W271-2 – Spring 2016 – **HW 1**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 3, 2016

Contents

Data

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.

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")`**

```
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)
```

```r
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
```

```r
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")
```

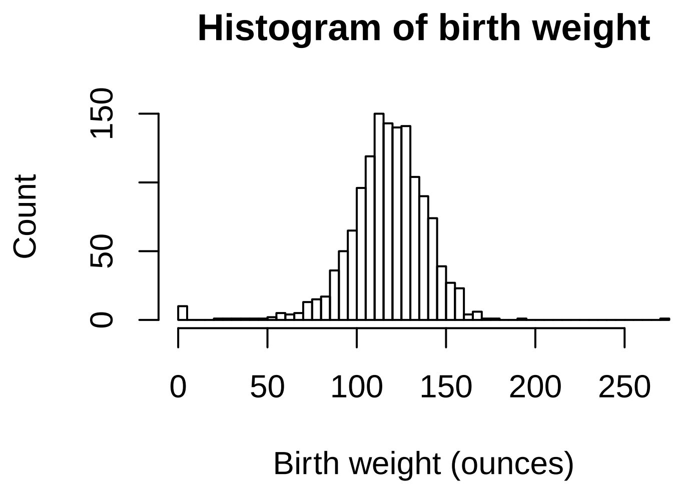

Figure 1: Histogram of birth weight (in ounces), using `hist` and bin width = 5

```
# 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")
```

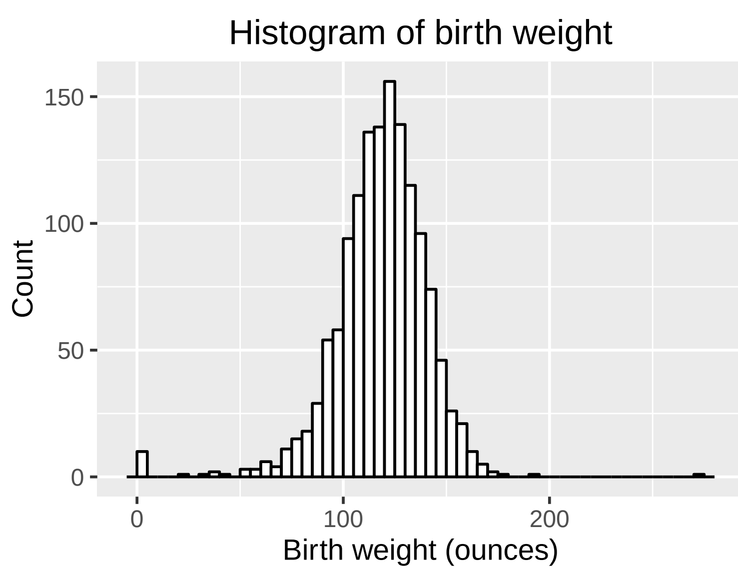

Figure 2: Histogram of birth weight (in ounces), using `ggplot` and bin width = 5

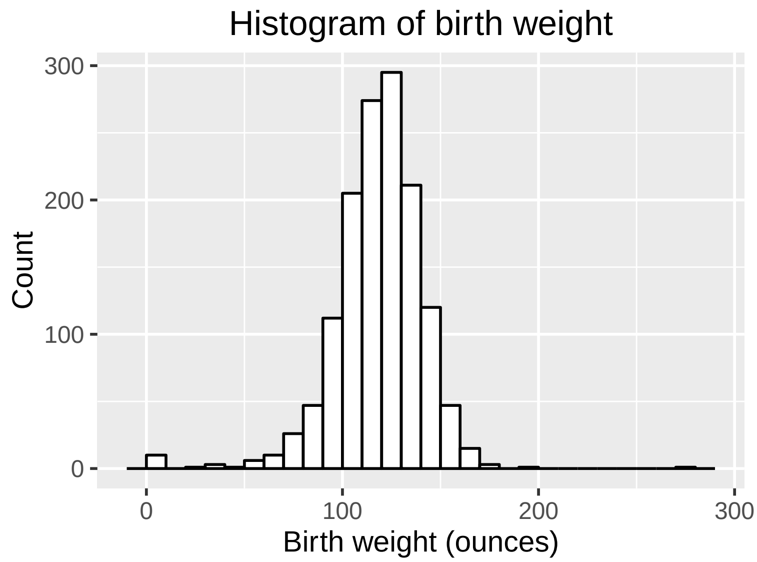

Figure 3: Histogram of birth weight (in ounces), using `ggplot` and bin width = 10

