DataSci 271: Exercise 1

May 08, 2016

The optional exercise serves two purposes: (1) Extend the materials taught in the asynchronized materials; some new concepts or techniques are introduced in the weekly assignment. (2) Ensure that you have learned the concepts, techniques, theories, statistical models covered in a specific week. Even though these are optional, the students are highly encouraged to work on these exercises on their own and then discuss their analyses with fellow classmates.

The file *birthweight_w271.RData* contains data from the 1988 National Health Interview 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 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 will treat the data as a true random sample.

This exercise corresponds to the materials taught in week 3. You will practice using exploratory data analysis (EDA), providing narrative with your analysis, start thinking about how the insights gained from the EDA will impact your specification of the regression model, interpreting OLS coefficients, and summarizing the results of your final model.

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“)

```r
library(data.table)
library(dplyr)
library(stargazer)
library(Hmisc)

rm(list = ls())
load('/Users/patrickng/Documents/mids/w271/Ex01/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? Examine the number of missing observations in each of the variables.

These commands will be useful:

- desc
- str(data)
- summary(data)

```r
desc
str(data)
summary(data)
describe(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 execrise, we are in the question: *"Do mothers who smoke have babies with lower birth weight?"*

The dependent variable of interested is *bwght*, representing birthweigt in ounces. Examine this variable using both tabulated summary and graphs. Specifically,

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.

In each of the tables and graphs, explain what you observe. Think about how they will affect the variables entering your regression models in the later stage of your analysis.

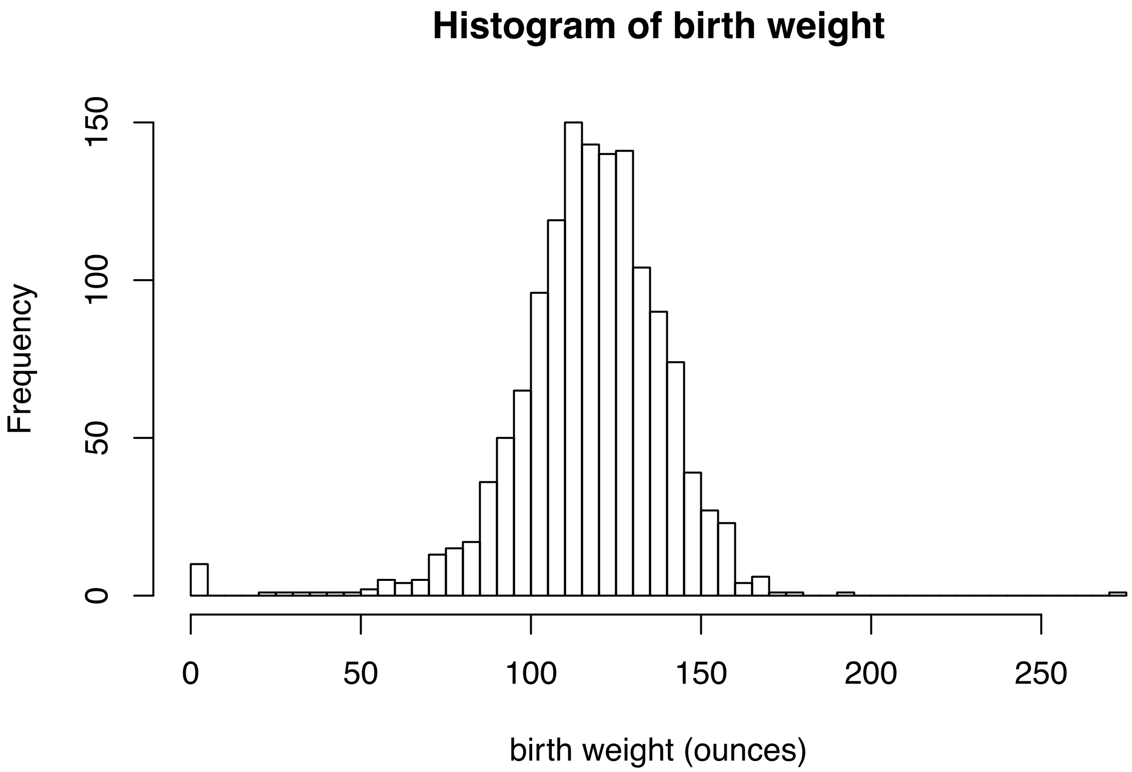
```r
summary(data$bwght)
```

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

```r
quantile(data$bwght, probs=c(0.01, 0.05, .1, .25, .5, .75, .9, .95, .99))
```

```
##     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
```

```r
hist(data$bwght, breaks=40, main="Histogram of birth weight", xlab="birth weight (ounces)")
```



