W271-2 – Spring 2016 – **HW 1**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 2, 2016

Contents

Data **1**

Exercises **1**

Question 1: 1

Question 2: 2

Question 3: 3

Question 4: 8

Question 5: 13

Question 6: 14

Question 7: 15

Question 8: 17

Question 9: 18

Question 10: 18

Data

The file birthweight w271.RData contains data from the 1988 National Health Inter- view Survey, which may have been modified by the instructors to test your proficiency. This survey is conducted by the U.S. Census Bureau and has collected data on individ- ual 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 exercise, 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:

Load the birthweight dataset. Note that the actual data is provided in a data table named "data".

Use the following procedures to load the data:

Step 1: put the provided R Workspace birthweight w271.RData in the directory of your choice.

Step 2: Load the dataset using this command: load(\birthweight:Rdata)

```
##setwd('.\\MIDS\\Semester3\\W271\\Homework'1)
load("birthweight_w271.rdata")
```

Question 2:

Examine the basic structure of the data set using desc, str, and summary to examine all of the variables in the data set. How many variables and observations in the data? These commands will be useful:

1. desc
2. str(data)
3. summary(data)

```
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
```

```
desc
```

```
##    variable                          label
## 1    faminc     1988 family income, $1000s
## 2    cigtax   cig. tax in home state, 1988
## 3  cigprice cig. price in home state, 1988
## 4     bwght            birth weight, ounces
## 5  fatheduc            father's yrs of educ
## 6  motheduc            mother's yrs of educ
## 7    parity            birth order of child
## 8      male                =1 if male child
## 9     white                     =1 if white
## 10     cigs  cigs smked per day while preg
## 11   lbwght                   log of bwght
## 12 bwghtlbs           birth weight, pounds
## 13    packs packs smked per day while preg
## 14  lfaminc                    log(faminc)
```

```
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  
## 
```

Question 3:

As we mentioned in the live session, it is important to start with a question (or a hy- pothesis) 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. Ex- amine this variable using both tabulated summary and graphs. Specifcally,

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 dif- ferent 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.**

```
attach(data)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

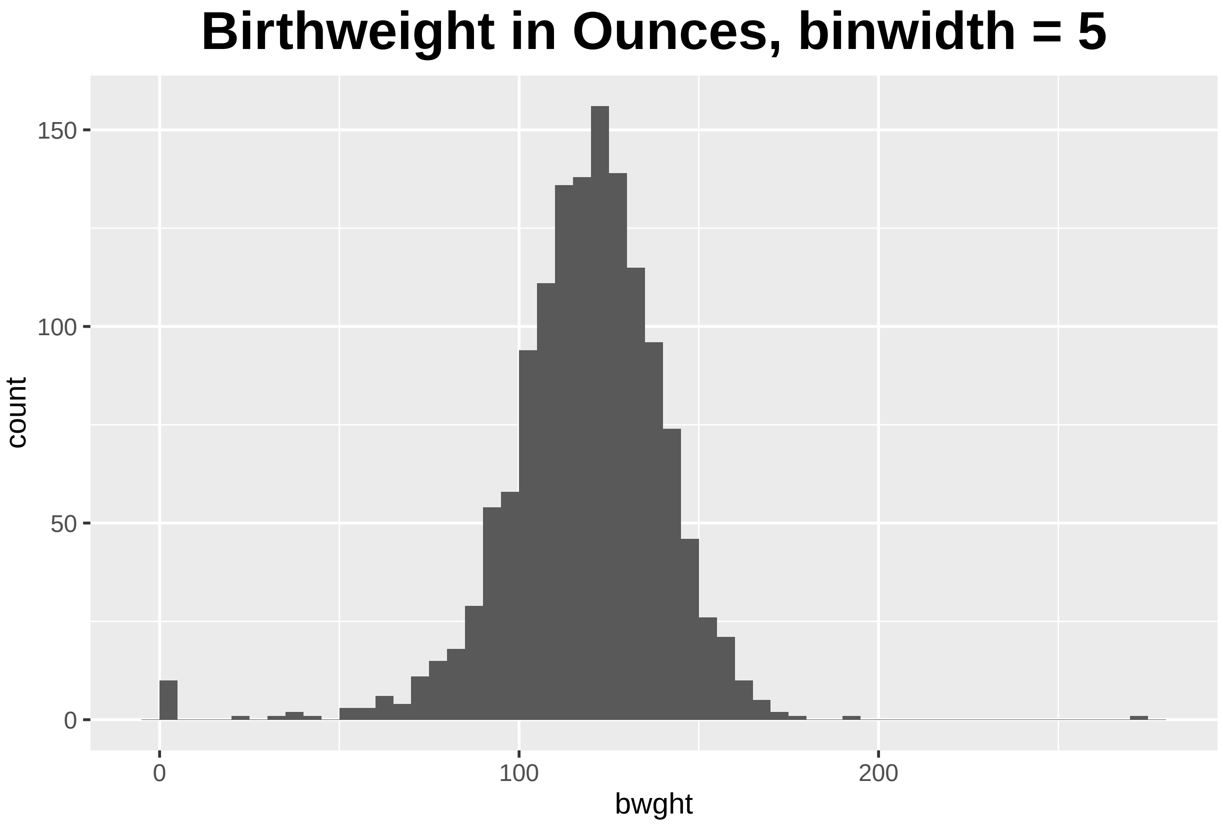
```
summary(bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   106.0   119.0   117.9   132.0   271.0
```

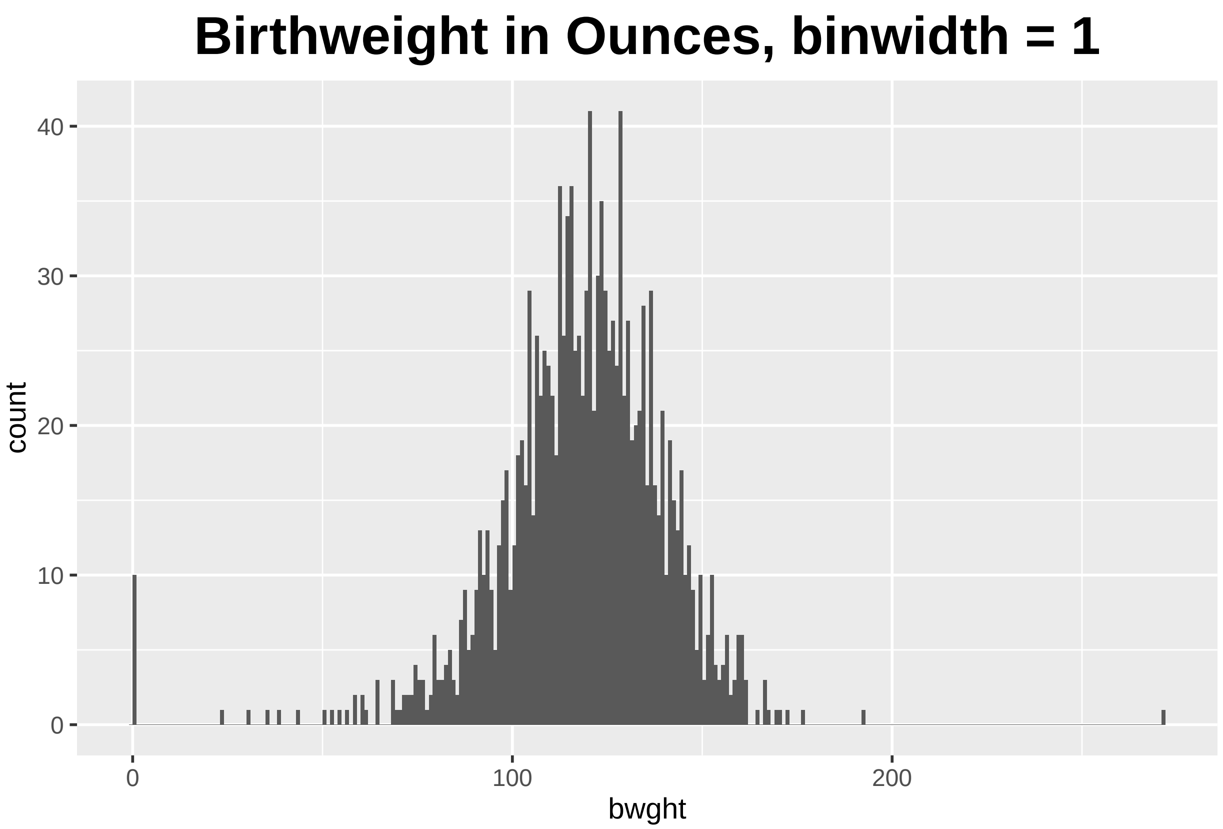
```
qnt <-  c(.01, .05, .1, .25, .5, .75, .9, .95, .99)
quantile(bwght, probs = qnt)
```