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 be representative of a few situations. There are no `NA` values for `data$bwght` so it seems likely that missing values have been coded as `0`, so we will exclude them from our analysis from now on. Alternatively, these values may be how mortality has been encoded. If we exclude those observations, the minimum birth weight is 23 ounces, which still seems very low but might be possible. Finally, some of us thought we should 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.

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
```

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

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

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

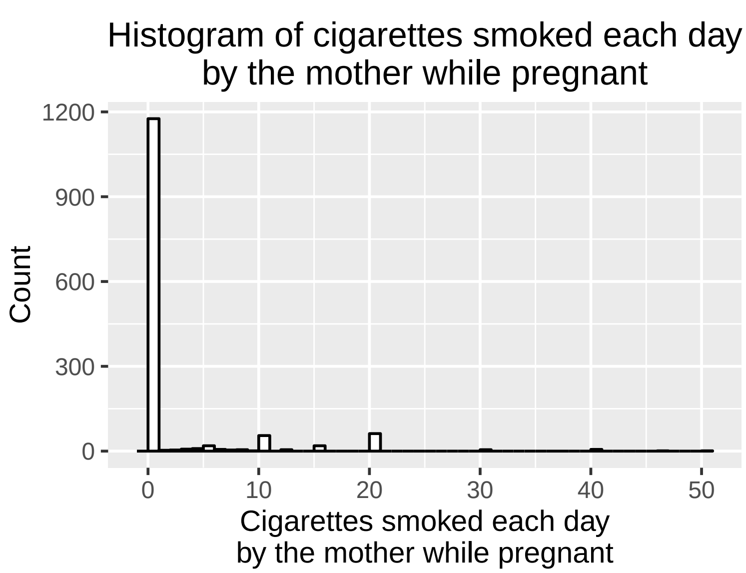

Figure 5: Histogram of cigarettes smoked each day by the mother while pregnant, bin width = 1

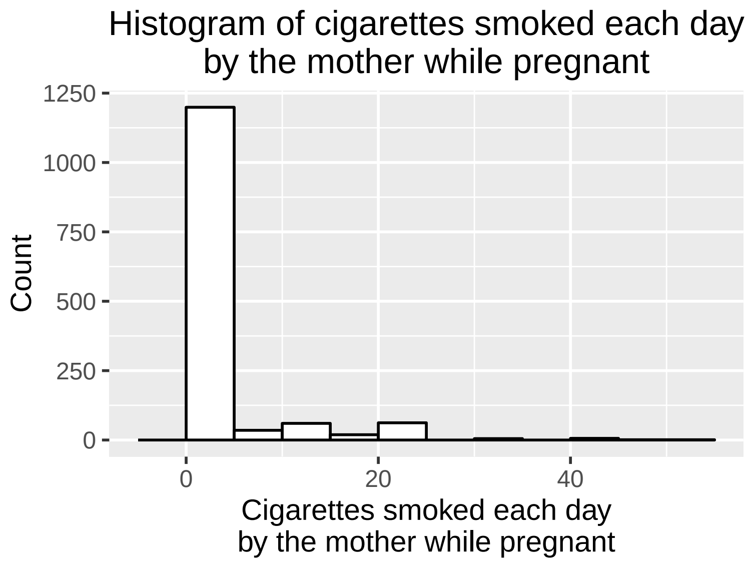

Figure 6: Histogram of cigarettes smoked each day by the mother while pregnant, bin width = 5

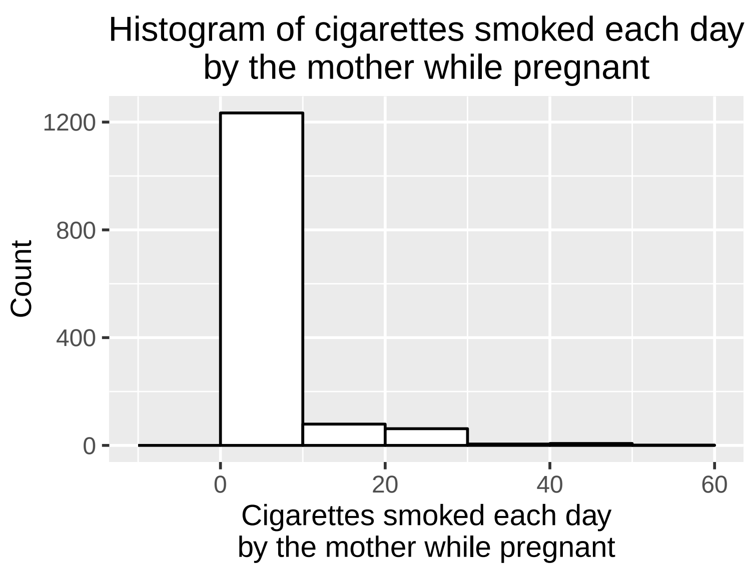