In the histogram, there are several records with weight near zero.

```
table(data$bwght)
```

```
##
##   0  23  30  35  38  43  50  52  54  56  58  60  61  64  68  69  70  71
##  10   1   1   1   1   1   1   1   1   1   2   2   1   3   3   1   1   2
##  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89
##   2   2   4   3   3   1   2   6   3   3   4   5   3   2   7   9   5   6
##  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107
##   9  13  10  13   9   5  12  15  17   9  12  18  19  16  29  14  26  22
## 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
##  25  24  22  18  36  26  34  36  25  26  22  29  41  21  30  35  29  25
## 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143
##  27  24  41  22  27  19  20  21  28  16  29  16  14  21  10  19  15  13
## 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161
##  17  10  12   9   5  10   3   6  10   4   3   4   6   2   3   6   6   3
## 164 166 167 169 170 172 176 192 271
##   1   3   1   1   1   1   1   1   1
```

From the output of the table() command, we see that 10 records have zero ounce. They seem to be missing data.

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.

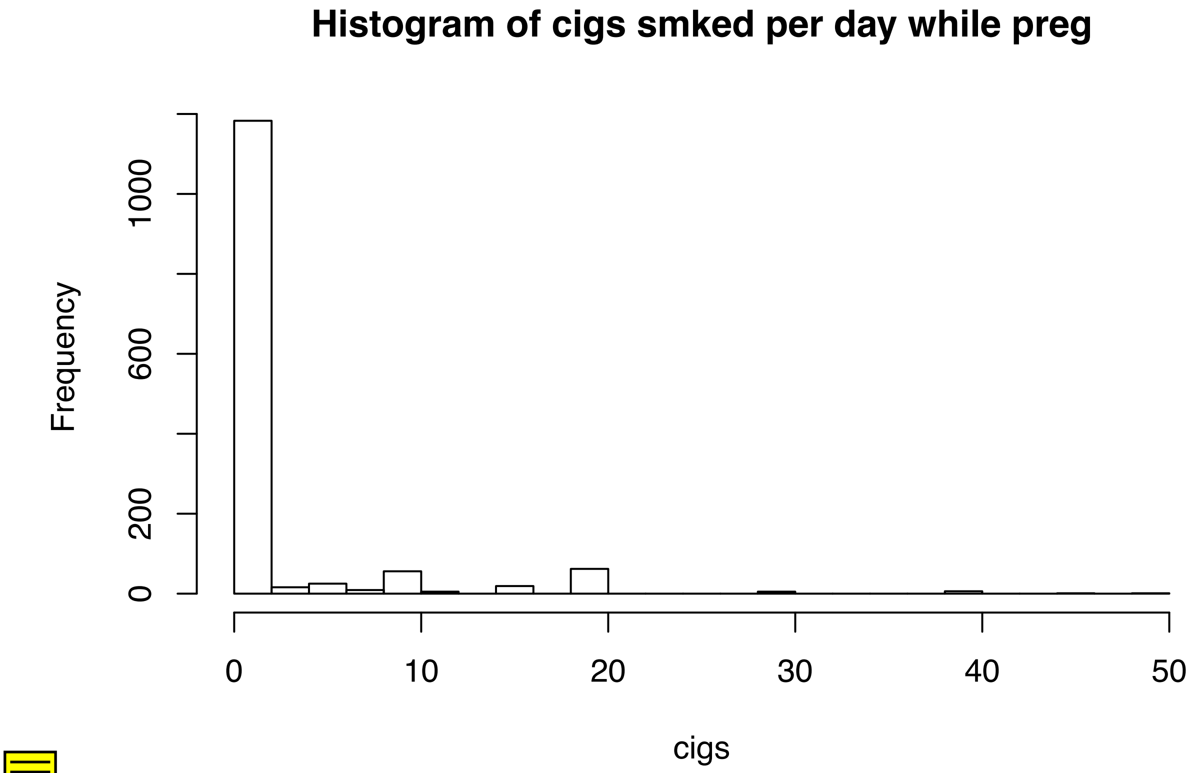
```r
summary(data$cigs)
```

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

```r
quantile(data$cigs, probs=c(0.01, 0.05, .1, .25, .5, .75, .9, .95, .99))
```

