W271-2 – Spring 2016 – **HW 1**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 3, 2016

Contents

The file birthweight w271.RData contains data from the 1988 National Health Interview Survey, which may have been modifed by the instructors to test your profciency. 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 exercise, however, you are to treat the data as a true random sample. You will use this dataset to practice interpreting OLS coeffcients.

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")`

```
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)`

```
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)
```

```
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
```

```
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
##
```

As shown by `desc` and `str(data)`, there are 14 variables and 1388 observations in the data.

Question 3

As we mentioned in the live session, it is important to start with a question (or a hypothesis) when conducting regression modeling. In this exercise, 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 summary and graphs. Specifically,

1. **Summarize the variable `bwght`: `summary(data$bwght)`**

```
summary(data$bwght)
```

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

2. **You may also use the `quantile` function: `quantile(data$bwght)`. List the following quantiles: 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%**

```
quantile(data$bwght, probs = c(1, 5, 10, 25, 50, 75, 90, 95, 99)/100)
```

```
##     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
```

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.**

We tested several bin widths, though here only three (5, 10, and 20) are plotted—they're enough to show that the smaller the bin size, the closer the histogram looks to the density plot (which is close to the normal distribution—except for a long left tail—in this case).

The first bin size (5) is plotted below using `hist` and `ggplot`. The rest are plotted using `ggplot` exclusively.

```
# Use hist and bin width = 5
bin_width = 5
hist(data$bwght, breaks = seq(floor(min(data$bwght)/bin_width)*bin_width,
                              ceiling(max(data$bwght)/bin_width)*bin_width,
                              by = bin_width),
     xlab = "Birth weight (ounces)", ylab = "Count",
     main = "Histogram of birth weight")
```

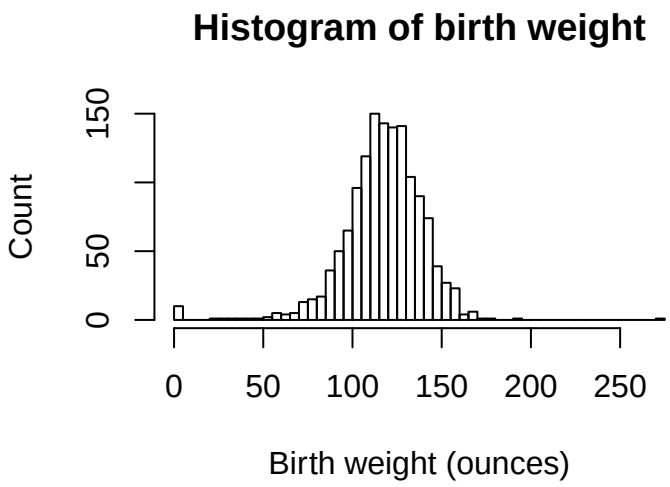


Figure 1: Histogram of birth weight (in ounces), using `hist` and bin width = 5

```
# Use ggplot and bin width = 5
ggplot(data = data, aes(bwght)) +
  geom_histogram(colour = 'black', fill = 'white',
                 binwidth = bin_width) +
  labs(x = "Birth weight (ounces)", y = "Count",
       title = "Histogram of birth weight")
```

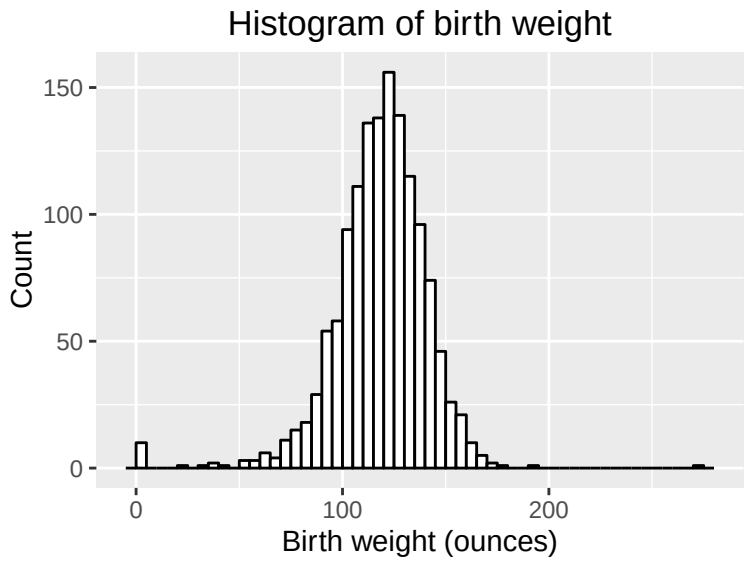


Figure 2: Histogram of birth weight (in ounces), using `ggplot` and bin width = 5