Figure 7: Histogram of cigarettes smoked each day by the mother while pregnant, bin width = 10

`cigs` has a heavy-tailed distribution, similar to a Pareto one. This makes sense, since most of the women do not smoke while pregnant, which is evident not only the histogram but also the quantiles of `cigs`. To better assess the shape of the distribution, it is more useful to look at the distribution among smokers.

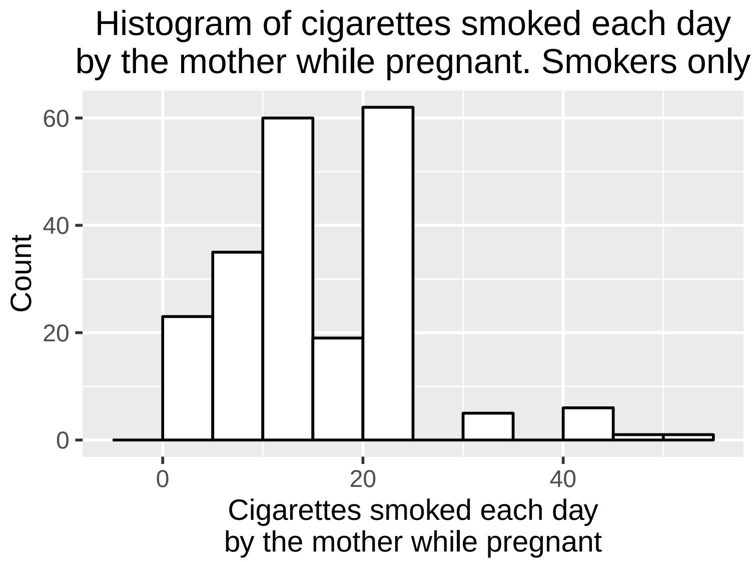

Figure 8: Histogram of cigarettes smoked each day by the mother while pregnant, bin width = 5, smokers only

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. Another alternative is making the cigs variable binary and just having a smokers vs non-smokers. However, since we lose some information in these approaches we kept origional cigs variable for further analysis.

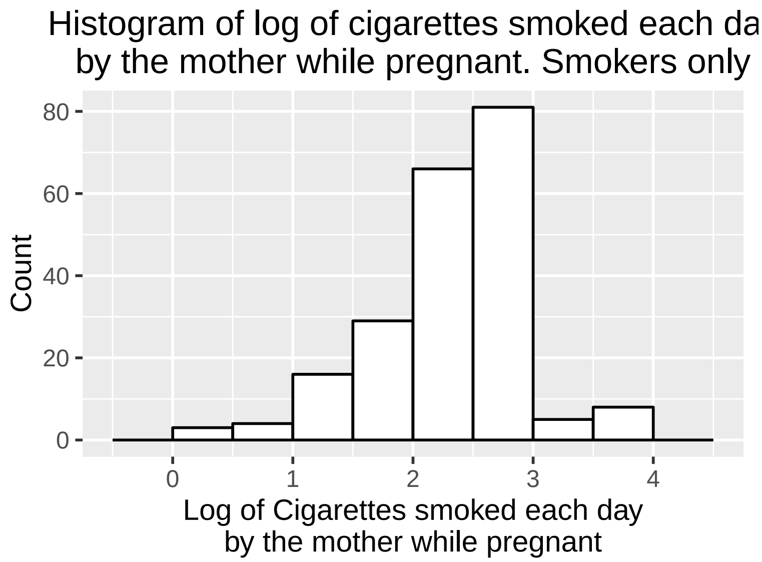

Figure 9: Histogram of log of cigarettes smoked each day by the mother while pregnant, bin width = 0.5, smokers only

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`?