```
##     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
```

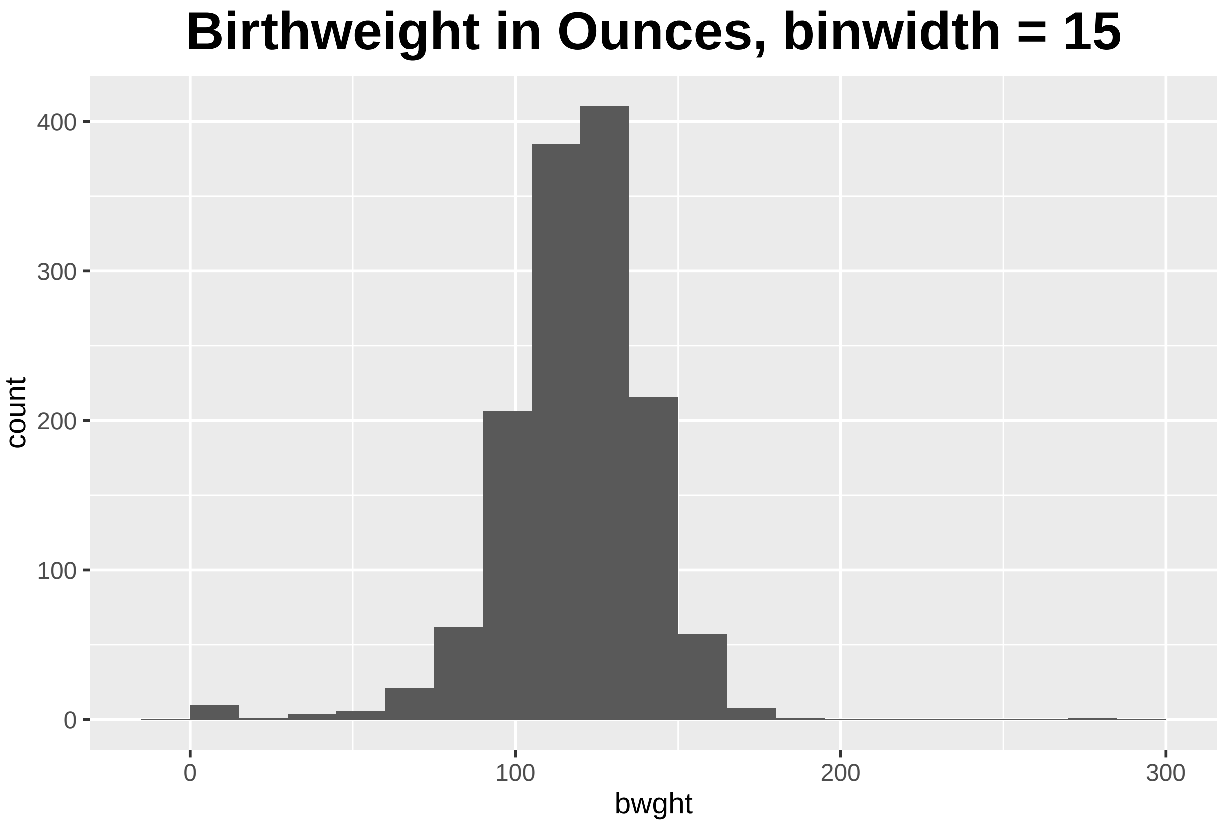
```
# Regular bin wdith
p1 <- ggplot(data, aes(bwght)) + geom_histogram(binwidth = 5) +
  ggtitle('Birthweight in Ounces, binwidth = 5')
p1 +theme(plot.title = element_text(size=20, face="bold",
    margin = margin(10, 0, 10, 0)))
```



