DataSci 271: Exercise 1

September 18, 2016

The file birthweight_w271.Rdatabirthweight_w271.Rdata contains data from the 1988 National Health Interview Survey 1988 National Health Interview Survey, which is modified by the instructor. This survey is conducted by the U.S. Census Bureau and has collected data on individual health metrics since 1957. Like all surveys, a full analysis would require advanced techniques such as those provided by the R survey package. For this homework, however, you are to treat the data as a true random sample. You will use this dataset to practice interpreting OLS coefficients.

Exercises

Question 1: Examine the basic structure of the data

Load the birthweight dataset. * Examine the basic structure of the data set. RR functions such as desc, str, summarydesc, str, summary may be useful. * Describe the number of variables and observations in the data. * Examine if there are any missing values in each of the variables.

```
library(dtplyr)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
rm(list = ls())
load('birthweight_w271.rdata')
```

```
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
##
```

```
describe(data)
```

```
## data
##
##  14  Variables      1388  Observations
## ---------------------------------------------------------------------------
## faminc
##        n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0      27    0.99   29.03     3.5     6.5    14.5    27.5
##     .75     .90     .95
##    37.5    65.0    65.0
##
## lowest :  0.5  1.5  2.5  3.5  4.5, highest: 32.5 37.5 42.5 47.5 65.0
## ---------------------------------------------------------------------------
## cigtax
##        n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0      28    0.99   19.55       3      10      15      20
##     .75     .90     .95
##      26      27      31
##
## lowest :  2.0  2.5  3.0  7.0  8.0, highest: 30.0 31.0 33.0 34.0 38.0
## ---------------------------------------------------------------------------
## cigprice
##        n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0      46       1   130.6   109.4   120.2   122.8   130.8
##     .75     .90     .95
##   137.0   142.0   148.6
##
```

```
## lowest : 103.8 107.6 109.4 111.9 118.6
## highest: 145.6 148.6 149.1 150.6 152.5
## ---------------------------------------------------------------------------
## bwght
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0     117       1   117.9      83      93     106     119
##     .75     .90     .95
##     132     143     149
##
## lowest :   0  23  30  35  38, highest: 170 172 176 192 271
## ---------------------------------------------------------------------------
## fatheduc
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1192     196      18    0.94   13.19       9      10      12      12
##     .75     .90     .95
##      16      17      18
##
##           1 2 3 4 5  6  7  8  9 10 11  12 13  14 15  16 17 18
## Frequency 1 2 4 3 4 10 10 22 17 49 64 443 87 115 43 189 32 97
## %         0 0 0 0 0  1  1  2  1  4  5  37  7  10  4  16  3  8
## ---------------------------------------------------------------------------
## motheduc
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1387       1      17    0.93   12.94       9      10      12      12
##     .75     .90     .95
##      14      16      17
##
##           2 3 4 5 6 7  8  9 10 11  12  13  14 15  16 17 18
## Frequency 1 1 1 2 9 7 21 43 70 67 562 122 151 41 198 37 54
## %         0 0 0 0 1 1  2  3  5  5  41   9  11  3  14  3  4
## ---------------------------------------------------------------------------
## parity
##       n missing  unique    Info    Mean
##    1388       0       6    0.79   1.633
##
##             1   2   3  4  5 6
## Frequency 795 389 146 39 15 4
## %          57  28  11  3  1 0
## ---------------------------------------------------------------------------
## male
##       n missing  unique    Info     Sum    Mean
##    1388       0       2    0.75     723  0.5209
## ---------------------------------------------------------------------------
## white
##       n missing  unique    Info     Sum    Mean
##    1388       0       2    0.51    1089  0.7846
## ---------------------------------------------------------------------------
## cigs
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0      18    0.39   2.087       0       0       0       0
##     .75     .90     .95
##       0      10      20
##
##              0 1 2 3 4  5 6 7 8 9 10 12 15 20 30 40 46 50
## Frequency 1176 3 4 7 9 19 6 4 5 1 55  5 19 62  5  6  1  1
## %           85 0 0 1 1  1 0 0 0 0  4  0  1  4  0  0  0  0
## ---------------------------------------------------------------------------
## lbwght
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0     117       1   4.726   4.419   4.533   4.663   4.779
##     .75     .90     .95
##   4.883   4.963   5.004
##
## lowest : 0.000 3.135 3.401 3.555 3.638
## highest: 5.136 5.147 5.170 5.257 5.602
## ---------------------------------------------------------------------------
## bwghtlbs
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0     117       1   7.366   5.188   5.812   6.625   7.438
##     .75     .90     .95
##   8.250   8.938   9.312
##
## lowest :  0.000  1.438  1.875  2.188  2.375
## highest: 10.625 10.750 11.000 12.000 16.938
## ---------------------------------------------------------------------------
## packs
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0      18    0.39  0.1044     0.0     0.0     0.0     0.0
##     .75     .90     .95
##     0.0     0.5     1.0
##
## 0 (1176, 85%), 0.0500000007450581 (3, 0%)
## 0.100000001490116 (4, 0%)
## 0.150000005960464 (7, 1%)
## 0.200000002980232 (9, 1%), 0.25 (19, 1%)
## 0.300000011920929 (6, 0%)
## 0.349999994039536 (4, 0%)
## 0.400000005960464 (5, 0%)
## 0.449999988079071 (1, 0%), 0.5 (55, 4%)
## 0.600000023841858 (5, 0%), 0.75 (19, 1%)
## 1 (62, 4%), 1.5 (5, 0%), 2 (6, 0%), 2.29999995231628 (1, 0%)
## 2.5 (1, 0%)
## ---------------------------------------------------------------------------
## lfaminc
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1388       0      27    0.99   3.071   1.253   1.872   2.674   3.314
##     .75     .90     .95
##   3.624   4.174   4.174
##
## lowest : -0.6931  0.4055  0.9163  1.2528  1.5041
## highest:  3.4812  3.6243  3.7495  3.8607  4.1744
## ---------------------------------------------------------------------------
```

Question 2: Exploratory Data Analysis (EDA)

As we mentioned in the live session, it is important to start with a question (or a hypothesis) when conducting regression modeling. In this execrise, we are in the question: "Do mothers who smoke have babies with lower birth weight?”

The dependent variable of interested is bwght, representing birthweight in ounces. Examine this variable using both tabulated and graphical summaries.