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

```r
hist(data$cigs, breaks=20, main="Histogram of cigs smked per day while preg", xlab="cigs")
```

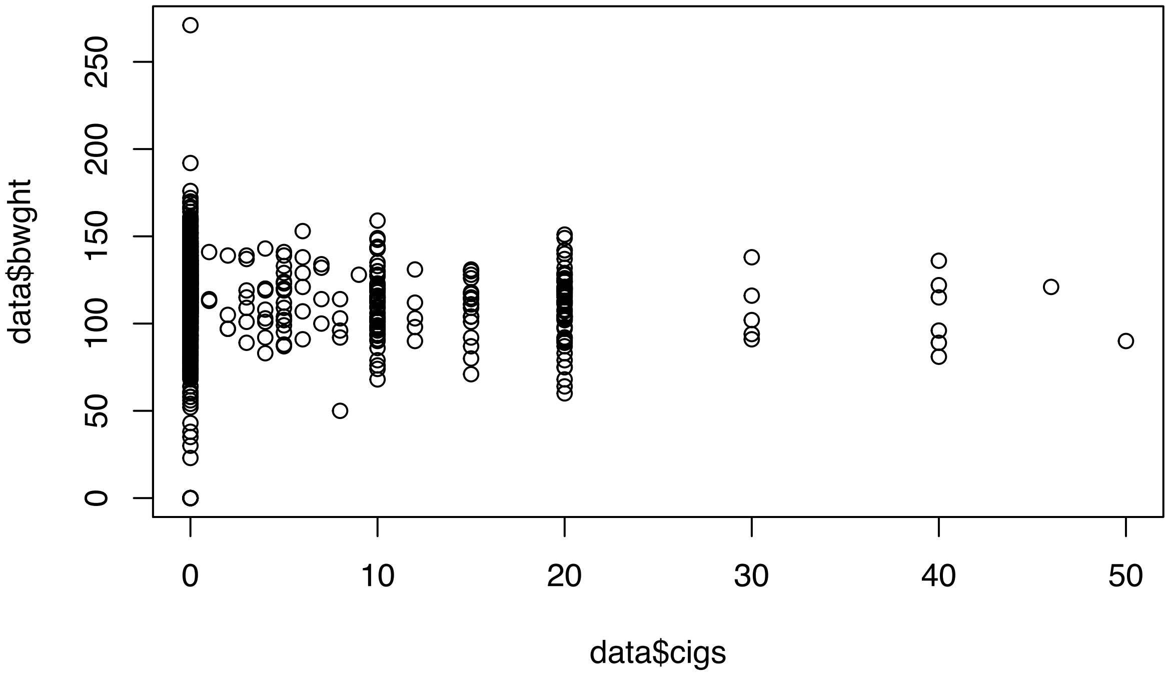


Cigs seems to follow a poisson distribution.

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*?

```r
plot(data$cigs, data$bwght)
abline()
```



The mean value of bwght seems to be quite uniform. In another words, a change in cigs does not seem to bring a large change in the mean value of bwght. There I think only a small variation in bwght can be explained by cigs.

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.

```
data2 <- data.table(data)
data2 <- data2[bwght>0]
m <- lm(bwght ~ cigs, data = data2)
summary(m)
```

```
##
## Call:
## lm(formula = bwght ~ cigs, data = data2)
##
## 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
```

Coefficient = -0.515, and SE = 0.091. It means that, with all other factors held constant, a decrease of 0.515 ounce in birth weight is seen with an increase in one cigerette smoked per day during pregency.

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.