```
#Narrow bin width
p2 <- ggplot(data, aes(bwght)) + geom_histogram(binwidth = 1) +
  ggtitle('Birthweight in Ounces, binwidth = 1')
p2 +theme(plot.title = element_text(size=20, face="bold",
    margin = margin(10, 0, 10, 0)))
```



```
#Wide bin width
p3 <- ggplot(data, aes(bwght)) + geom_histogram(binwidth = 15) +
  ggtitle('Birthweight in Ounces, binwidth = 15')
p3 +theme(plot.title = element_text(size=20, face="bold",
    margin = margin(10, 0, 10, 0)))
```



The bwght variable has the general shape of a normal distribution with some outliers on both sides of the distribution when the histogram is viewed at an average bin width. At very wide bin sizes, the shape is compressed and outliers are difficult to see. At very narrow bin sizes the distribution becomes very jagged, representing that some neighboring bwght values do not necessarily occur with similar frequency.

Based on the histogram, the primary values of interest are the birth weight values that seem implausably small and large. There is a spike in frequency at 0 ounces and a few observations above 250 ounces that merit further investigation.

```
outliers <- data[data$bwght < 10 | data$bwght > 200,]
outliers[,c("bwght", "lbwght", "bwghtlbs")]
```

```
##      bwght   lbwght bwghtlbs
## 85       0 0.000000   0.0000
## 128      0 0.000000   0.0000
## 230      0 0.000000   0.0000
## 352      0 0.000000   0.0000
## 377    271 5.602119  16.9375
## 567      0 0.000000   0.0000
## 570      0 0.000000   0.0000
## 730      0 0.000000   0.0000
## 834      0 0.000000   0.0000
## 1069     0 0.000000   0.0000
## 1255     0 0.000000   0.0000
```

Examing the various outlier points there doesn't seem to be any obvious evidence of errant data entry. Without knowing if 0 values represent mortality or missed observations, I would be hesitant to just throw them out of the dataset, but would consider running the analysis keeping these points in as well as without them. I would also reach out to the data provider to see if I could get clarification on the meaning of those points. Finally, I would remove the outlier at 271 oz because it is likely to have undue influence on the relationship between weight and cigarette smoking and is a true outlier in the sense that from a population sample these large, the odds of a baby having that birth weight are astronomically low.

Question 4:

Examine the variable cigs, which represents number of cigarettes smoked each day by the mother while pregnant. Conduct the same analysis as in question 3.

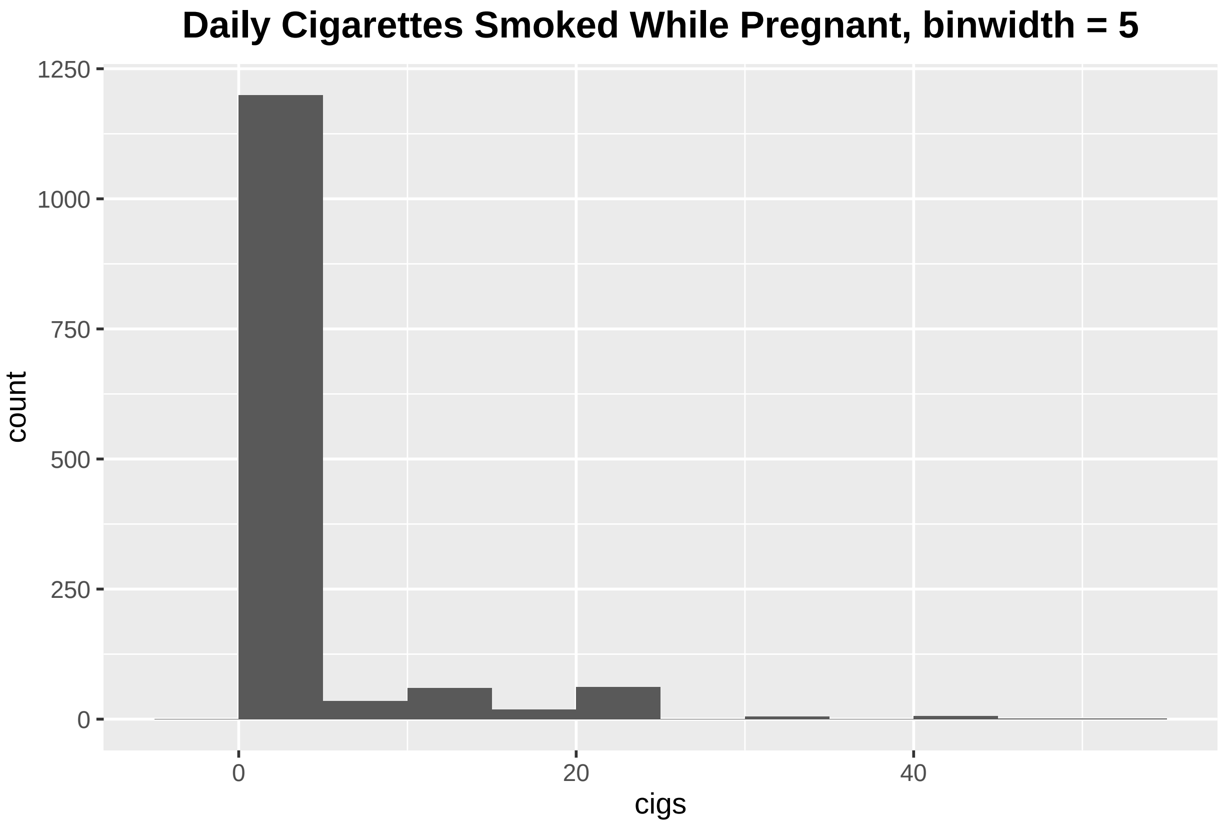
```
summary(cigs)
```

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

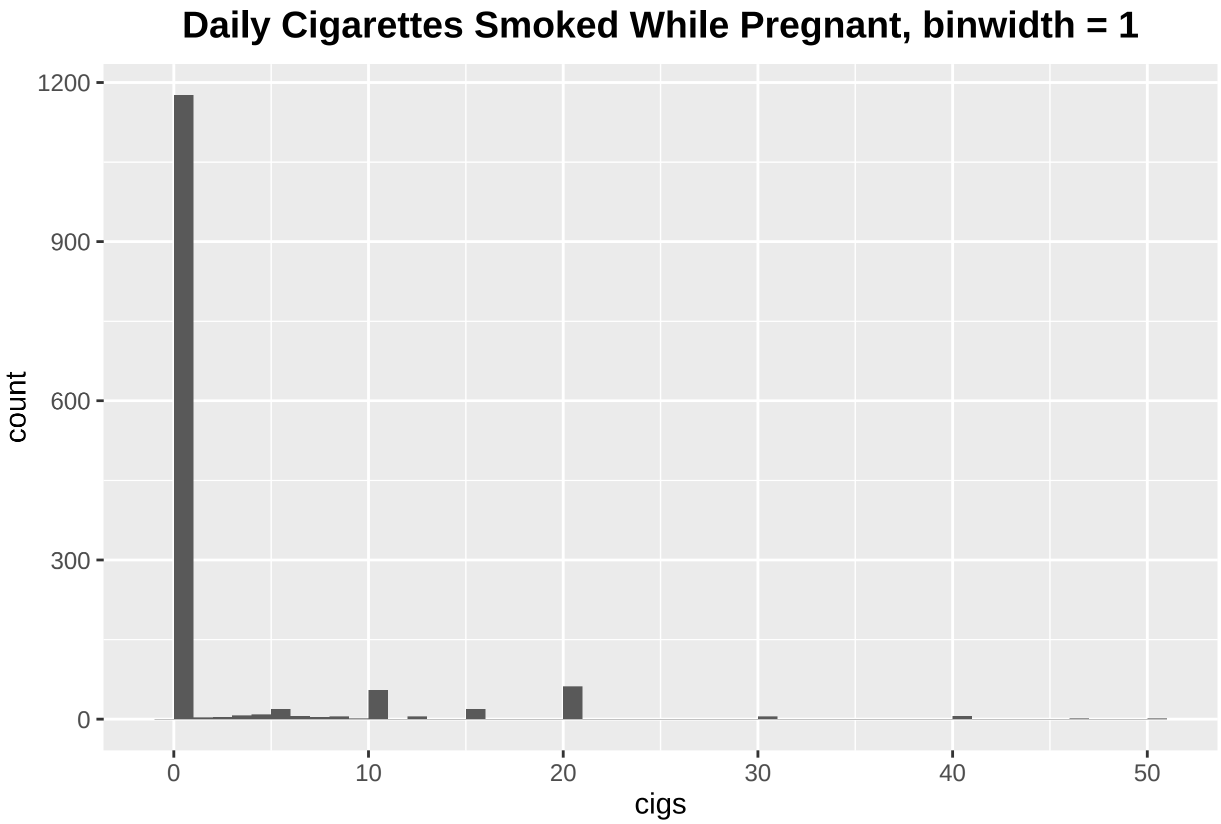
```
quantile(cigs, qnt)
```