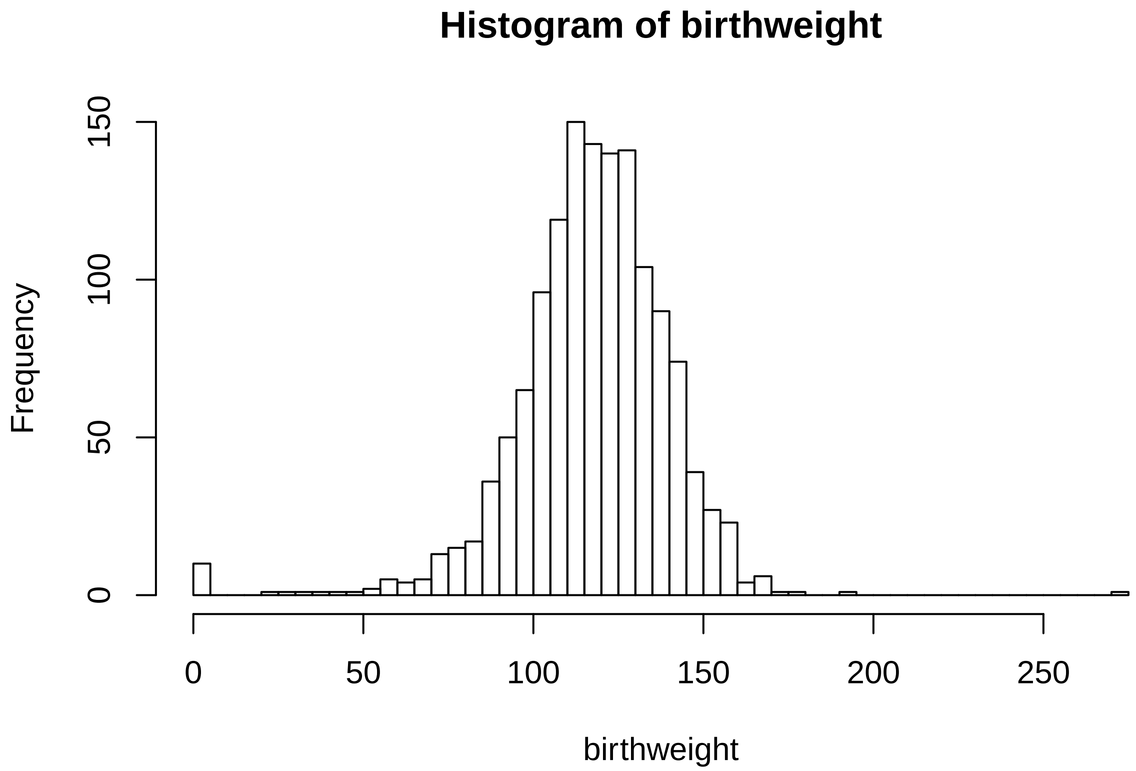
1. Summarize the variable *bwght*: *summary*(*data$bwght*)
2. You may also use the quantile function: *quantile*(*data$bwght*). List the following quantiles: 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%
3. Plot the histogram of *bwght* and comment on the shape of its distribution. Try different bin sizes and comment how it affects the shape of the histogram. Remember to label the graph clearly. You will also need a title for the graph.
4. This is a more open-ended question: Have you noticed anything "strange" with the *bwght* variable and the shape of histogram this variable? If so, please elaborate on your observations and investigate any issues you have identified.
5. Is the variable skewed?
6. Does the variable have extreme values or values that seem unreasonable?
7. Does the variable appear to be top- or bottom-coded?

```
EDA = function (d, label){
print("Summary")
print (summary(d))
print("Quantiles")
print(quantile(d, probs=c(0.01, 0.05, .1, .25, .5, .75, .9, .95, .99)))

hist(d, breaks=40, main=paste("Histogram of",label), xlab=label)

}
EDA(data$bwght,"birthweight")
```

```
## [1] "Summary"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   106.0   119.0   117.9   132.0   271.0
## [1] "Quantiles"
##     1%      5%     10%     25%     50%     75%     90%     95%     99%
##  42.35  83.00  93.00 106.00 119.00 132.00 143.00 149.00 160.13
```

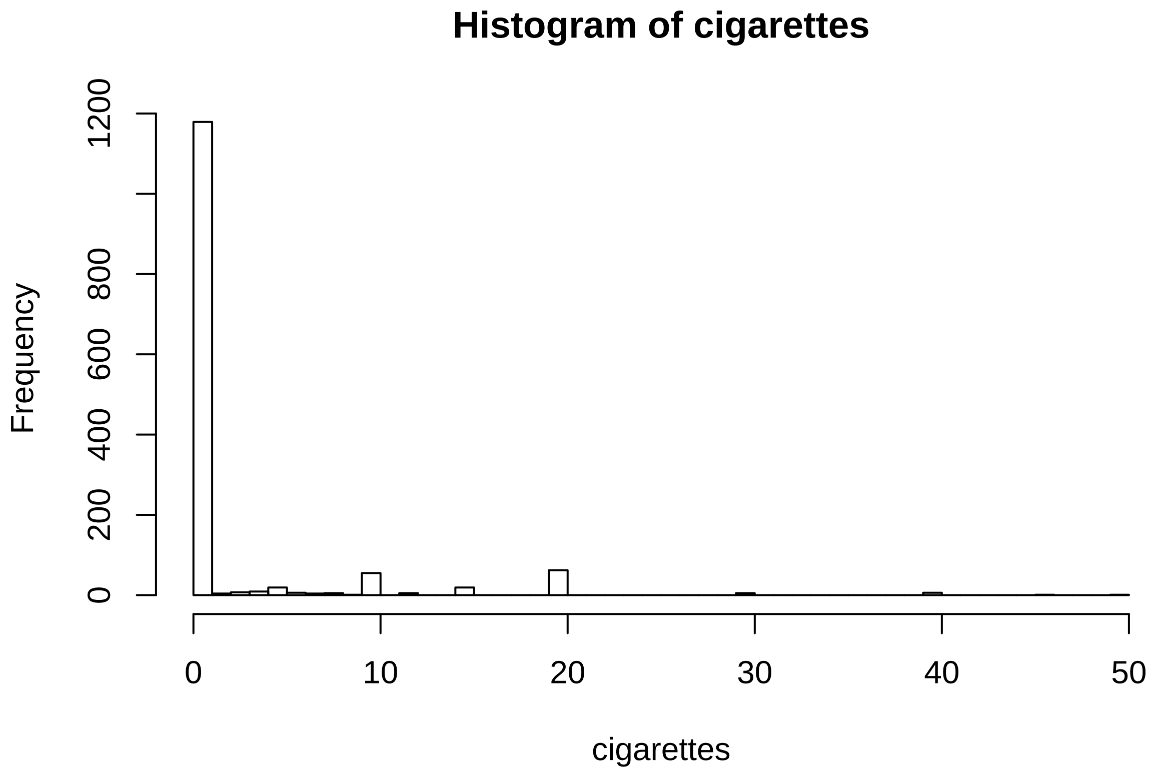


The bwght variable is not skewed nor does it appear to be top or bottom coded. However, it's unusual that there are so many observations with 0 birthweight. There's also one outlier on the other end of the graph at 271. These must be incorrectly recorded observations

The key explanatory variable of interest is cigs, which represents number of cigarettes smoked each day by the mother while pregnant. Conduct the same EDA analysis as that of the dependent variable.

```
EDA(data$cigs,"cigarettes")
```

```
## [1] "Summary"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.087   0.000  50.000
## [1] "Quantiles"
##  1%  5% 10% 25% 50% 75% 90% 95% 99%
##   0   0   0   0   0   0  10  20  20
```