When producing a scatterplot matrix, make sure that each of your graphs is legible. It may note be an issue in this exercise, but when you have a datasets of many potential independent variables, this becomes an issue.

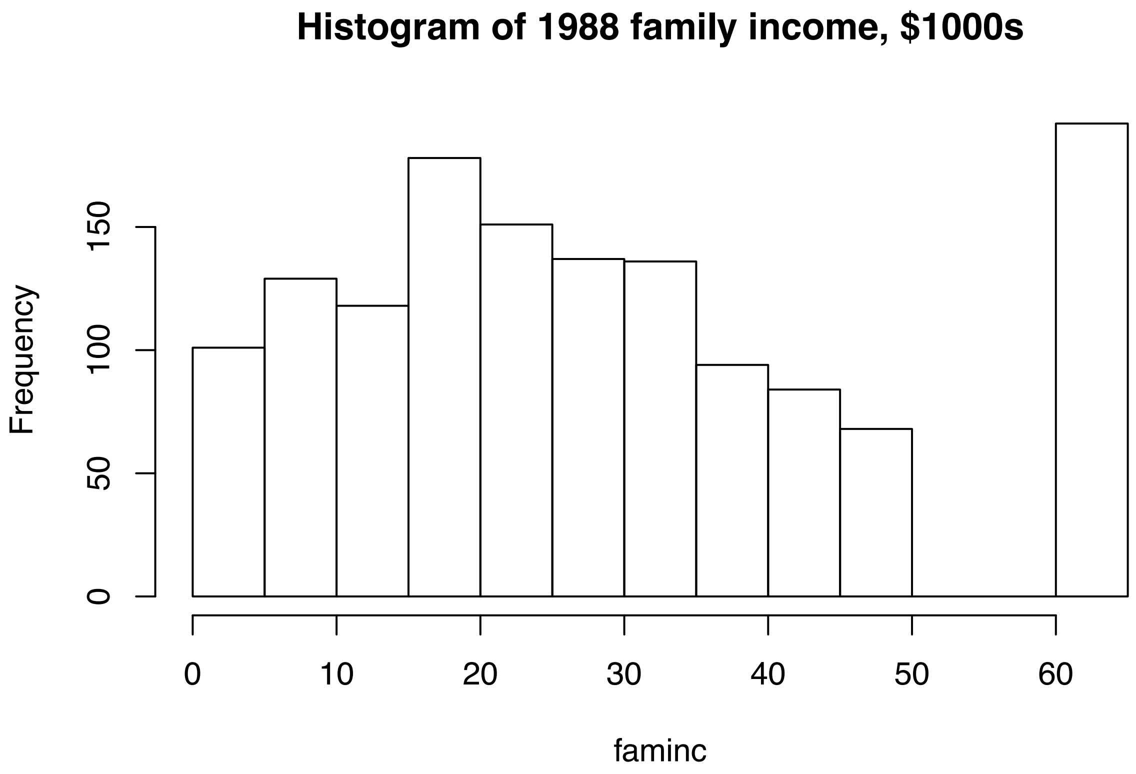
```r
summary(data$faminc)
```

```
##    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(0.01, 0.05, .1, .25, .5, .75, .9, .95, .99))
```

```
##   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
hist(data$faminc, breaks=20, main="Histogram of 1988 family income, $1000s", xlab="faminc")
```