```
##  1%  5% 10% 25% 50% 75% 90% 95% 99%
##   0   0   0   0   0   0  10  20  20
```

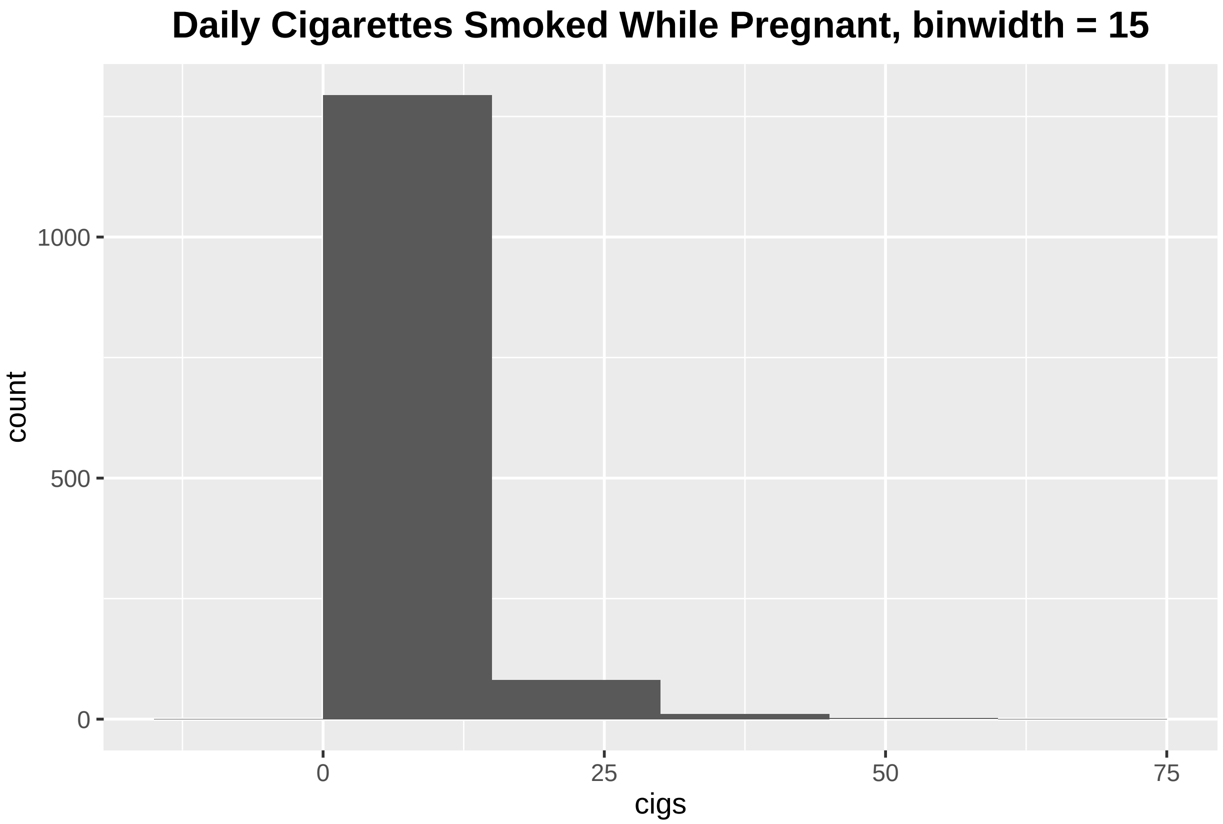
```
c1 <- ggplot(data, aes(cigs)) + geom_histogram(binwidth = 5) +
  ggtitle('Daily Cigarettes Smoked While Pregnant, binwidth = 5')
c1 +theme(plot.title = element_text(size=14, face="bold",
    margin = margin(10, 0, 10, 0)))
```