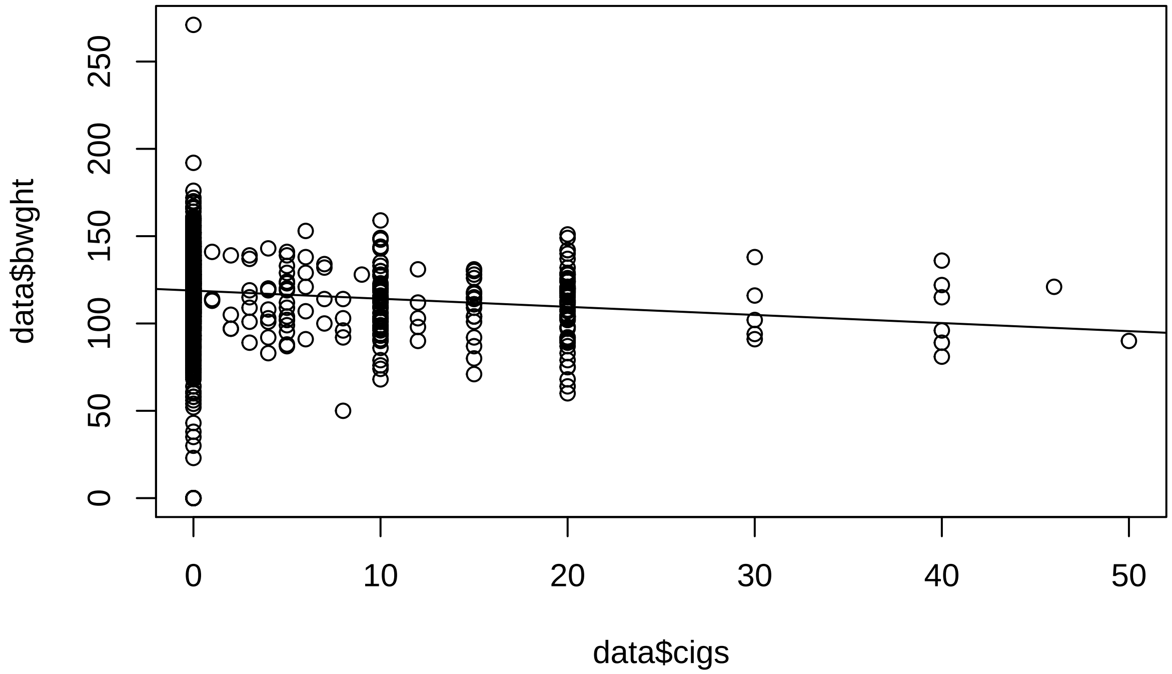
This variable is skewed towards the left, as more than 75% of the respondents reported 0. This variable may have been bottom coded, as I it's possible that the 0 value represents both mothers who don't smoke and mothers who smoke less than one cigarette a day on average (for example, maybe the smoke three times a week). The high values like 30 and 40 appear to appear to be somewhat extreme, but not impossible

After conducting the univariate analysis of both the dependent variable and the explanatory variable of interest, examine the relationship between the dependent variable and explanatory variable of interest. In this simple case, we only examine the dependent variable and one explanatory variable.

Start with generating a scatterplot of *bwght*} against *cigs*. Based on the appearance of this plot, how much of the variation in bwght do you think can be explained by *cigs*? Do the relationship appear to be linear?

Based on the plot, I doubt that cigs will be able explain much of the variation in bwght. Not only is the linear relationship weak, but cigs just does not have much variation in it.

```
plot(data$cigs, data$bwght)
abline(lm(data$bwght~ data$cigs))
```



The relationship does not appear to be linear. In fact, there does not seem to be a relationship at all. The trend line is slightly negative, but there is too much variance and not enough data at higher cig levels to actually see a relationship

Question 3: Build a Simple Linear Regression Model

Estimate the simple linear regression of *bwght* on cigs. That is, we enter the explanatory variable as its raw form. What coefficient estimates and the standard errors associated with the coefficient estimates do you get? Interpret the results.

```
m <- lm(bwght ~ cigs, data = data[data$bwght>0,])
summary(m)
```

```
##
## Call:
## lm(formula = bwght ~ cigs, data = data[data$bwght > 0, ])
##
## 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
```

The p value for cigs is statitically signficant, so there appears to be some relationship between cigarettes smoked each day and birthweight. It looks to be a slightly negative one, since the coefficient on cigs is negative. The standard error is relatively large compared to the coefficient, but not so large that one would doubt the sign of the coefficient.

However, do you think it a simple linear regression of *bwght* on *cigs* in their raw forms is an appropriate way to capture the effect of cigarettes smoked per day during pregnant on child birthweight? If not, please explain. You don't have to build another regression. Note: I have to emphasize again that we have not yet used the insights we generated from the EDA in specify the regression function. We will do so in week 5 when we learn about variable transformation and featuere engineering in general.

To capture the effect of cigarettes on birthweight, one needs to run an actual experiment where mothers are given either real or "placebo" cigarettes to smoke. A regression alone does not indicate causality.

Question 4: Regression Diagnostic

Conduct regression diagnostic of the above model. Try to use each of the diagnostic plots to "examine" the underlying assumptions of the Classical Linear Regression Models. Note that I use the term "examine" rather than "test" because regression diagnostic is a diagnostic tool and not a formal statistical testing, but they are extremely useful and should be a part of any regression model building process. Interpret each of your graphs and comments on whether the corresponding underlying assumptions make sense.

