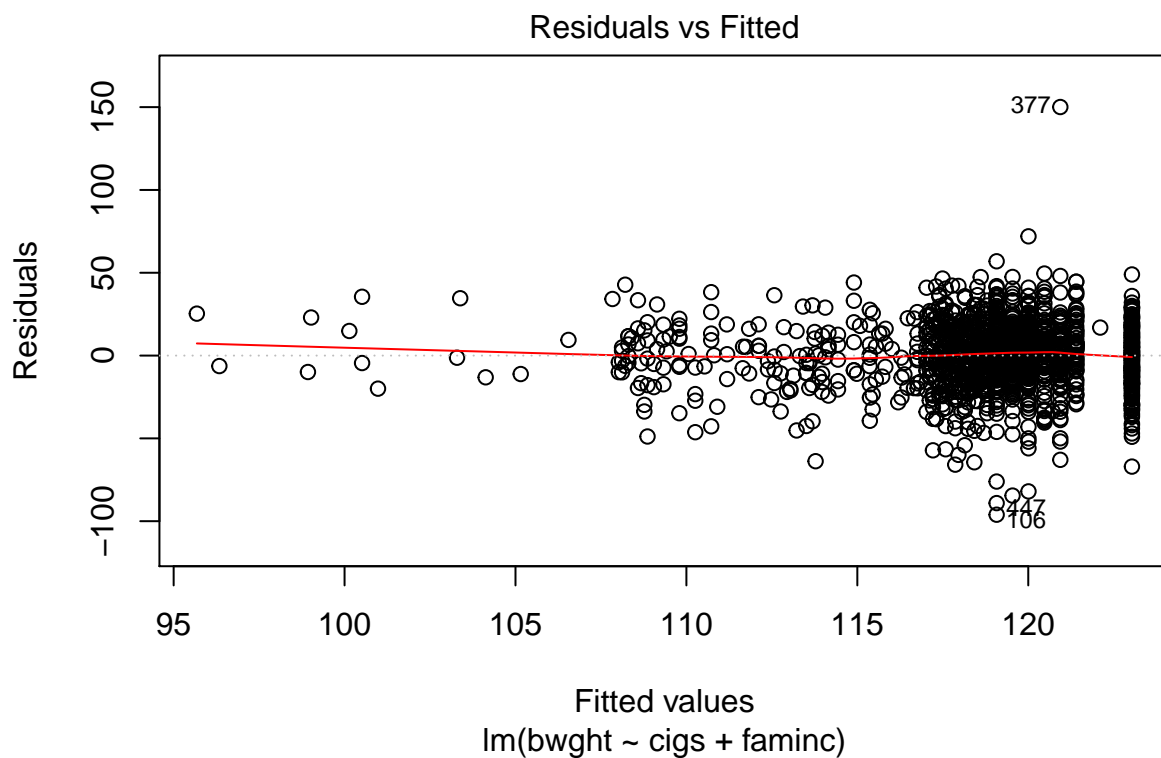
Question 8:

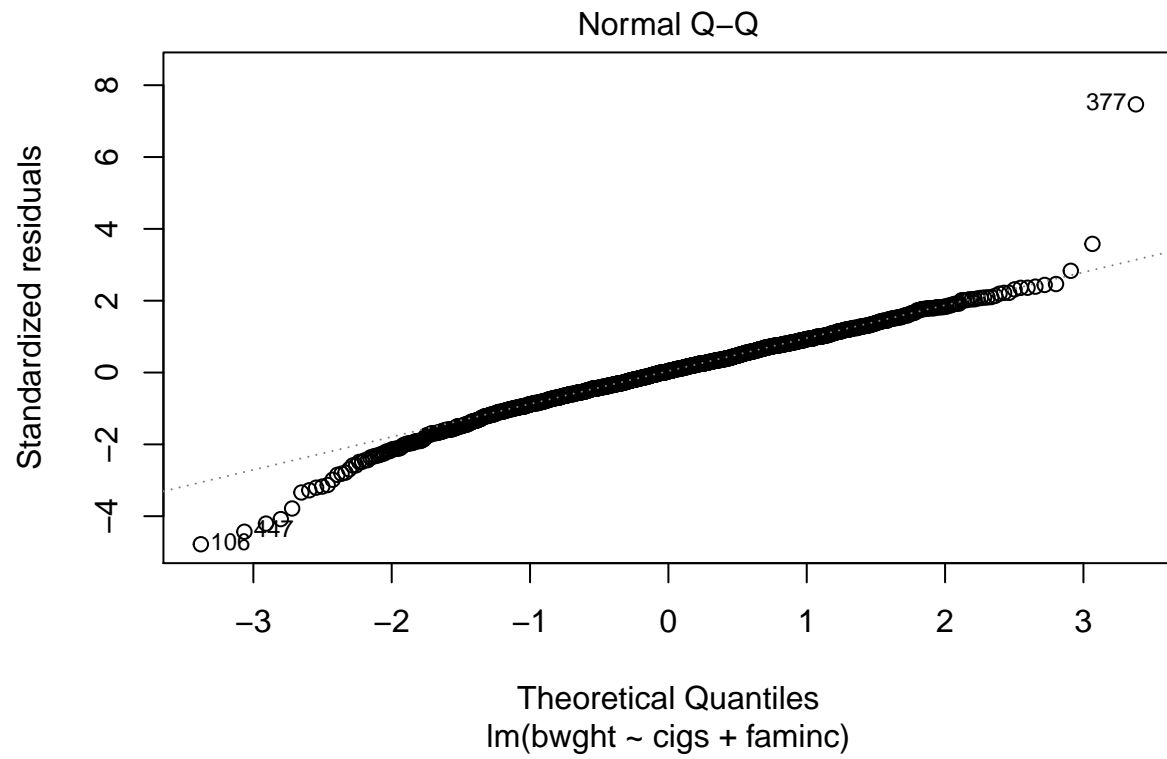
```
#Introduce a new independent variable to the model
multiple.ols.cigs.faminc.bwght <- lm(bwght ~ cigs + faminc, data = data)
print(summary.lm(multiple.ols.cigs.faminc.bwght))
```