```
c2 <- ggplot(data, aes(cigs)) + geom_histogram(binwidth = 1) +
  ggtitle('Daily Cigarettes Smoked While Pregnant, binwidth = 1')
c2 +theme(plot.title = element_text(size=14, face="bold",
    margin = margin(10, 0, 10, 0)))
```

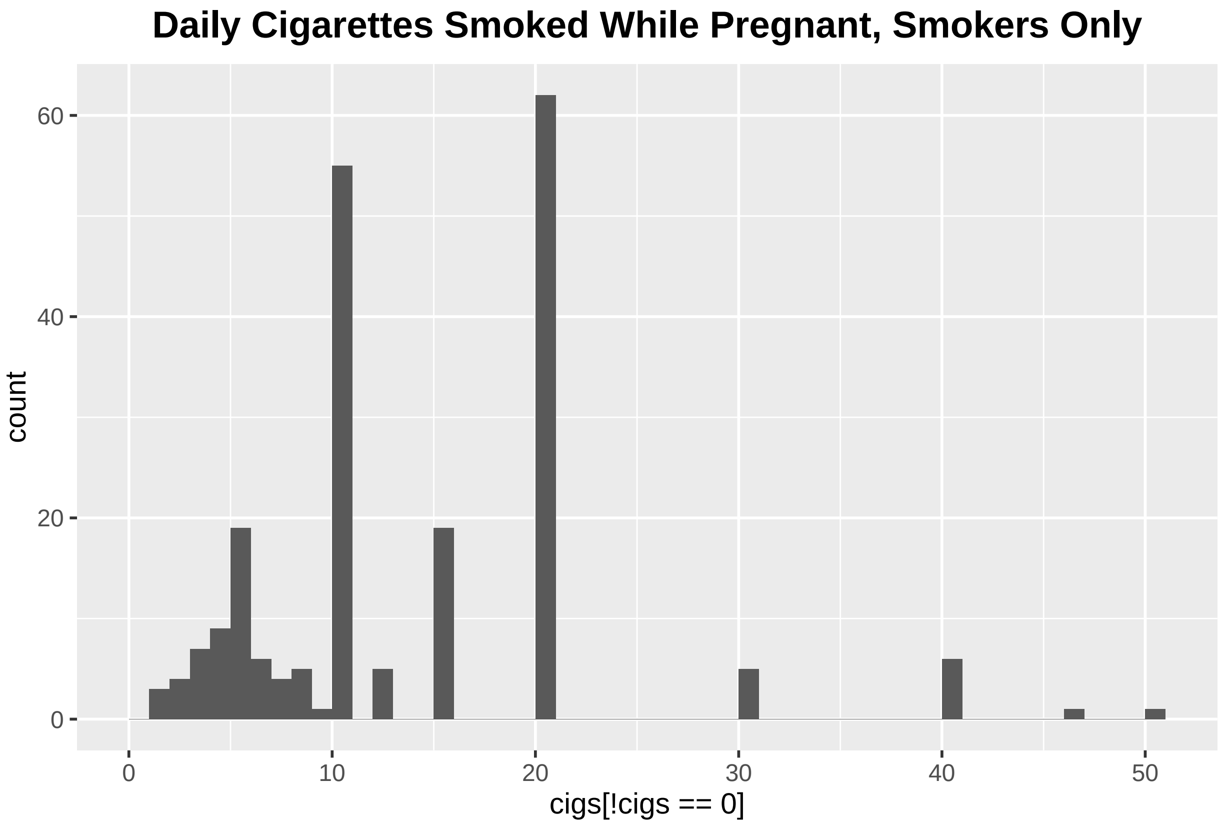