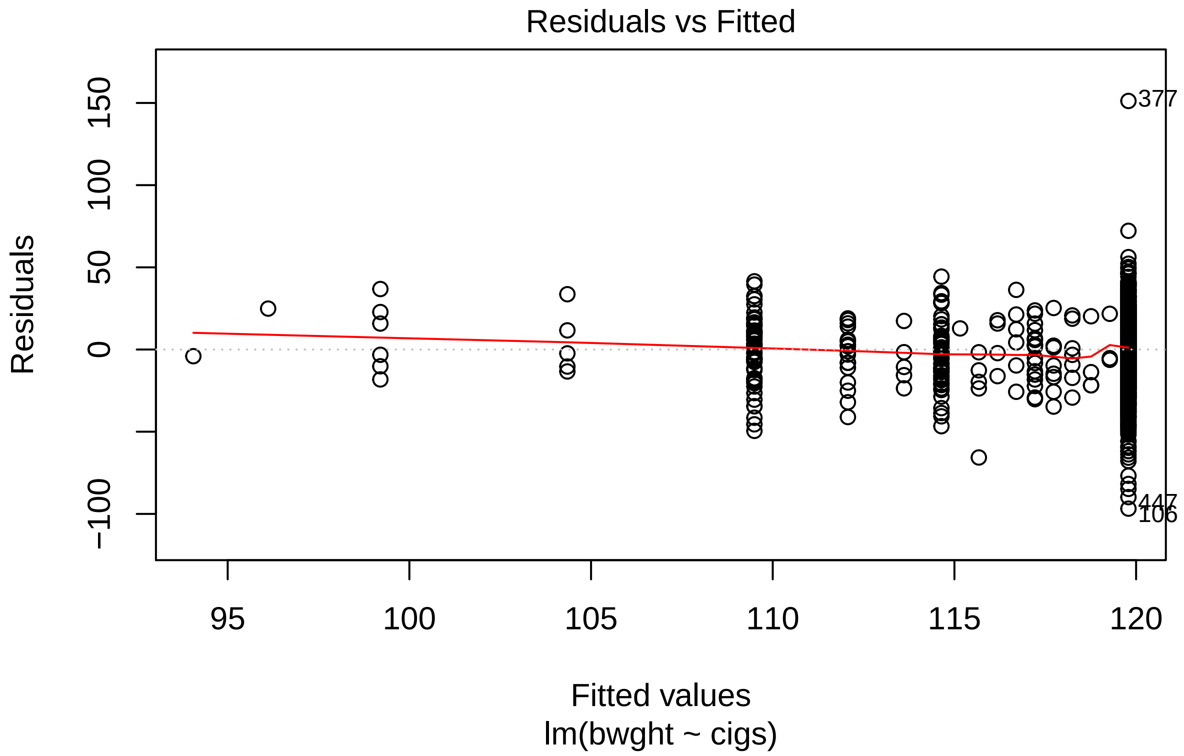
```
library(car)
print("Durbin Watson")
```