```
ggplot(data = data, aes(cigs, bwght)) +
  geom_point() +
  labs(x = "Cigarettes smoked each day by the mother while pregnant",
       y = "Birth weight (ounces)",
       title = "Cigarettes smoked by the mother\nagainst birth weight") +
  geom_smooth(method = "lm")
```

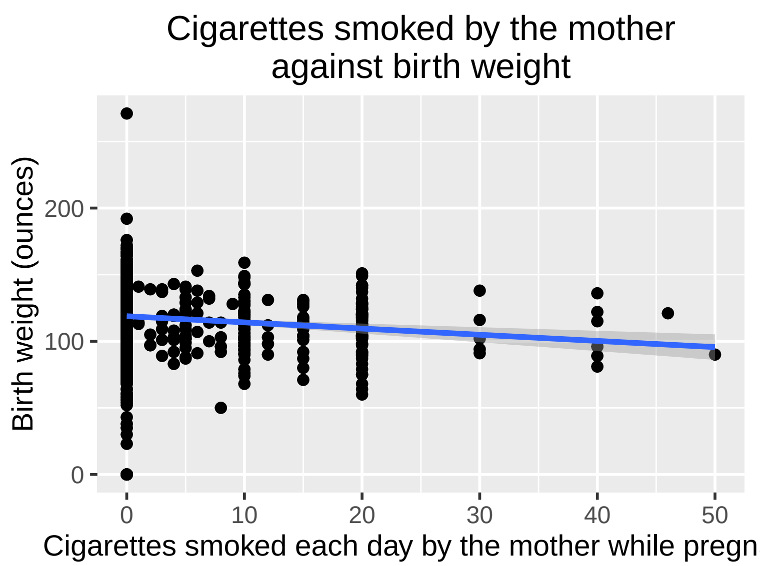

Figure 10: Scatterplot of birth weight (in ounces) against cigarettes smoked each day by the mother while pregnant

There seems to be a (small) negative relationship between the number of cigarettes smoked each day by the mother while pregnant and the birth weight of the child (i.e., the more a mother smokes, the less her child will weigh), but since `cigs` only takes a few values (mainly 0) we don't think it explains a lot of the variation in `bwght`: there is a huge 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.

```
# Regressor
params <- "cigs"
# Excluding bwght == 0 (possible missing observations)
```

```
##
## Call:
## lm(formula = as.formula(paste("bwght", paste(params, sep = "",
##     collapse = " + "), sep = " ~ ")), 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
```

Table 1: Effect of the number of cigarettes smoked each day by the mother while pregnant on the birth weight

	Birth weight (ounces)
Cigarettes smoked each day by the mother	-0.515***
	(0.091)
Baseline (Intercept)	119.790***
	(0.576)
R^2	0.023
F	32.179
p	0.000
N	1378

Regression showed a small negative effect of maternal cigarette smoking on birthweight ($\beta_1 = -0.515$ (0.091)). This represents a practically small but not meaningless effect. For example, among smokers, the average daily cigarettes smoked is 13.7. Thus, the child of an average smoker would have a 7.0 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):

```r
library(car)
scatterplot.matrix(bwght + cigs + faminc, data = data2)
```

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

```r
summary(data$faminc)
```

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

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