```
c3 <- ggplot(data, aes(cigs)) + geom_histogram(binwidth = 15) +
  ggtitle('Daily Cigarettes Smoked While Pregnant, binwidth = 15')
c3 +theme(plot.title = element_text(size=14, face="bold",
    margin = margin(10, 0, 10, 0)))
```

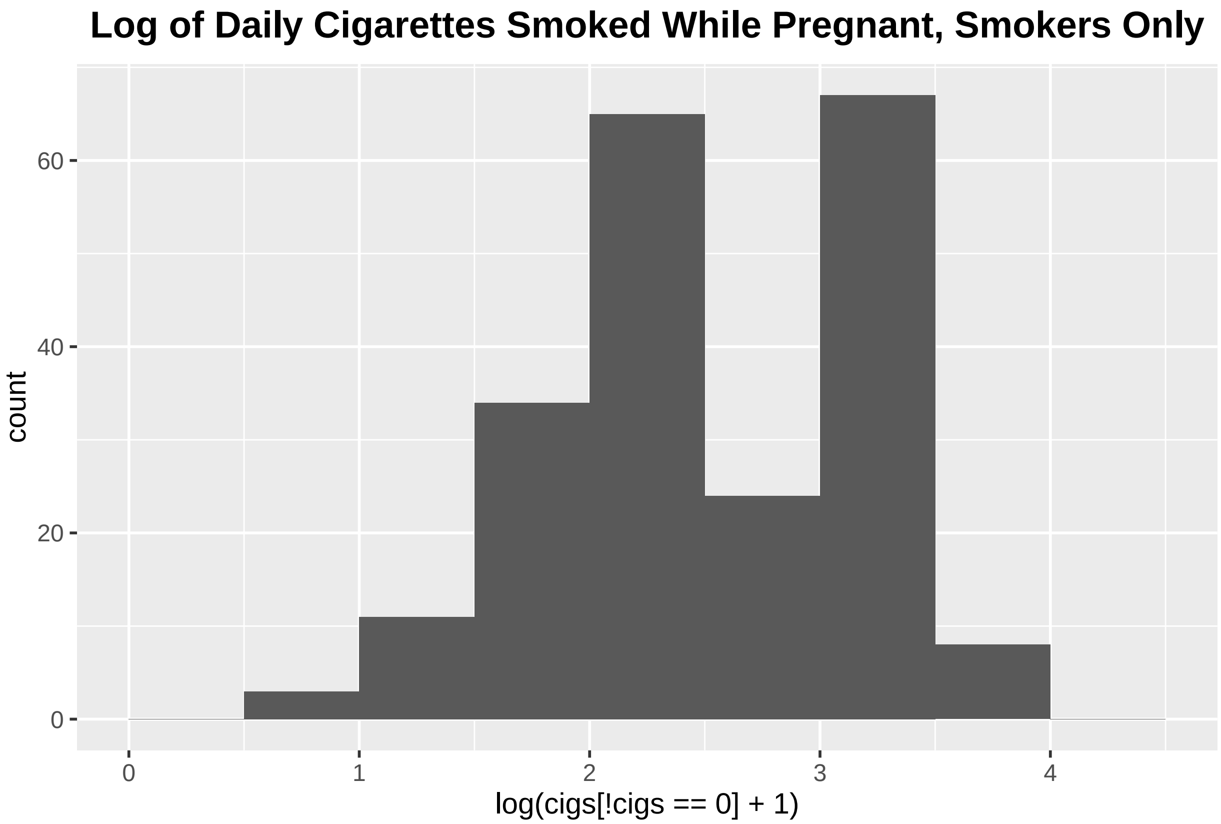


The histogram and quantiles of the cigs variable tell us that the vast majority of women in this sample did not smoke while pregnant. To better assess the shape of the distribution, it is more useful to look at the distribution among smokers.

```
c4 <- ggplot(data[!cigs==0,], aes(cigs[!cigs==0])) + geom_histogram(binwidth= 1) +
  ggtitle('Daily Cigarettes Smoked While Pregnant, Smokers Only')
c4 + theme(plot.title = element_text(size=14, face="bold",
    margin = margin(10, 0, 10, 0)))
```



```
c5 <- ggplot(data[!cigs==0,], aes(log(cigs[!cigs==0] + 1))) + geom_histogram(binwidth= 0.5) +
  ggtitle('Log of Daily Cigarettes Smoked While Pregnant, Smokers Only')
c5 + theme(plot.title = element_text(size=14, face="bold",
    margin = margin(10, 0, 10, 0)))
```



```r
#Hartigan's dip test for multimodality
require(diptest)
```

```
## Loading required package: diptest
```

```r
dip.test(log(cigs[!cigs==0] + 1))
```

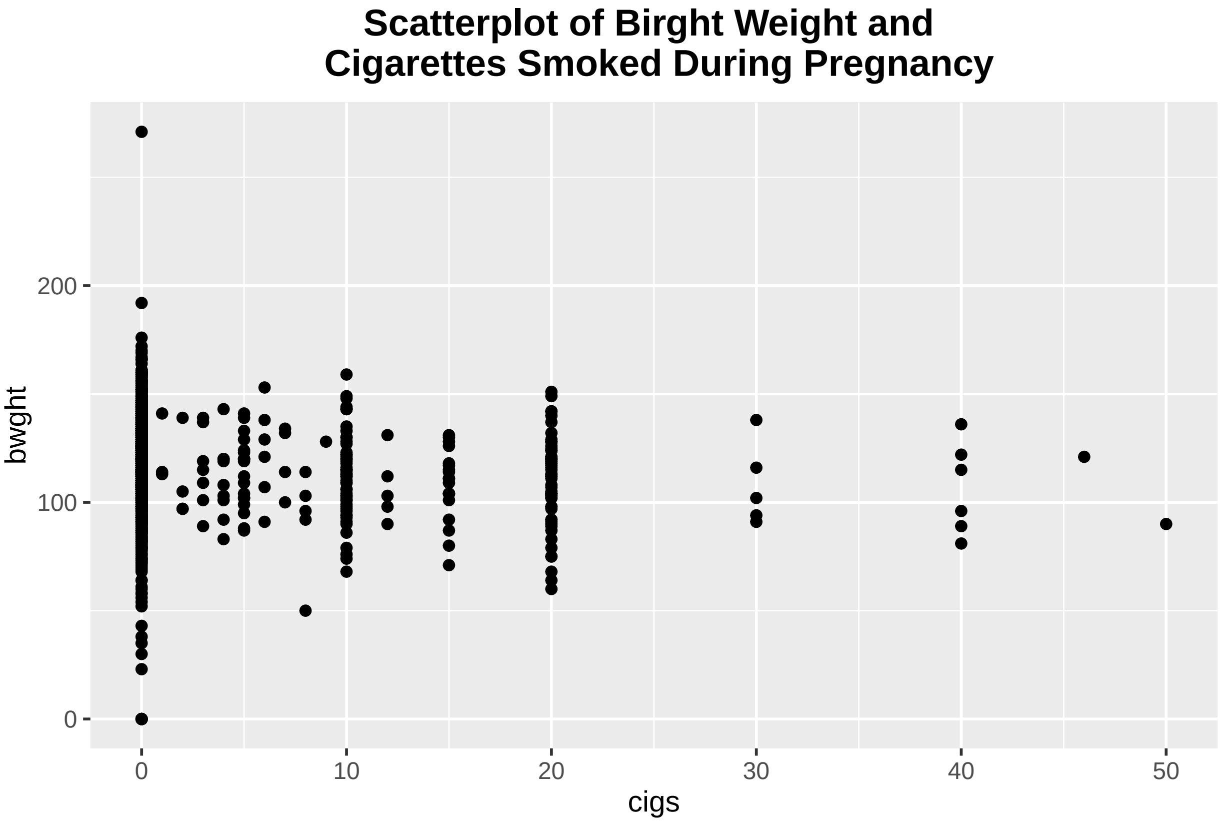
```
##
##  Hartigans' dip test for unimodality / multimodality
##
## data:  log(cigs[!cigs == 0] + 1)
## D = 0.12972, p-value < 2.2e-16
## alternative hypothesis: non-unimodal, i.e., at least bimodal
```

Among smokers, the distribution of cigarettes smoked is heavily right skewed. Log transformation gives the data a bimodel appearance.

Question 5:

Generate 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?