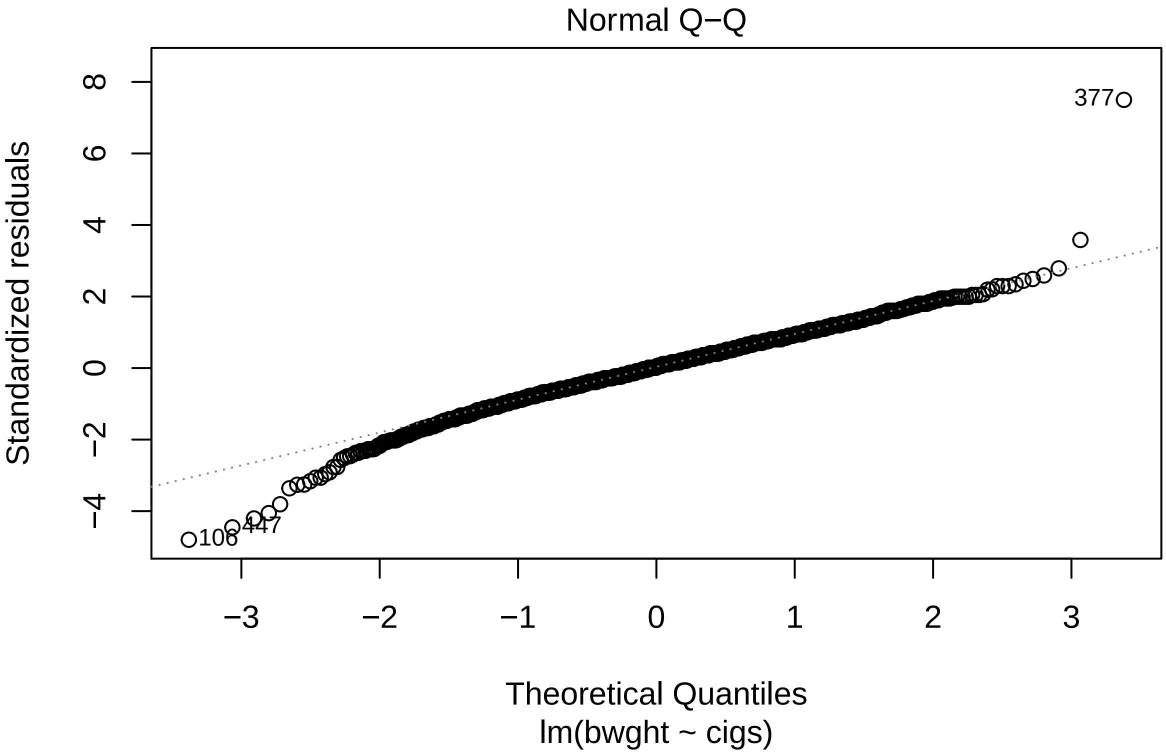
```
## [1] "Durbin Watson"
```

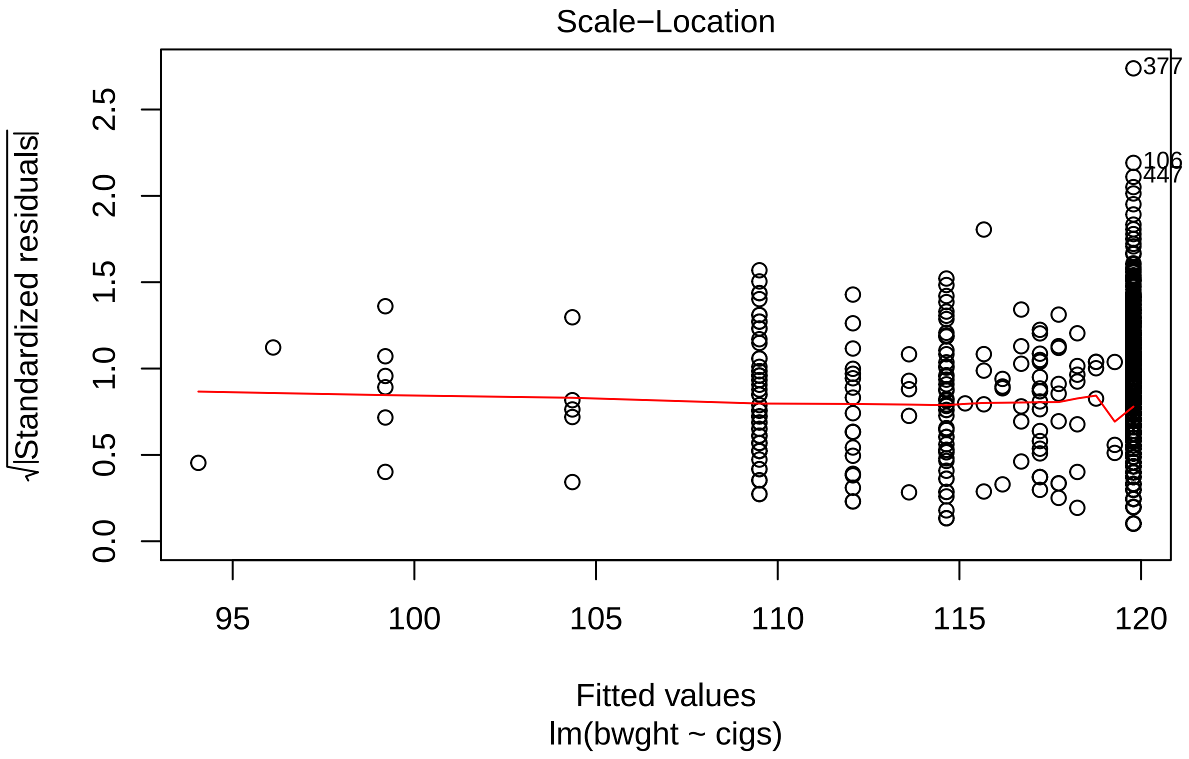
```
durbinWatsonTest(data$bwght)
```

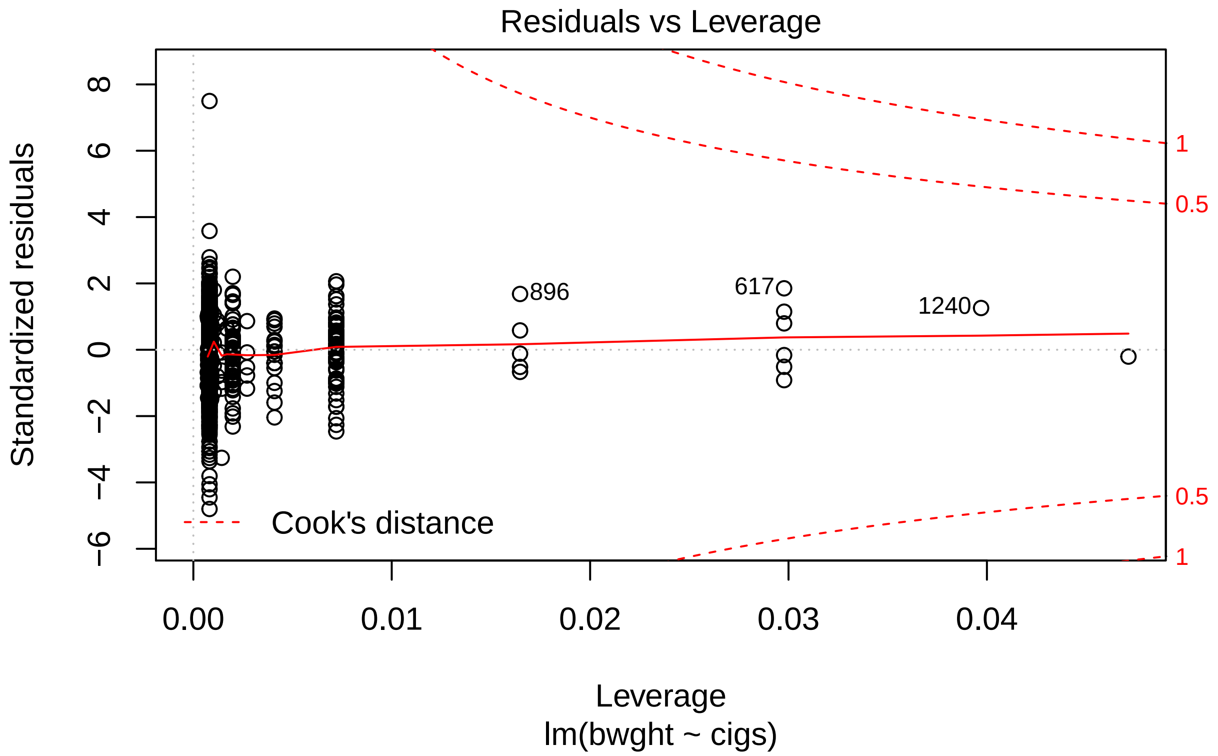
```
## [1] 0.06988125
```

```
plot(m)
```


Residuals vs Fitted
Residuals
150
100
50
0
−100
95
100
105
110
115
120
Fitted values
lm(bwght ~ cigs)
377
447
106


Normal Q−Q
377
106
447
Standardized residuals
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bwght ~ cigs)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
95
100
105
110
115
120
Fitted values
lm(bwght ~ cigs)
377
106
447



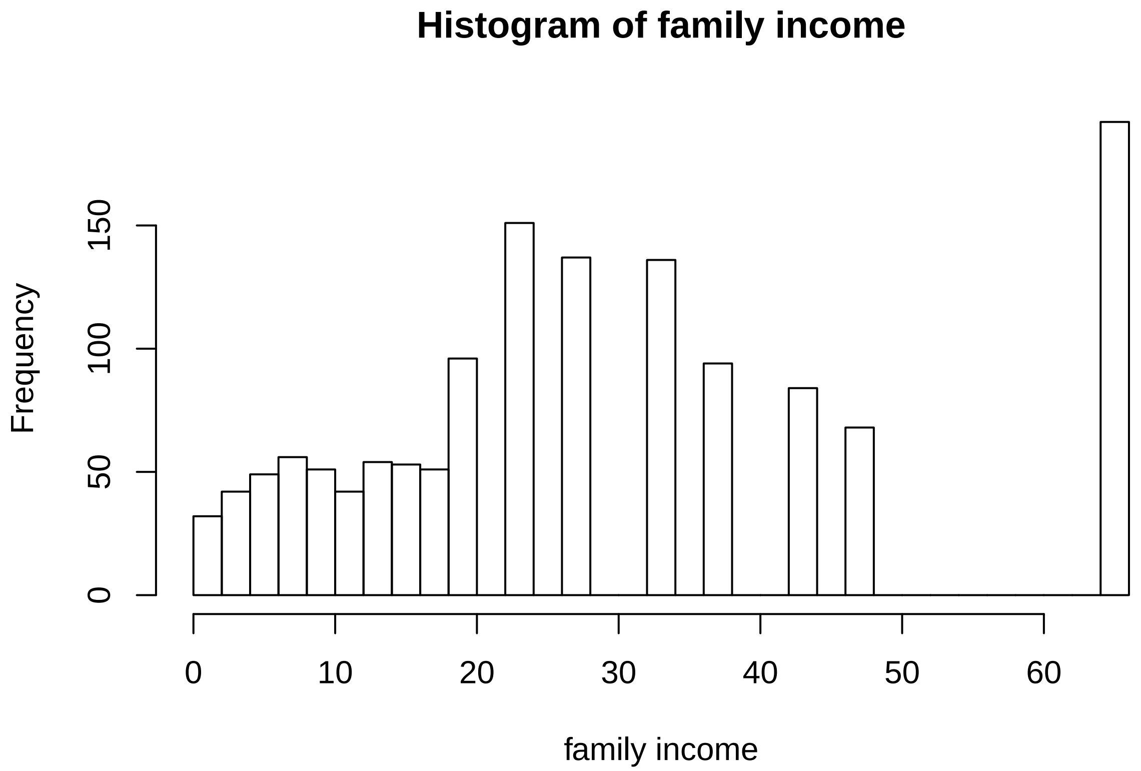
Since there is only one explanatory variable, we don't need to worry about collinearity. The Durbin Watson test is not statistically significant, so we can accept the null that this data is a random sample The residuals vs fitted graph shows that the expected error is 0 for almost all values of the explanatory variable. Although the expected error goes positive for the lower values of bwght, this may just be because there is not as much data in this area.

