W271-2 – Spring 2016 – **HW 1**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 2, 2016

Contents

Data

The file birthweight w271.RData contains data from the 1988 National Health Inter- view 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 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 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 the data
setwd("C:/Users/songminghu/UCB_DataScience/W271_ApplyRegressionTimeSeriesAnalysis/data")
load("birthweight_w271.Rdata")




# QUESTION 2 --------------------------------------------------------------
# check how many variables and observations in the data -------------------
# there are  14 variables and 1388   observations in the data -------------
```

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)

```
dim(desc)[1] # check number of variables, or use str(data) command
```

```
## [1] 14
```

```
dim(data)[1] # check number of observations, or use str(data) command
```

```
## [1] 1388
```

```
# QUESTION 3 --------------------------------------------------------------
```

therefore, there are totally 14 variables and 1388 observations.

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)**

```
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(0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99) )
```

```
##     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
```

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.**

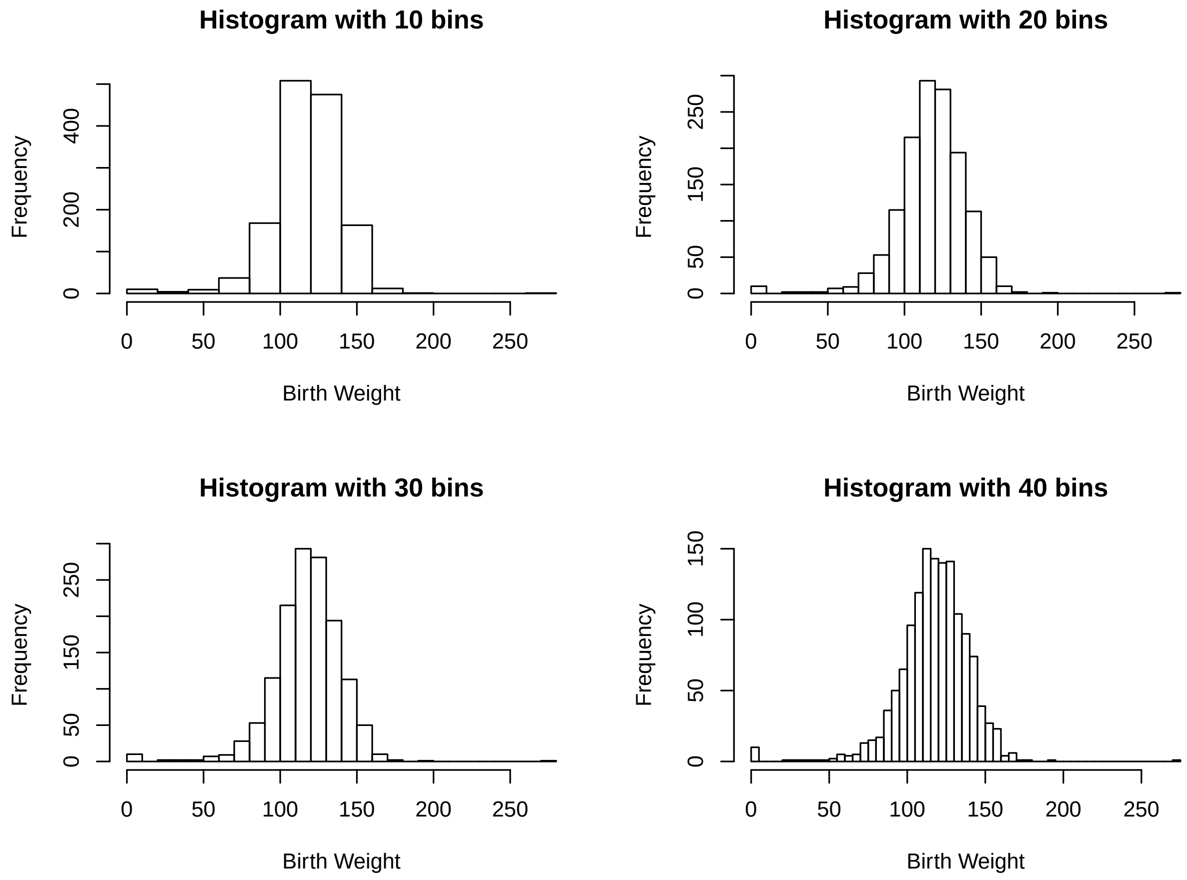

Figure 1: Histogram plots of Birth Weight with varied bins

As shown in the above histrogram plots, the overall shape of those histograms are similar and all of them seem to follow the normal distribution. Those distribution plots with the increasing number of bins becomes much smoother.

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 identifed.**

Strangely, there are around 10 outliers with zero birth weight as shown in the left part of histograms.

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.

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

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

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

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