```
# Use ggplot and bin width = 10
bin_width = 10
```

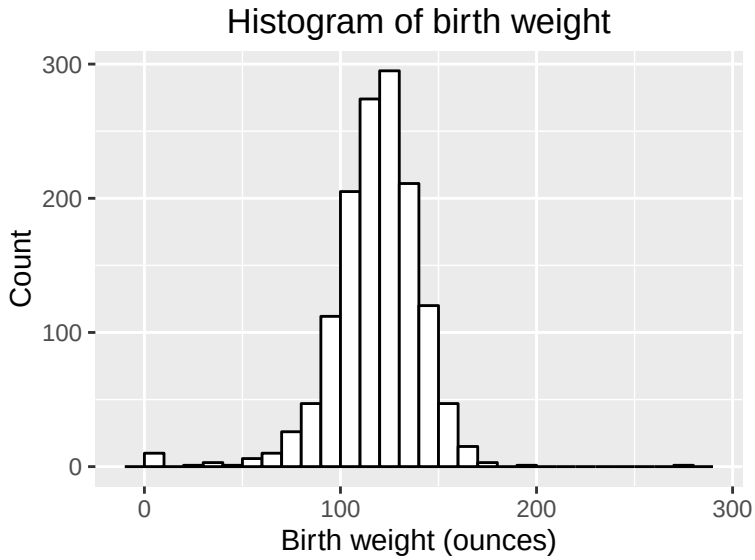


Figure 3: Histogram of birth weight (in ounces), using `ggplot` and bin width = 10

```
# Use ggplot and bin width = 20
bin_width = 20
```

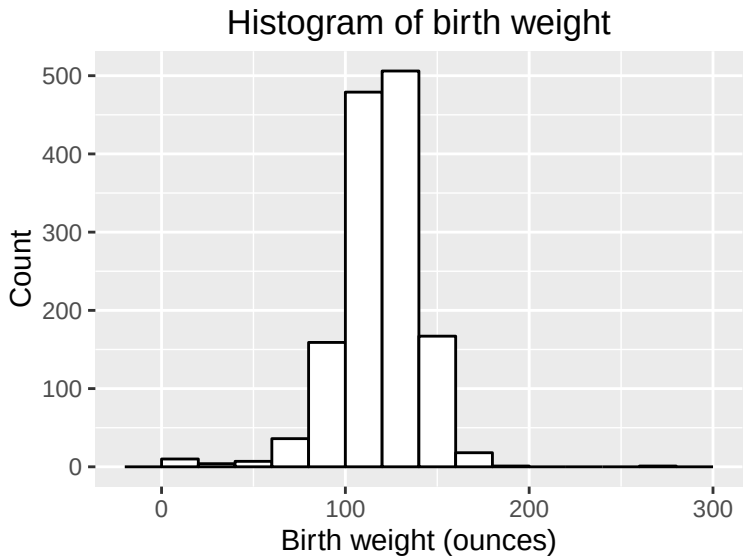


Figure 4: Histogram of birth weight (in ounces), using `ggplot` and bin width = 20

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.**

The left tail of the distribution is quite long for such variable. Actually, there are 10 observations with a weight equal to zero, which could code for missing values or could code for mortality. As these observations

are outside of the influence of cigarette smoking on birth weight, they could be excluded. If we exclude those observations, the minimum birth weight is 23 ounces, which still seems very low but might be possible. 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 this large, the odds of a baby having at that birth weight are astronomically low.

There are no `NA` values for `data$bwght` so it seems likely that missing values have been coded as `0`.

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(data$cigs)
```

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

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

```
# Use ggplot and bin width = 1
ggplot(data = data, aes(cigs)) +
  geom_histogram(colour = 'black', fill = 'white',
                 binwidth = bin_width) +
  labs(x = "Cigarettes smoked each day by the mother while pregnant",
       y = "Count",
       title = "Histogram of cigarettes smoked each day\nby the mother while pregnant")
```