Question 5: Multiple Linear Regression

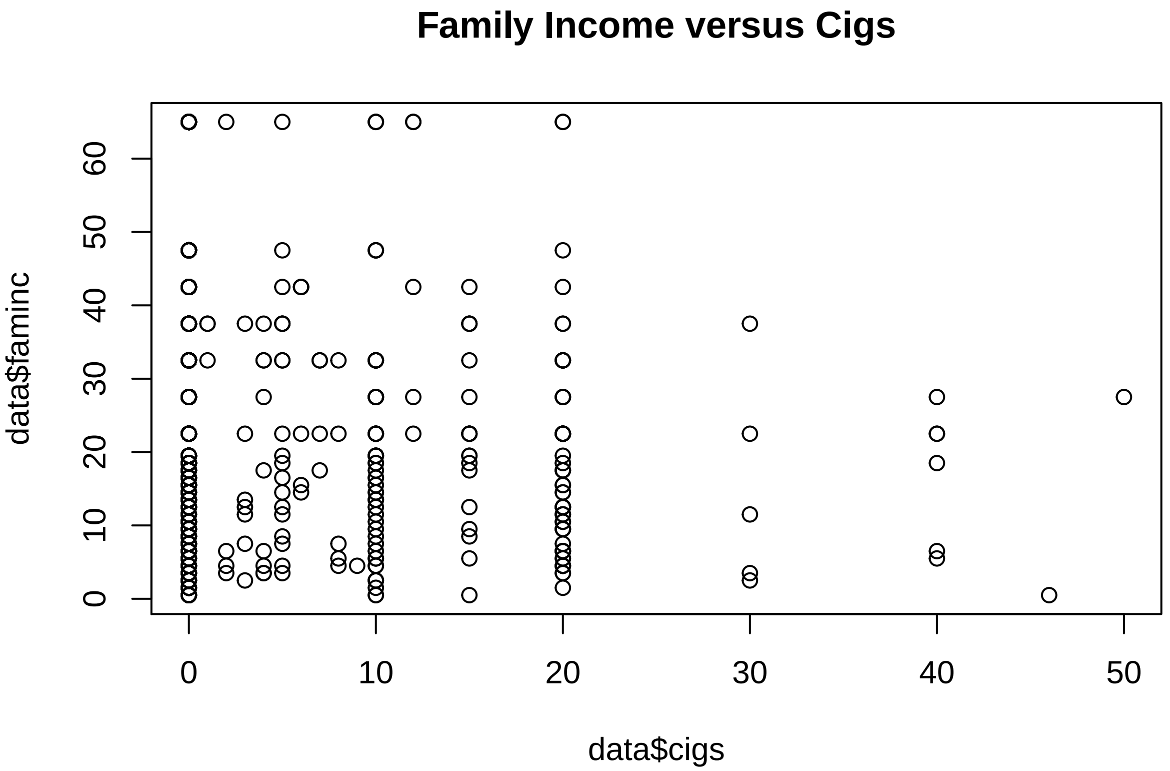
Despite child birth weight could be a function of mother's smoking behavior during pregnancy, other factors may influence child birth weight. As an exercise, let's introduce a new explanatory variable, *faminc*, representing family income in thousands of dollars and is serving as a proxy of other variables not captured in the model. Examine this variable (as well as its relationship with other variables in the model) using the same EDA as above. Note: In general, we will attempt multiple regression model specifications, test each one of them, and conduct model selection. We will study this between week 5 and 7.

```
EDA(data$faminc,"family income")
```

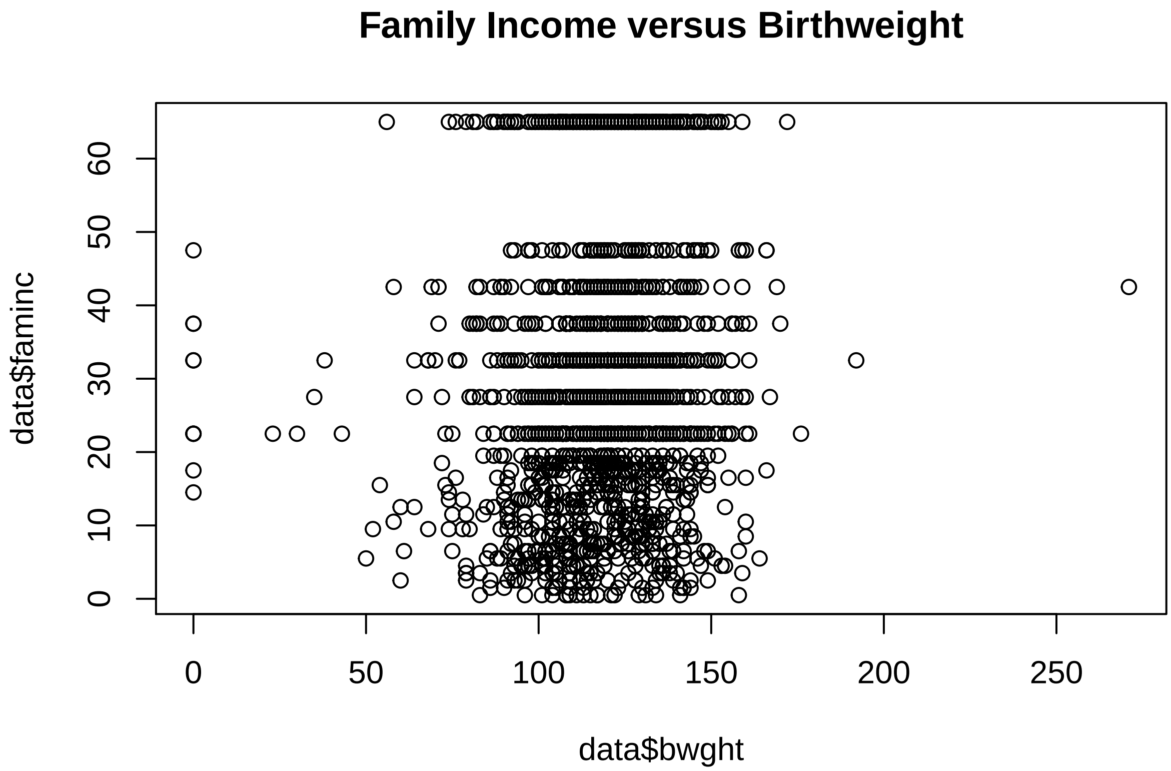
```
## [1] "Summary"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.50   14.50   27.50   29.03   37.50   65.00 
## [1] "Quantiles"
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 
##  0.5  3.5  6.5 14.5 27.5 37.5 65.0 65.0 65.0
```



```r
plot(data$cigs,data$faminc, main=paste("Family Income versus Cigs"))
```



```
plot(data$bwght,data$faminc, main=paste("Family Income versus Birthweight"))
```



The faminc variable is skewed right, and I suspect that there must be some data quality issues. I initially thought that the faminc variable must have been top coded, but the fact that there are no mothers between 50 and 60 thousand suggests that the skew is probably a result of incorrect recording.