```
##   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
```

```
bin_width = 2
ggplot(data = data, aes(faminc)) +
  geom_histogram(colour = 'black', fill = 'white',
                 binwidth = bin_width) +
  labs(x = "Family income (thousands of dollars)", y = "Count",
       title = "Histogram of family income")
```

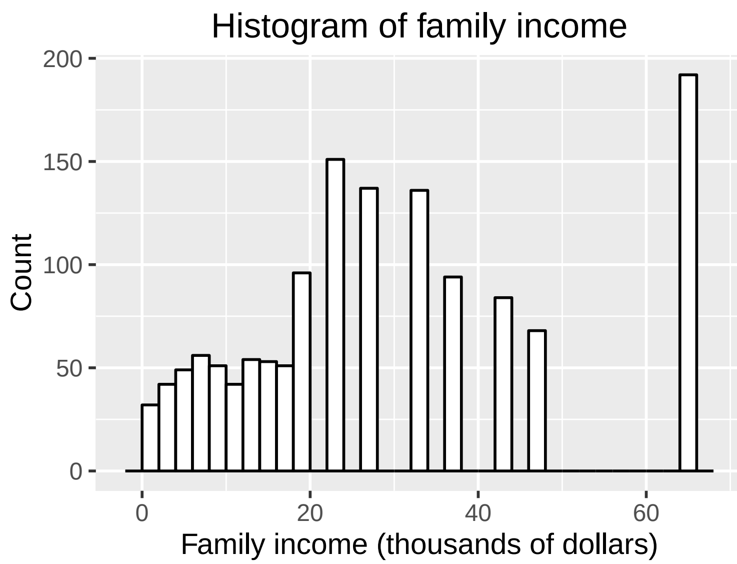

Figure 11: Histogram of family income (in thousands of dollars), bin width = 2

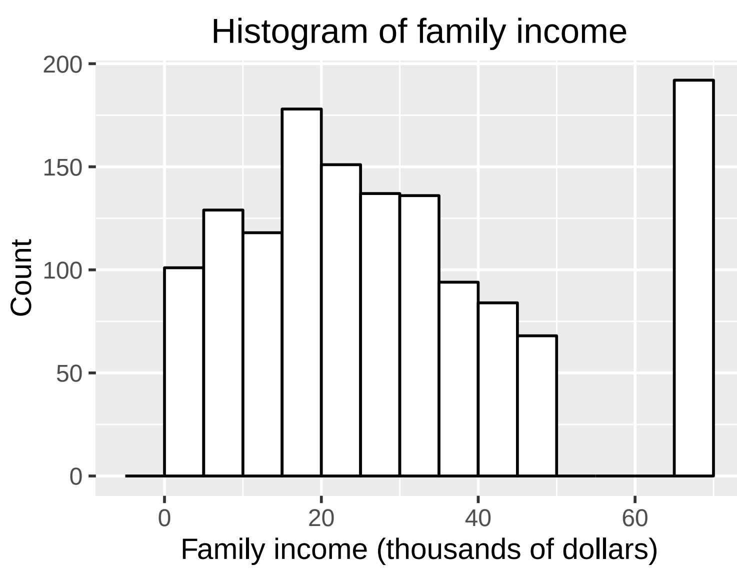

Figure 12: Histogram of family income (in thousands of dollars), bin width = 5

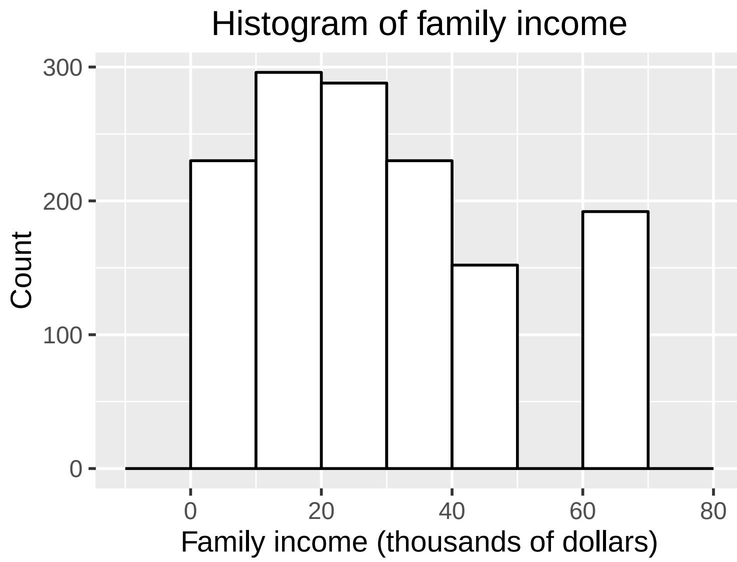

Figure 13: Histogram of family income (in thousands of dollars), bin width = 10