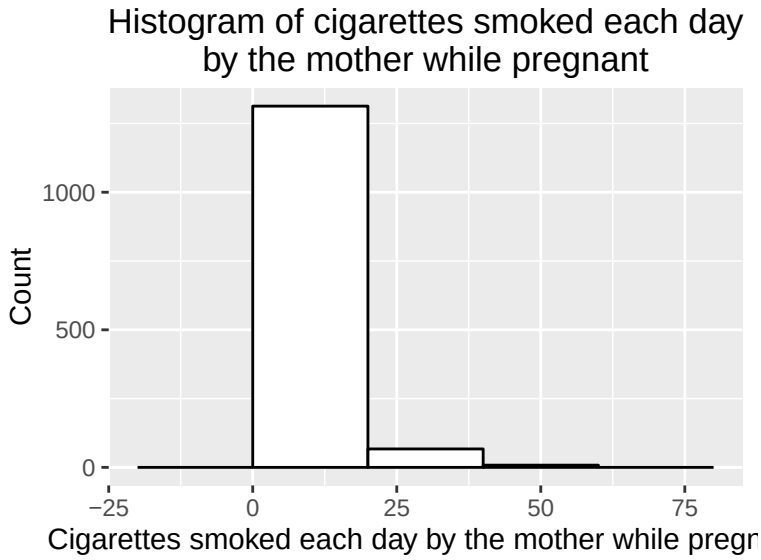


Figure 5: Histogram of cigarettes smoked each day by the mother while pregnant, using `ggplot` and bin width = 1

```
# Use ggplot and bin width = 5
bin_width = 5
```

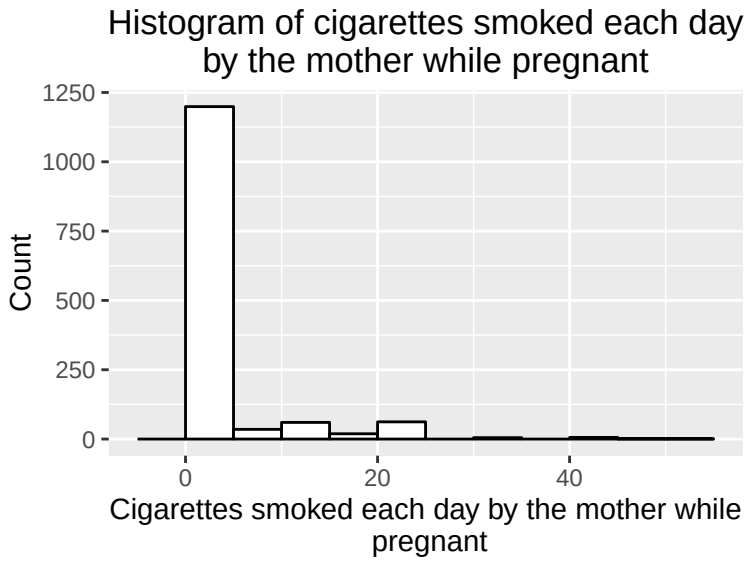


Figure 6: Histogram of cigarettes smoked each day by the mother while pregnant, using `ggplot` and bin width = 5

```
# Use ggplot and bin width = 10
bin_width = 10
```

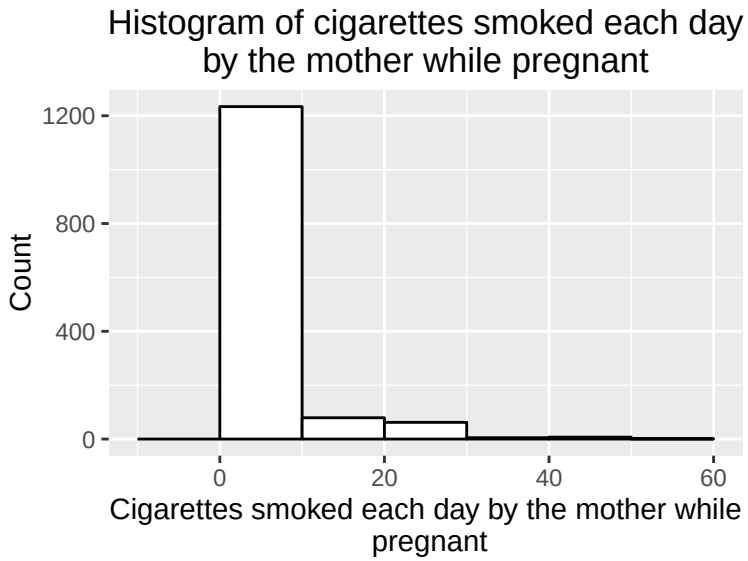


Figure 7: Histogram of cigarettes smoked each day by the mother while pregnant, using `ggplot` and bin width = 10