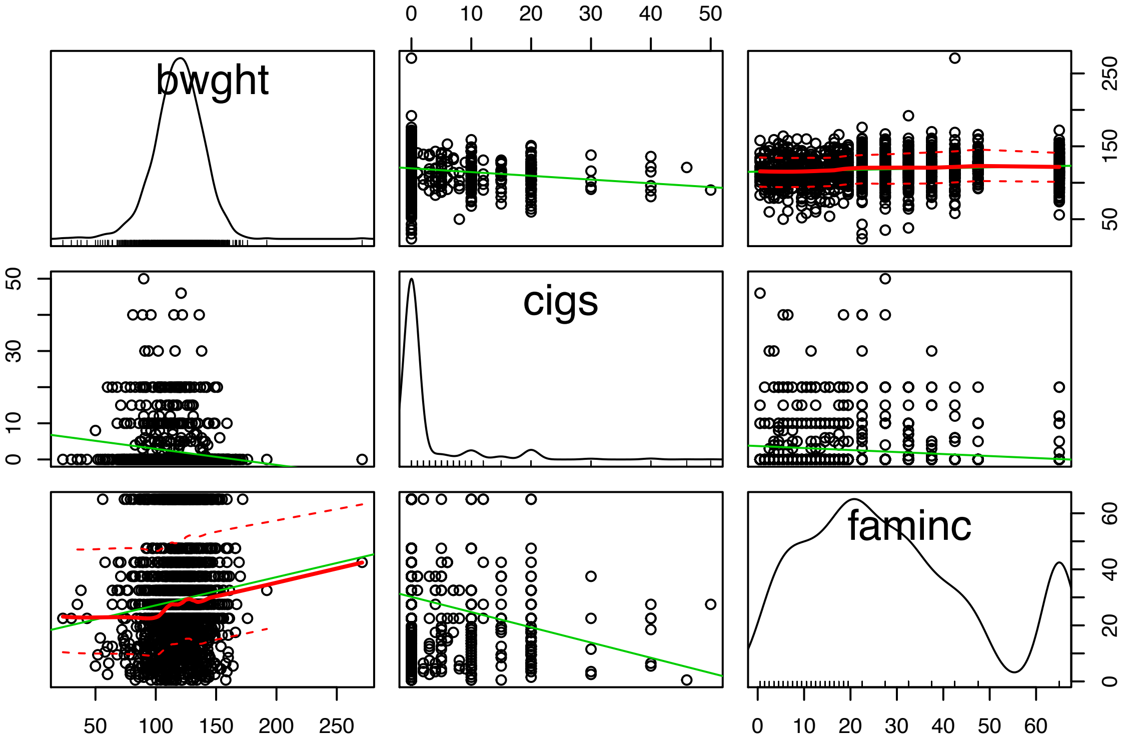
```r
library(car)
scatterplotMatrix(~bwght + cigs + faminc, data=data2)
```

```
## 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:

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
library(stargazer)
m2 <- lm(bwght ~ cigs + faminc, data = data2)
stargazer(m2, type='text', single.row = T)
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                bwght
## -----------------------------------------------
## cigs                       -0.464*** (0.092)
## faminc                     0.093*** (0.029)
## Constant                  116.979*** (1.054)
## -----------------------------------------------
## Observations                     1,378
## R2                               0.030
## Adjusted R2                      0.029
## Residual Std. Error       20.107 (df = 1375)
## F Statistic           21.255*** (df = 2; 1375)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Coefficients estimates (SE):

- cigs: -0.464 (0.092)
- faminc: 0.093 (0.029)

It means that, with all other factors held constant:

- A decrease of 0.464 ounce in birth weight is seen with an increase in one cigerette smoked per day during pregency.
- An increase of 0.093 ounce in birth weight is seen with an increase in $1000 in family income.

Question 9:

Explain, in your own words, what does the coefficient of *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.

```r
stargazer(m, m2, type='text', single.row = T)
```

```
##
## =====================================================================
##                                    Dependent variable:
##                     -------------------------------------------------
##                                           bwght
##                               (1)                      (2)
## ---------------------------------------------------------------------
## cigs                   -0.515*** (0.091)        -0.464*** (0.092)
## faminc                                           0.093*** (0.029)
## Constant               119.790*** (0.576)       116.979*** (1.054)
## ---------------------------------------------------------------------
## Observations                 1,378                    1,378
## R2                           0.023                    0.030
## Adjusted R2                  0.022                    0.029
## Residual Std. Error    20.173 (df = 1376)       20.107 (df = 1375)
## F Statistic         32.179*** (df = 1; 1376) 21.255*** (df = 2; 1375)
## =====================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

Let b1 and b2 be the coefficients of cigs and faminc. The value of b1 is the change seen in *bwght* when a unit change is seen in *cigs*.

From the scatterplot matrix, we can see a -ve correlation between cigs and faminc. Because b2 is +ve, it means omitting faminc will have a -ve bias on b1, which explains why the estimated b1 (-0.515) in the single regression model is less than that (-0.464) in the multiple regression model.