```r
ggplot(data = data, aes(faminc, bwght)) +
  geom_point() +
  labs(x = "Family income (thousands of dollars)",
       y = "Birth weight (ounces)",
       title = "Family income against birth weight") +
  geom_smooth(method = "lm")
```

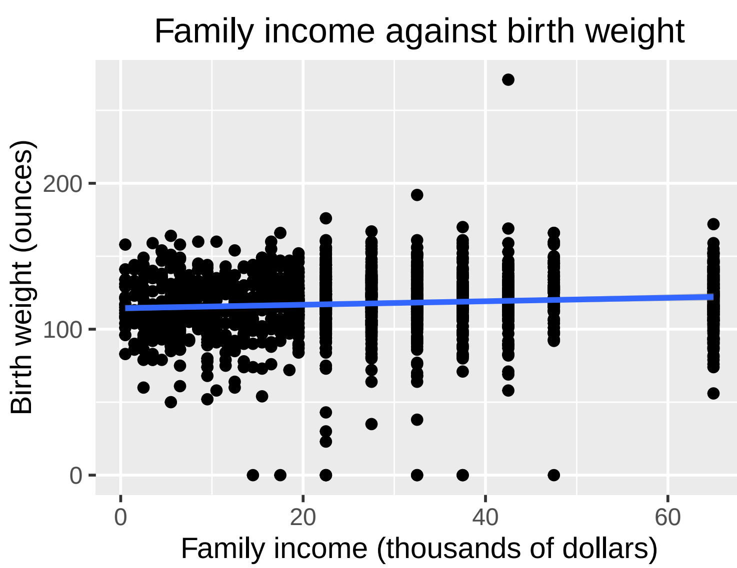

Figure 14: Scatterplot of birth weight (in ounces) against family income (in thousands of dollars)

This graph above also appears in the first row and second column of the scatterplot matrix in the next page: family income has a positive effect on birth weight, though again there is a lot of variation in the latter variable that may not be explained.

```r
scatterplotMatrix(~ bwght + cigs + faminc, data[data$bwght !=0, ])
```

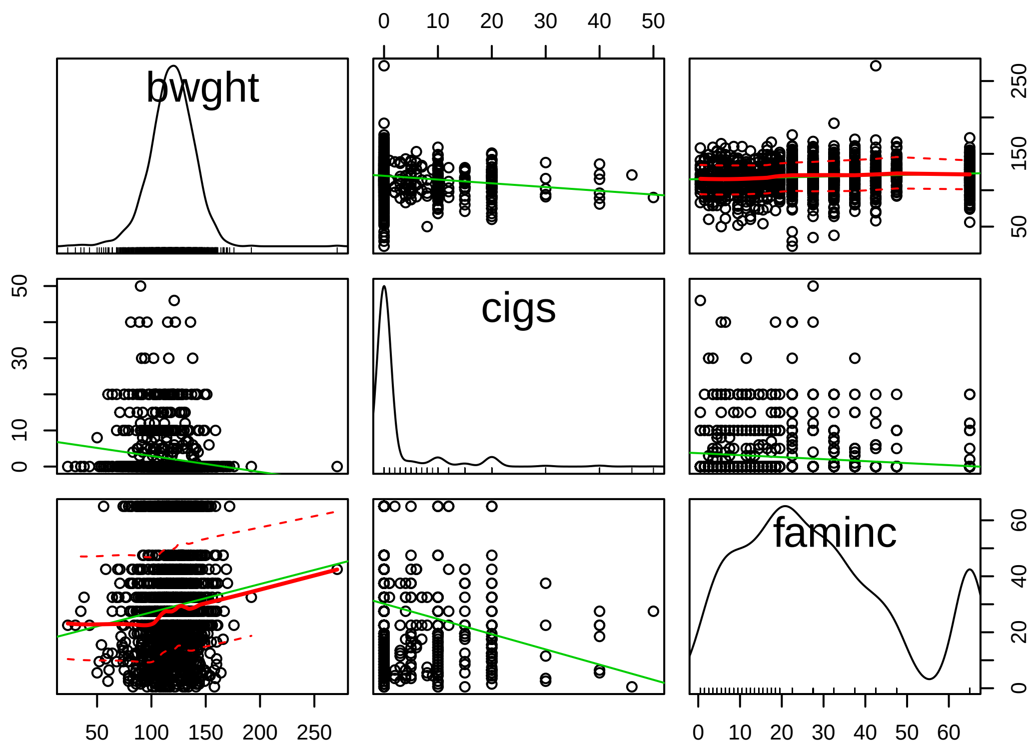

Figure 15: Scatterplot matrix of birth weight, cigarettes smoked each day by the mother while pregnant, and family income

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.