```
s <- ggplot(data, aes(cigs, bwght)) + geom_point() +
  ggtitle('Scatterplot of Birght Weight and \n Cigarettes Smoked During Pregnancy') +
  theme(plot.title = element_text(size=14, face="bold", margin = margin(10, 0, 10, 0)))
s
```



Looking at the scatterplot, there seems to be a small negative relationship between birth weight and cigarettes smoked during pregnancy. The relationship looks weak because there is still wide variation in birth weight at a given level of cigarette smoking, and thus the cigarettes probably account for a small share of the variation.

Question 6:

Estimate the simple linear regression of bwght on cigs. What coefficient estimates and the standard errors associated with the coefficient estimates do you get Interpret the results. Note that you may have to "take care of" any potential data issues before build- ing a regression model.

```
# Exclude any data where there is no observation for cigs or bwght
data = data[complete.cases(data$bwght, data$cigs),]
# Exclude the upper outlier for bwght
data = data[data$bwght < 200, ]
m <- lm(data$bwght~data$cigs)
summary.lm(m)
```

```
##
## Call:
```

```
## lm(formula = data$bwght ~ data$cigs)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -118.698  -11.121    1.302   14.302   73.302 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 118.69796    0.62990 188.439  < 2e-16 ***
## data$cigs    -0.45774    0.09956  -4.598 4.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 22.14 on 1385 degrees of freedom
## Multiple R-squared:  0.01503,    Adjusted R-squared:  0.01432 
## F-statistic: 21.14 on 1 and 1385 DF,  p-value: 4.659e-06
```

Regression showed a small negative effect of maternal cigarette smoking on birthweight (β_1 = -0.45 (0.10), p < .001, R^2 = 0.014). This represents a practically small but not meaningless effect. For example, among smokers, the average daily cigarettes smoked is 13.7. Thus, the mean cigarette smoker would have a 6 Oz. lower expected birth weight, other factors held constant.

Question 7:

Now, introduce a new independent variable, faminc, representing family income in thou- sands of dollars. Examine this variable using the same analysis as in question 3. In addition, produce a scatterplot matrix of bwght, cigs, and faminc. Use the following command (as a starting point):

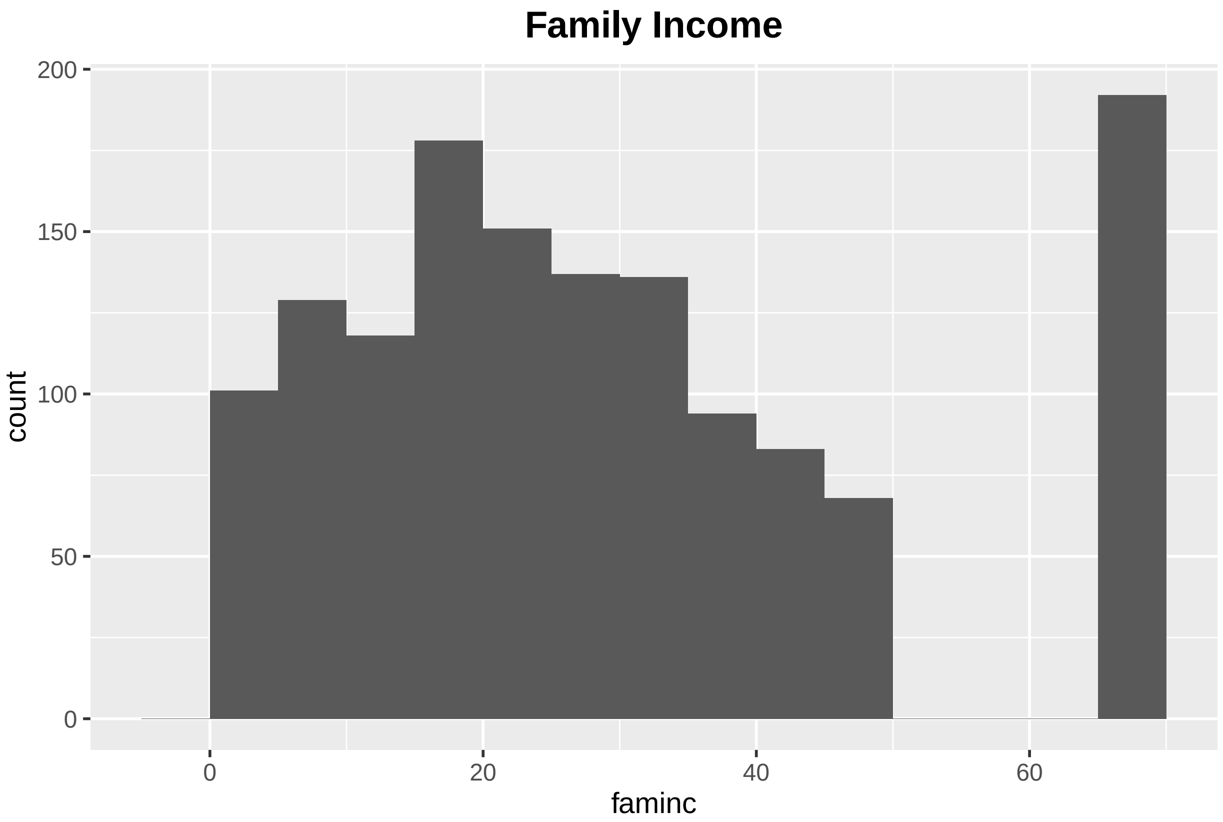
```
summary(faminc)
```

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