```
quantile( data$cigs, probs = c(0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99) )
```

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

3. **Plot the histogram of cigs 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.**

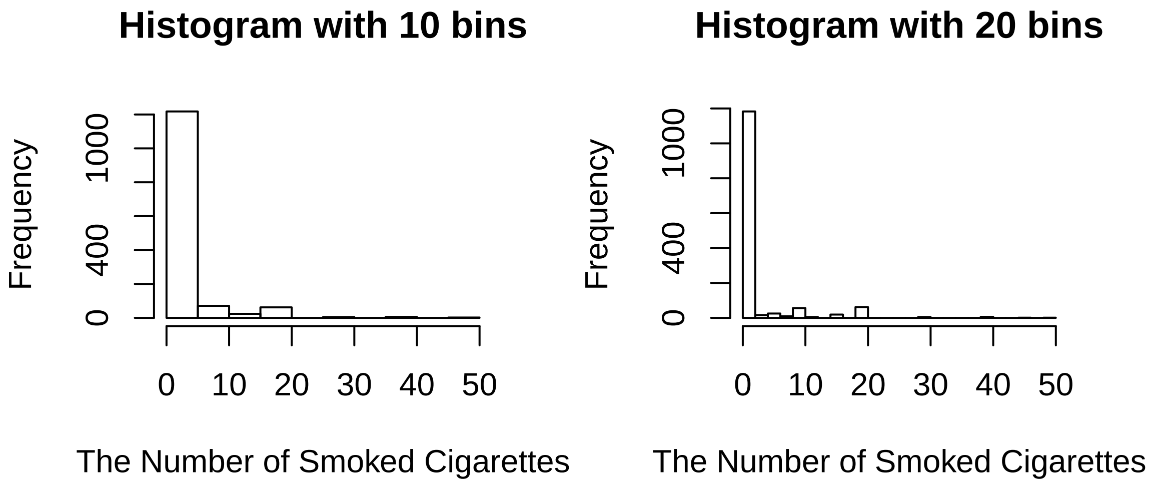

Figure 2: Varied Histogram plots for the number of cigarettes smoked each day

As shown in the above histrogram plots, their histrogram distribution for the number of cigarettes smoked each day are skewed toward zero, The overall shape of those histograms are similar as well.

4. **This is a more open-ended question: Have you noticed anything "strange" with the cigs variable and the shape of histogram this variable? If so, please elaborate on your observations and investigate any issues you have identifed.**

????

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?

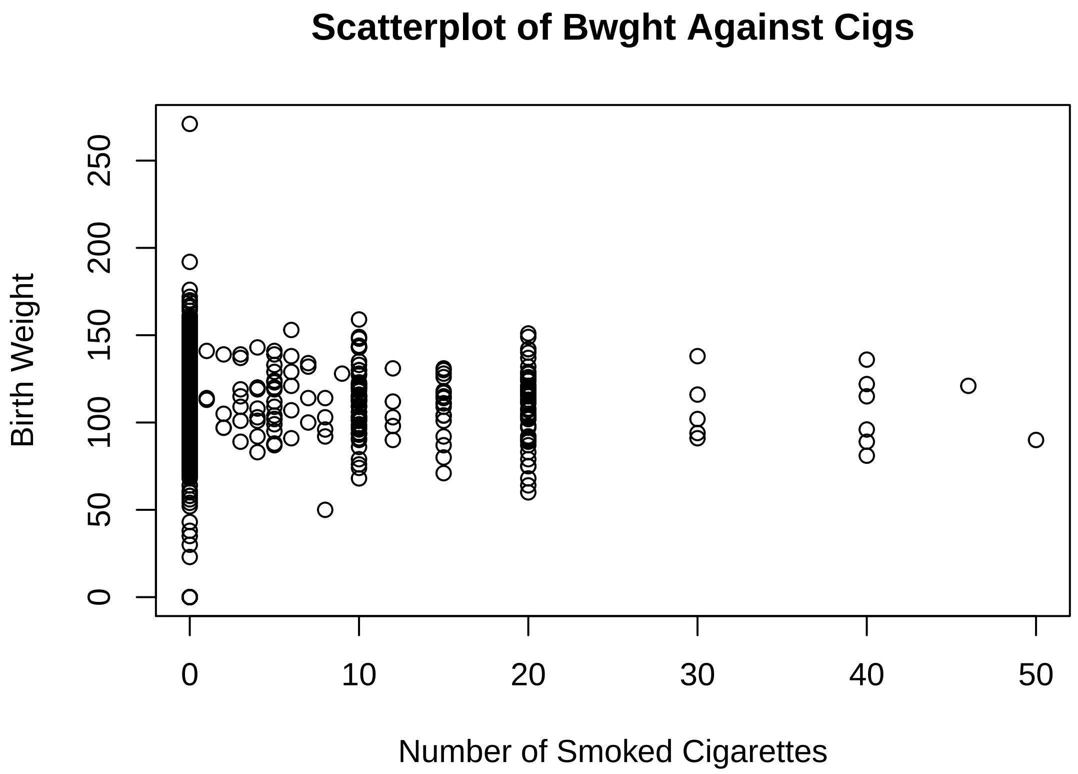

Figure 3: Varied Histogram plots for the number of cigarettes smoked each day

```
bwght_cigs.lm = lm(bwght ~ cigs, data=data)
summary(bwght_cigs.lm)$r.squared
```

```
## [1] 0.01495144
```

```
# QUESTION 6 ------------------------------------------------------------
```

The variable Cigs explains only about 1.5% of variation in the variable of bwght.

Question 6:

Estimate the simple linear regression of bwght on cigs. What coeffcient estimates and the standard errors associated with the coeffcient 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.

Since there are 10 observations with zero body weights, we need to take care of those obvious outliers. we remove those observations and assign the left data to a new data frame variable, newdata.