```r
# New regressors
params <- c("cigs", "faminc")
model2 <- lm(as.formula(paste("bwght", paste(params, sep = "",
                                             collapse = " + "), sep = " ~ ")),
             data = data[data$bwght !=0, ])
summary(model2)
```

```
##
## Call:
## lm(formula = as.formula(paste("bwght", paste(params, sep = "",
##     collapse = " + "), sep = " ~ ")), 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
```

Table 2: Effect of the number of cigarettes smoked each day by the mother while pregnant and the family income on the birth weight

	Birth weight (ounces)
Cigarettes smoked each day by the mother	-0.464***
	(0.092)
Family income (thousands of dollars)	0.093**
	(0.029)
Baseline (Intercept)	116.979***
	(1.054)
R^2	0.030
F	21.255
p	0.000
N	1378

Maternal cigarette smoking still has a small negative effect on birthweight, slightly smaller (in absolute value) than when family income is not considered ($\beta_1 = -0.464$ (0.092)).

The effect of income on birth weight ($\beta_2 = 0.093$ (0.029)), though statistically significant, has a small practical significance, as we would have to move from the median income in the sample ($\$27,500$) to the 95th percentile ($\$65,000$) to have an expected increase in birth weight of 3.5 Oz. Following what we mentioned in Question 6, the effect of smoking is more practically significant, since the child of a median smoker (i.e., excluding non-smokers; 10 cigarettes per day) would have about a 4.6 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 of `cigs` represents the mean change in birth weight for one unit of change in `cigs` (i.e., for one cigarette more smoked per day by the mother while pregnant), holding familiy income constant (which may be possible for this particular variables but not always: sometimes we cannot change the value of one regressor while leaving the other(s) unchanged). This differs from the simple regression as the coefficient in it represents the association of cigarettes smoked and birth weight without holding any other measured variables constant (where we make the assumption that `cigs` is not related to any other variable).

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. Regressing the explained variable on a single single regressor, without any other predictors, may produce a very different coefficient, because those other predictors are not held fixed. When we omit a variable X_2 in a model, the estimator of the slope of X_1, $\tilde{\beta}_1$, is:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the slope estimators from the multiple regression and $\tilde{\delta}_1$ is the slope from the simple regression of X_2 on X_1.

That means that, in this case, the relationship between `cigs` and `faminc` is negative ($\tilde{\delta}_1 < 0$; the higher the family income of the mother, the less she smokes while pregnant); that is why $\tilde{\beta}_1$ is more negative than $\hat{\beta}_1$.

Let's confirm the expression above (we had all coefficients but $\tilde{\delta}_1$ from Table 1 and Table 2, but we'll recalculate anyway):

```
# Demonstration of the differences in the value of one coefficient
# when another variable is included or omitted
tilde_delta_1 <- summary(lm(faminc ~ cigs,
                            data = data[data$bwght != 0, ]))$coefficients[2, 1]
hat_beta_1 <- summary(lm(bwght ~ cigs + faminc,
                         data = data[data$bwght != 0, ]))$coefficients[2, 1]
hat_beta_2 <- summary(lm(bwght ~ cigs + faminc,
                         data = data[data$bwght != 0, ]))$coefficients[3, 1]
hat_beta_1 + hat_beta_2 * tilde_delta_1
```

```
## [1] -0.5146957
```

```
(tilde_beta_1 <- summary(lm(bwght ~ cigs,
                            data = data[data$bwght != 0, ]))$coefficients[2, 1])
```

```
## [1] -0.5146957
```

$$-0.464 + 0.093 \cdot (-0.544) = -0.515$$