There appears to be a somewhat negative relationship between cigarettes smoked per day and family income. But it's not perfect collinearity, so we can still use linear regression.

Regress *bwght* on both *cigs* and *faminc*. What coefficient estimates and the standard errors associated with the coefficient estimates do you get? Interpret the results. Again, we will study model specification in week 55. For now, just focus on the interpretation of the model results.

```
m <- lm(bwght ~ cigs + faminc, data = data[data$bwght>0,])
summary(m)
```

```
##
## Call:
## lm(formula = bwght ~ cigs + faminc, data = data[data$bwght >
##     0, ])
##
## 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
```

Explain, in your own words, what the coefficient on *cigs* in the multiple regression means, and how it is different than the coefficient on *cigs* in the simple regression? Please provide the intuition to explain the difference, if any.

The coefficient on cigs is the amount we can expect birthweight to decrease if a mother starts smoking one more cigarette a day, holding everything else constant. It's not as large in magnitude as the coefficient from the previous regression because some of the variation in birthweight is no explained by family income.

Question 6: Regression Diagnostic

Conduct regression diagnostic of the above model.

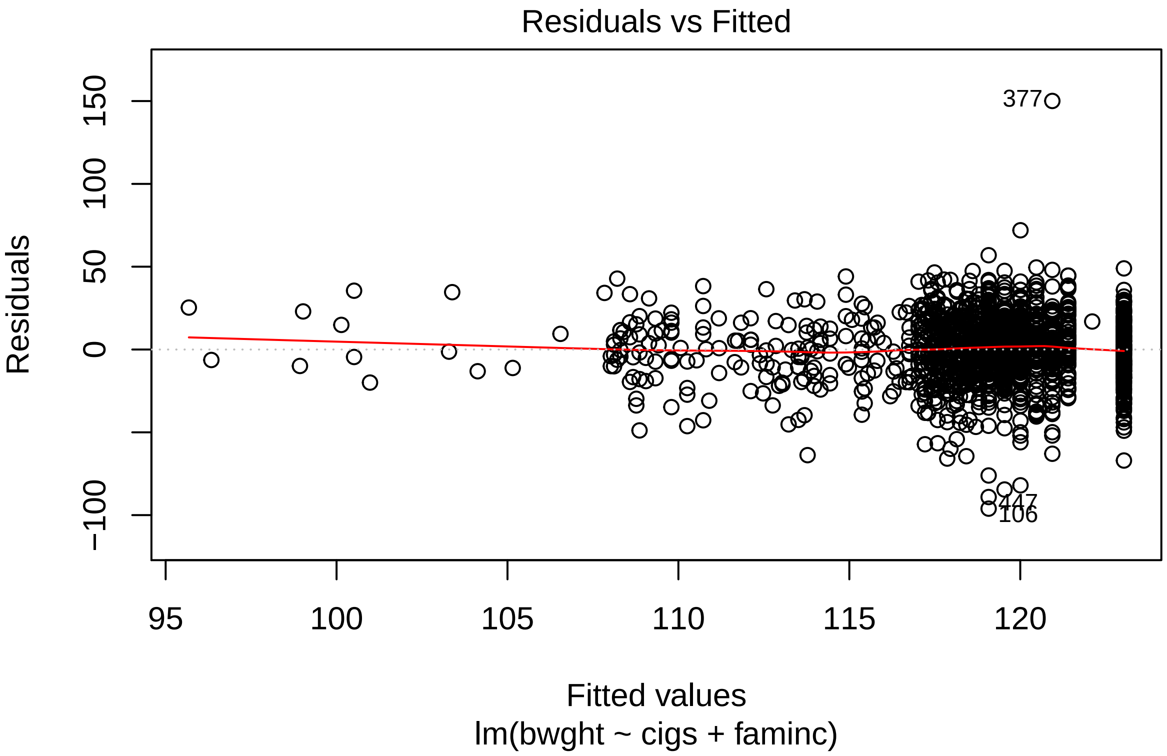
```r
library(car)
print("Durbin Watson")
```