```
quantile(faminc, qnt)
```

```
##   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 
```

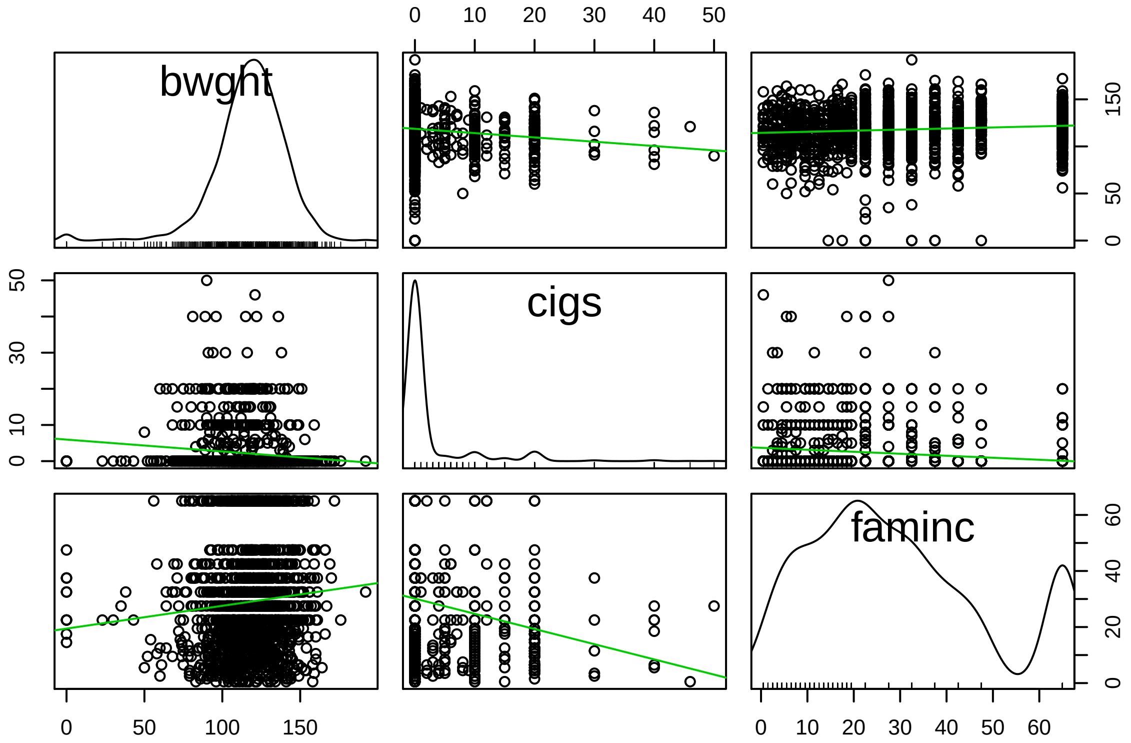
```
h <- ggplot(data, aes(faminc)) + geom_histogram(binwidth= 5) +
  ggtitle('Family Income') +
  theme(plot.title = element_text(size=14, face="bold", margin = margin(10, 0, 10, 0)))
h
```



```
require(car)
```

```
## Loading required package: car
```

```
scatterplotMatrix(data[, c("bwght", "cigs", "faminc")], smoother = F)
```



Question 8:

Regress bwgth on both cigs and faminc. What coefficient estimates and the standard errors associated with the coefficient estimates do you get? Interpret the results.

```
data = data[complete.cases(data$bwght, data$cigs, data$faminc),]
m2 <- lm(bwght~cigs + faminc, data = data)
summary.lm(m2)
```

```
##
## Call:
## lm(formula = bwght ~ cigs + faminc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.306  -10.929    1.382   14.050   73.084
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 115.90312    1.15475 100.371  < 2e-16 ***
## cigs         -0.40746    0.10081  -4.042  5.6e-05 ***
## faminc        0.09270    0.03214   2.885  0.00398 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.09 on 1384 degrees of freedom
## Multiple R-squared:  0.02092,    Adjusted R-squared:  0.01951
## F-statistic: 14.79 on 2 and 1384 DF,  p-value: 4.427e-07
```

Regression showed that maternal cigarette smoking had a small negative association with birth weight and family income had a small positive association with birth weight (β_1 = -.41 (.10), P<.001), β_2 = .09 (.03), p=.004, R^2 = .02). The effect of income on birth weight is practically very small, as moving from the median income in the sample to the 95th percentile would only increase expected birth rate by 3.5 Oz. The effect of smoking is more practically significant as the median smoking mother would have an expected birth weight about .4 Oz lower than a non smoker, and a smoker in the 95% percentile would have a 12 Oz. decrease in expected birth weight.

Question 9:

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.

In the multiple regression the coefficient represents the association of maternal smoking on birth weight, holding familiy income constant. This differs from the simple regression as that coefficient represents the association of cigarettes smoked and birth weight without holding any other measured variables constant.

Question 10:

Which coefficient for cigs is more negative than the other? Suggest an explanation for why this is so.

The coefficient in the simple regression model is more negative. An explaination for this is that familiy income also has a negative relationship with cigarettes smoked, and thus some of the variation that was accounted for by only cigarettes smoked in the simple model is accounted for by family income in the multiple regression model, lowering the coefficient for cigarettes smoked.