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.

```
# Use ggplot and bin width = 5, smokers only
bin_width =5
```

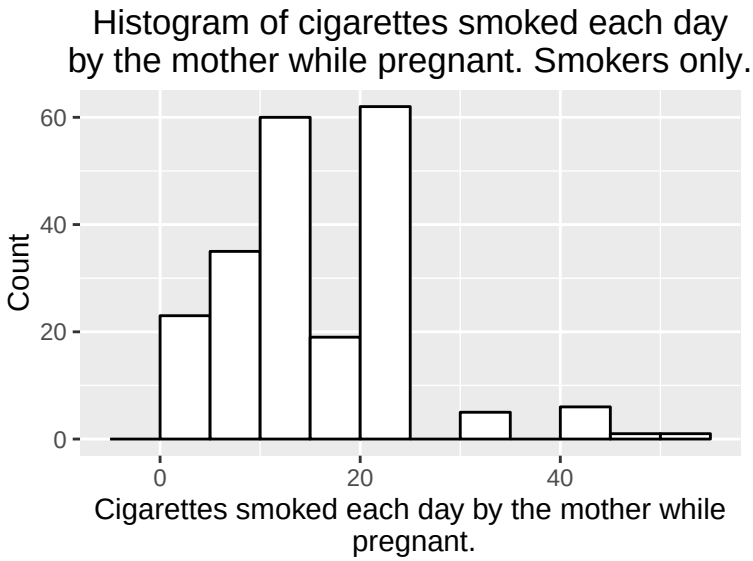


Figure 8: Histogram of cigarettes smoked each day among smokers by the mother while pregnant, using `ggplot` and bin width = 5

```
# Use ggplot and bin width = 0.5, smokers only
bin_width =0.5
```

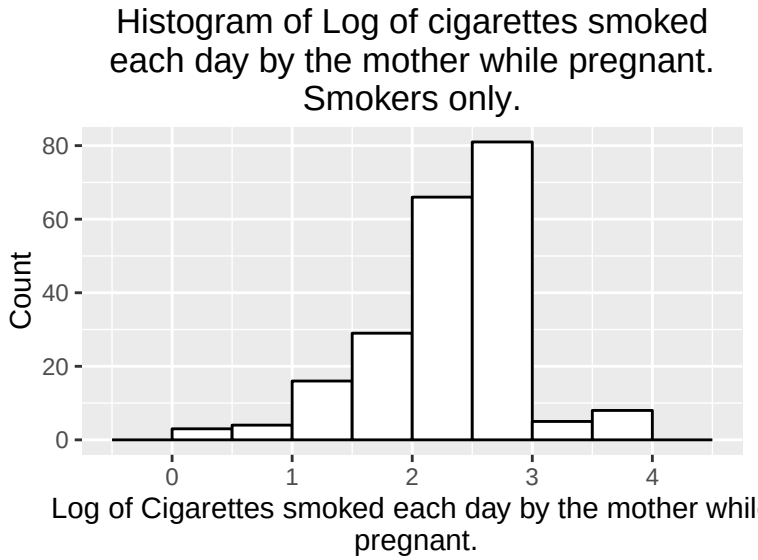


Figure 9: Histogram of Log of cigarettes smoked each day among smokers by the mother while pregnant, using `ggplot` and bin width = 0.5

Among smokers, the distribution of cigarettes smoked is right skewed. Log transformation gives the data a more approximately normal appearance. Log transformation could be considered for the cigs variatble, but given that the resulting variable is still non normal and would make interpretation of the model less clear, using the non-transformed variable seems more appropriate.