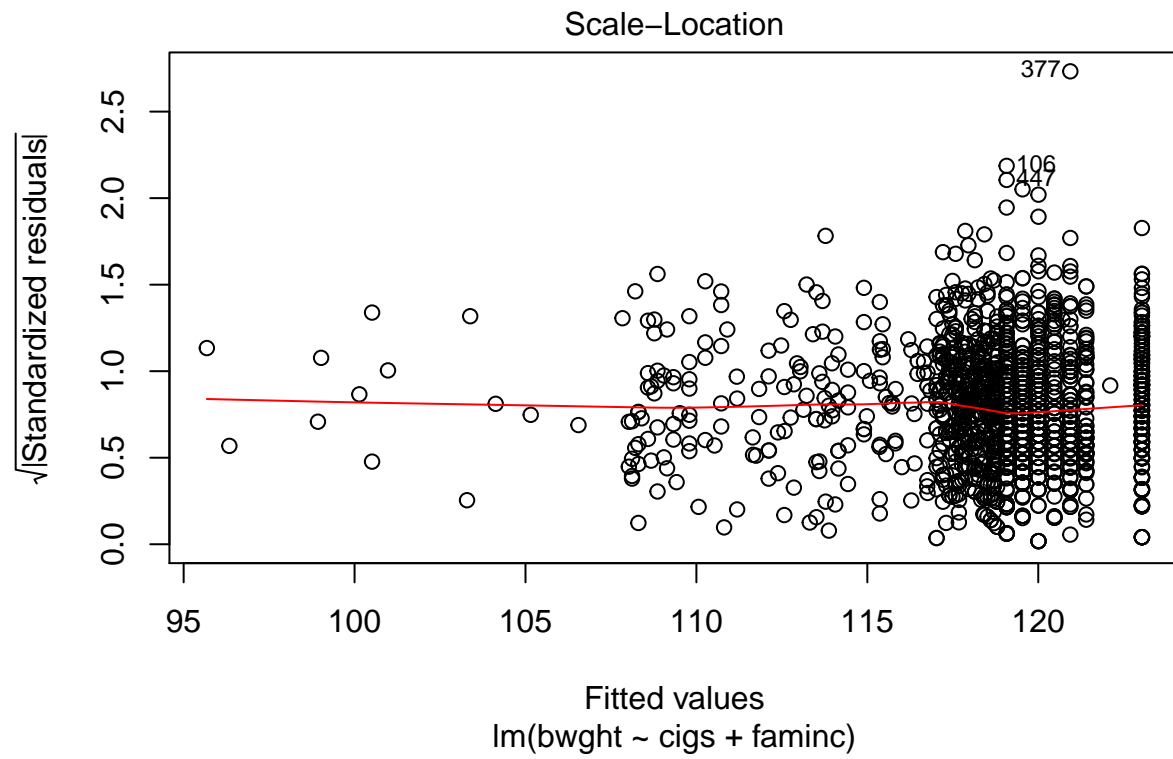


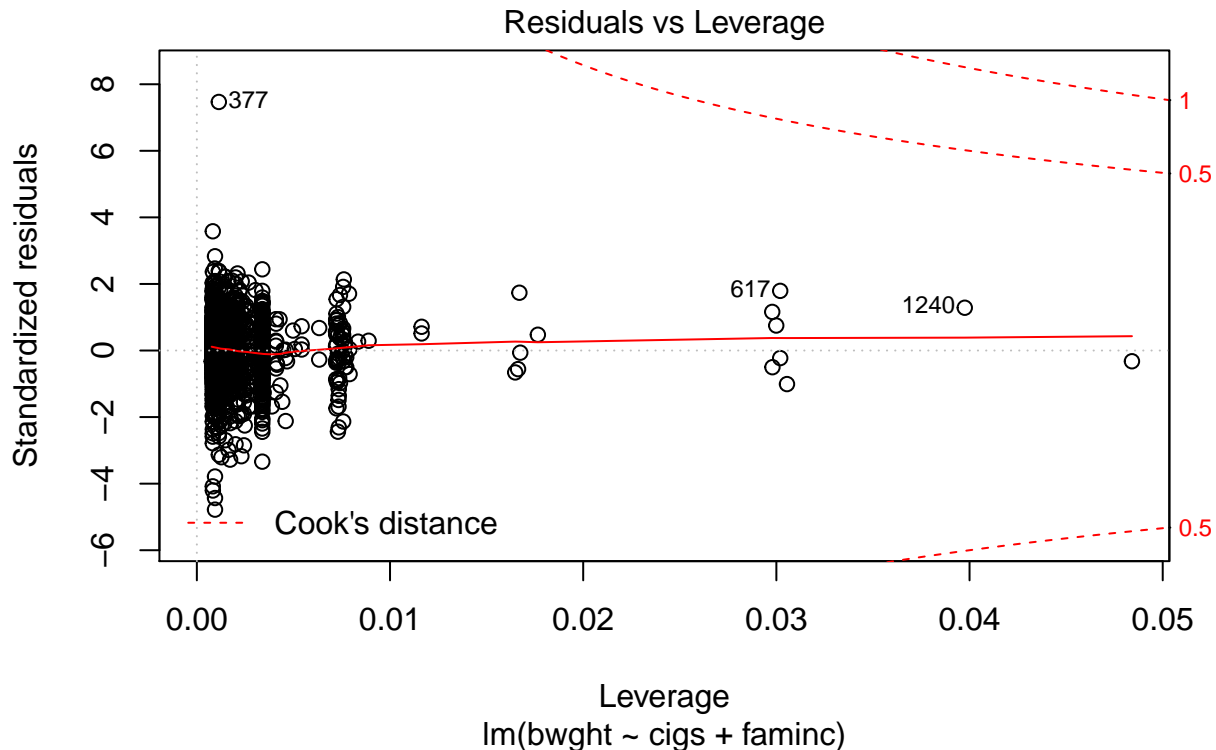
```
##
## Call:
## lm(formula = bwght ~ cigs + faminc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -96.075 -11.592   0.722  13.262 150.062
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.97933    1.05363  111.025  < 2e-16 ***
## cigs         -0.46407    0.09182   -5.054 4.91e-07 ***
## faminc        0.09314    0.02928    3.181  0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.11 on 1375 degrees of freedom
## Multiple R-squared:  0.02999,    Adjusted R-squared:  0.02858
## F-statistic: 21.25 on 2 and 1375 DF,  p-value: 8.109e-10
```

```
plot(multiple.ols.cigs.faminc.bwght)
```









Our null hypothesis is that there is no relationship between the `cigs` variable and the `bwght` and `faminc` variables. The p-value of the f-statistic of the model is significant at 8.109×10^{-10} , indicating that we should reject the null hypothesis. The model accounts for 2.99% of the variability in the variable `bwght`. That value is given by a R^2 value of 0.299 and a R value of 0.173 indicating a relatively small effect size.

The model shows a negative coefficient for the variable `cigs` with a value of -0.4640. It shows a positive coefficient for the variable `faminc` of 0.093. The coefficients of the model are statistically significant.

Question 9/10

The coefficient on `cigs` in the multiple regression means that for every additional cigarette smoked per day by the pregnant mother during their maternity, and leaving the income variable constant, the body weight of the infant children decreases by 0.4640 pound.

In the simple regression, we saw that this coefficient was also negative and had a value of -0.5147.

In the simple regression case, the coefficient on `cigs` meant that for every additional cigarette smoked per day, independent of any other condition of the mother, the birth weight of the infant was reduced by -0.5147 lbs.

In the multiple regression, the variance explained by the `faminc` variable was captured in the residuals of the simple model and partially also in the coefficient of the `cigs` variable as we can suspect there is some correlation between the two variables.

Therefore the introduction of the additional variable `faminc` has reduced the contribution of the `cigs` variable to the infant body weight. We can hypothesize that there is some correlation between the number of cigarettes smoked per day and the family income, where mothers at higher family income have better health habits and therefore smoke less. And introducing the family income variable thus takes away some of the variance explanation previously captured by the `cigs` variable.