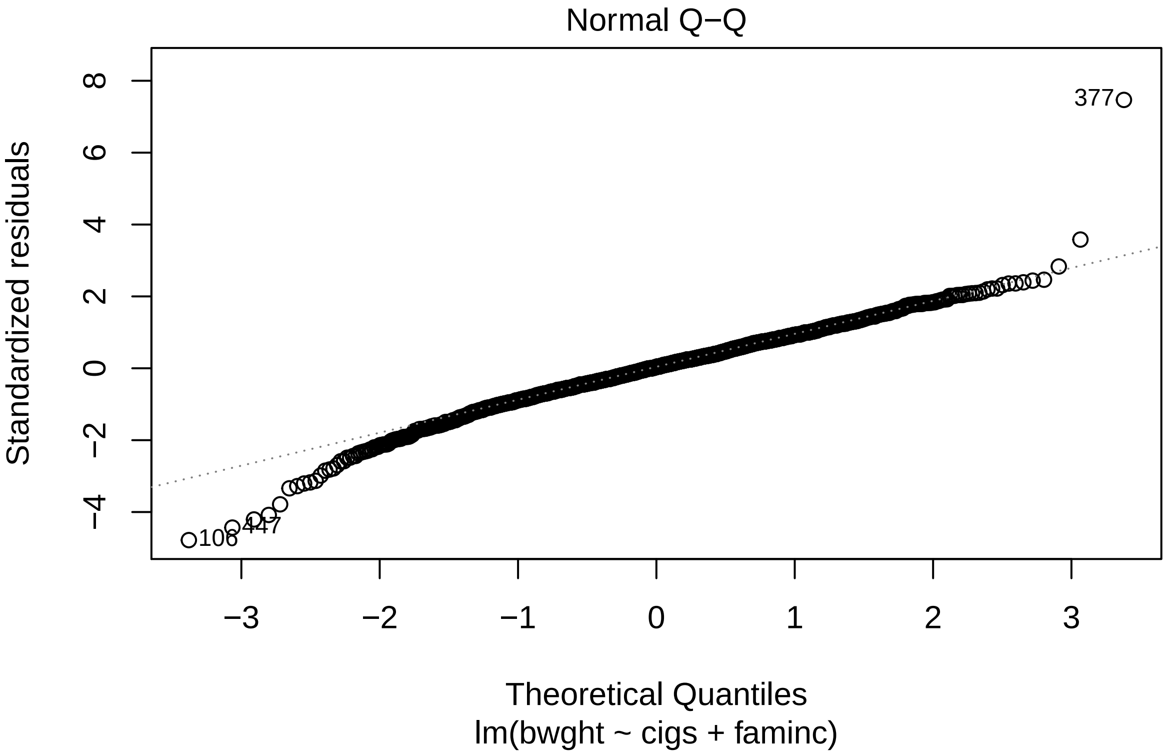
```
## [1] "Durbin Watson"
```

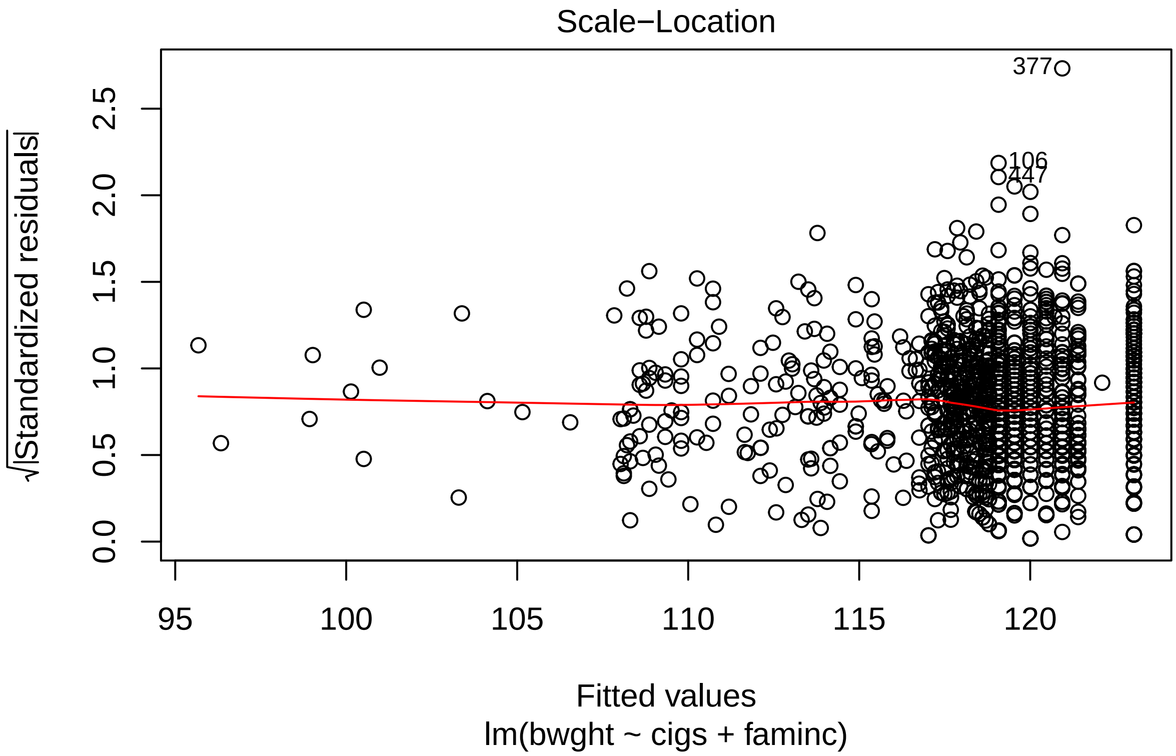
```r
durbinWatsonTest(data$faminc)
```

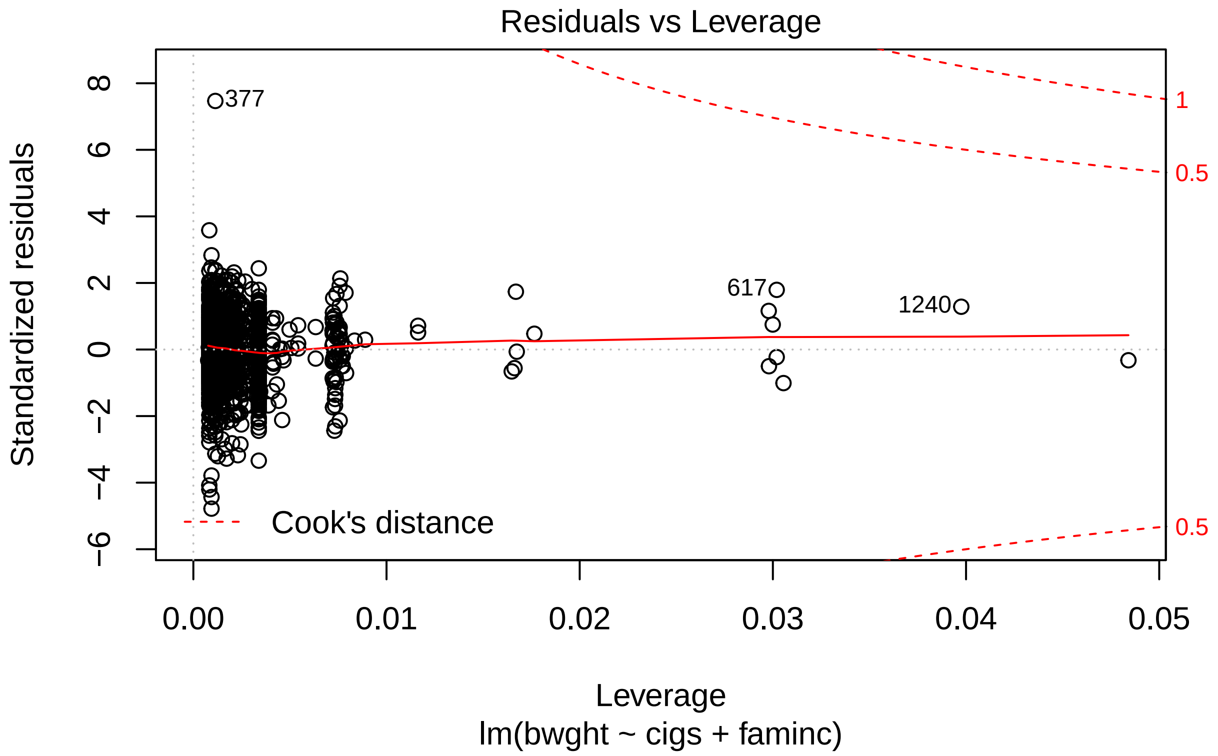
```
## [1] 0.5655137
```

```r
plot(m)
```


Residuals vs Fitted
Residuals
−100
0
50
100
150
95
100
105
110
115
120
Fitted values
lm(bwght ~ cigs + faminc)
377
447
106


Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
−4
377
106
447
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bwght ~ cigs + faminc)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
95
100
105
110
115
120
377
106
447
Fitted values
lm(bwght ~ cigs + faminc)



Neither faminc nor cigs show serial correlation according to the Durbin Watson test, so we can asume that they come from random samples. We also already showed in a previous problem that faminc and cigs are not perfectly collinear. Finally, the residuals vs fitted graph still shows that there is still no significant relationship between the values of the explanatory variables and the error.