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`?

```
# Use ggplot
```

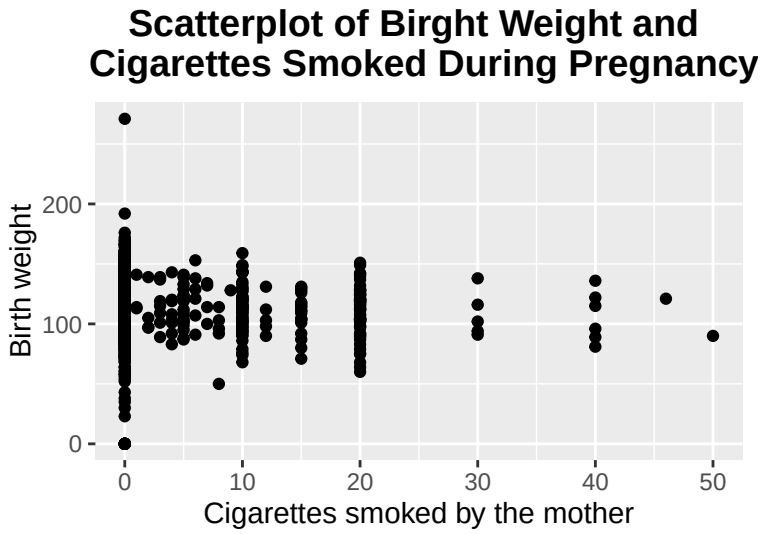


Figure 10: Scatterplot of birth weight and cigarettes smoked each day by the mother.

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 only 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 building a regression model.

```
# Exclude any data where there is no observation for cigs or bwght. Exclude outliers and 0s.
```

```
##
## Call:
## lm(formula = data$bwght ~ data$cigs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.666 -11.666   0.416  13.334  72.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119.6663     0.5645 211.989  < 2e-16 ***
## data$cigs    -0.5083     0.0889  -5.717 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 19.76 on 1375 degrees of freedom
## Multiple R-squared:  0.02322,    Adjusted R-squared:  0.02251
## F-statistic: 32.69 on 1 and 1375 DF,  p-value: 1.324e-08
```

Regression showed a small negative effect of maternal cigarette smoking on birthweight (β_1 = -0.51 (0.09), p < .001, R^2 = 0.02). 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 thousands 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):

```
library(car)
scatterplot:matrix( bwght + cigs + faminc; data = data2)
```

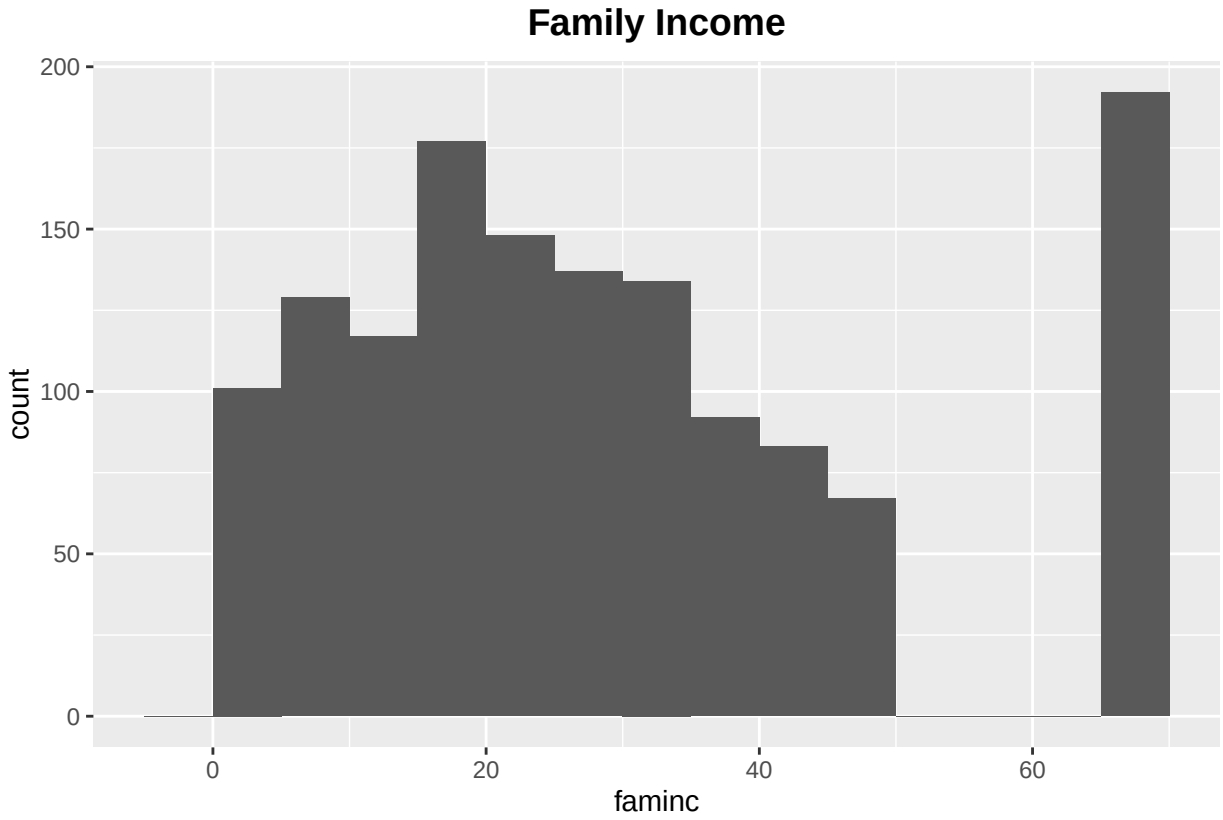
Note that the car package is needed in order to use the `scatterplot.matrix` function.

```
#scatterplotMatrix in car. Show linear trend lines.
summary(data$faminc)
```

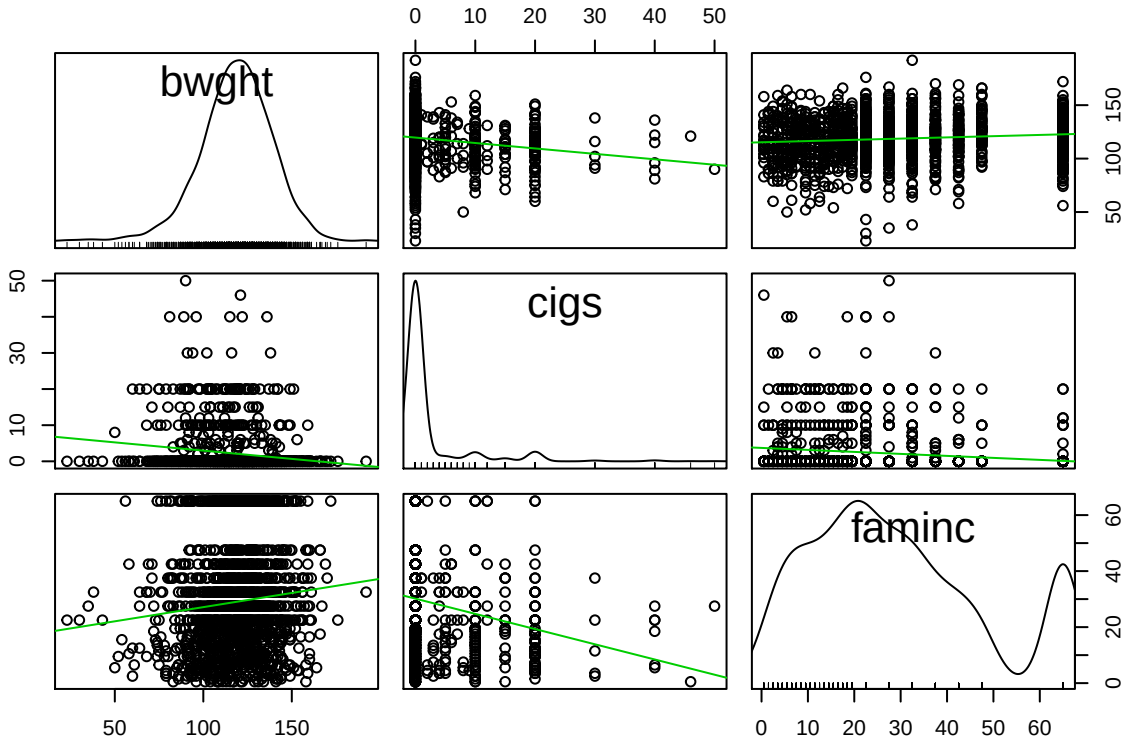
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50   14.50   27.50   29.02   37.50   65.00
```

```
quantile(data$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
```




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



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
## -95.983 -11.537   0.824  13.298  72.125
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.97540    1.03242 113.302  < 2e-16 ***
## cigs         -0.45981    0.08998  -5.110 3.67e-07 ***
## faminc        0.08921    0.02870   3.109  0.00192 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.7 on 1374 degrees of freedom
## Multiple R-squared:  0.03004,    Adjusted R-squared:  0.02863
## F-statistic: 21.28 on 2 and 1374 DF,  p-value: 7.916e-10
```

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 = -.46 (.09), P<.001), β_2 = .09 (.03), p=.002, R^2 = .03). The effect of income on birth weight is practically very small, as moving from the median income in the sample to the 95^{th} 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 about 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 coeffcient 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.