```
newdata = data[data[,"bwght"] != 0,]
```

The estimated coefficient and its standard error are listed in the belowed coefficient matrix:

```
bwght_cigs_2.lm = lm(bwght ~ cigs, data=newdata)
summary(bwght_cigs_2.lm)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 119.7896035 0.57594713 207.987151 0.000000e+00
## cigs         -0.5146957 0.09073251  -5.672671 1.711428e-08
```

```
# QUESTION 7 -------------------------------------------------------------
```

For cigs, the estimate coefficient is about -0.515 and its associated standard error is 0.091. this means that if mothers increase their number smoked cigarettes by one each day during pregnancy, on average the predicted birth weight of their new babies would decrease about one-half a percentage point (0.515) ounces.

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):

library(car)

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

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

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

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

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

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

```
quantile( data$faminc, probs = c(0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99) )
```

```
##   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
```

3. **Plot the histogram of faminc 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.**

As shown in the above histrogram plots, their histograms looks like the bimodal distribution. Those distribution plots with the increasing number of bins becomes much more spiky.

4. **Produce a scatterplot matrix of bwght, cigs and faminc.**

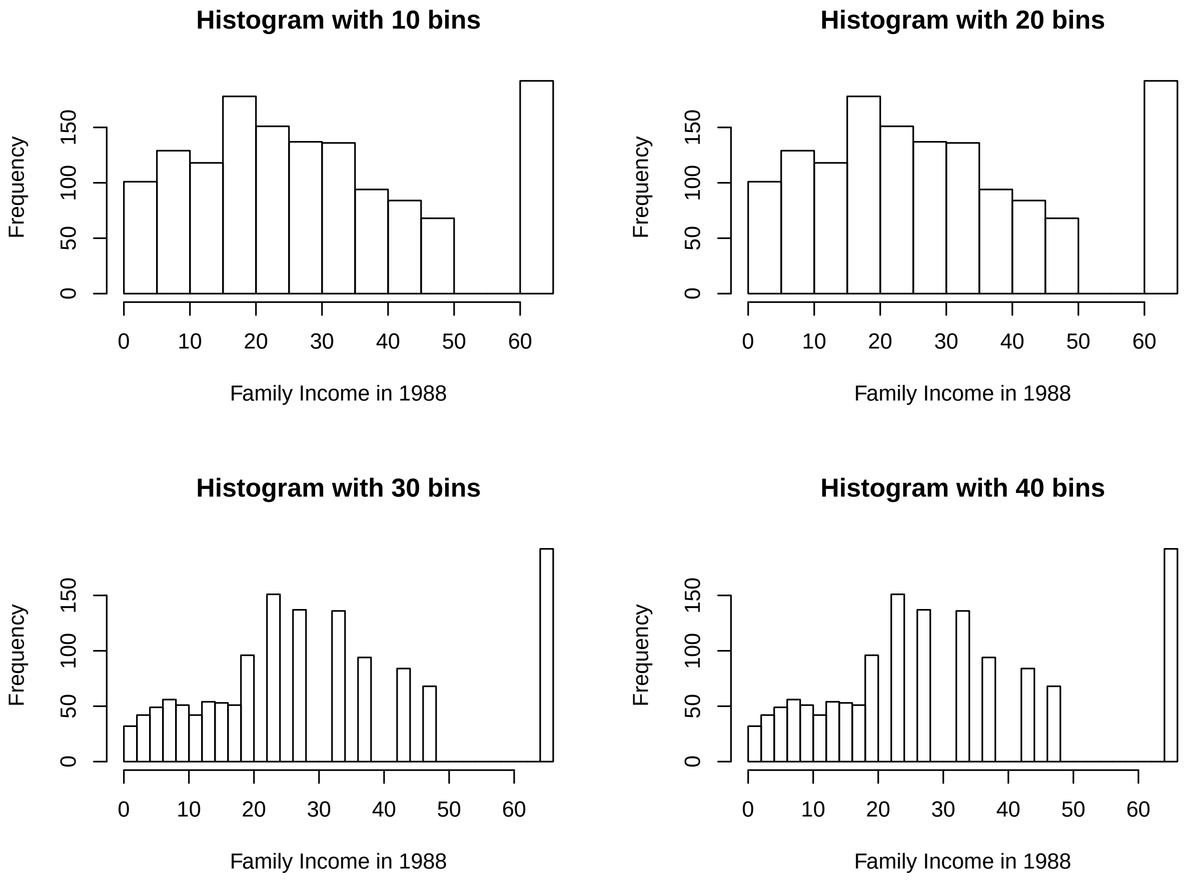

Figure 4: Histogram plots of family income in 1988 with varied bins

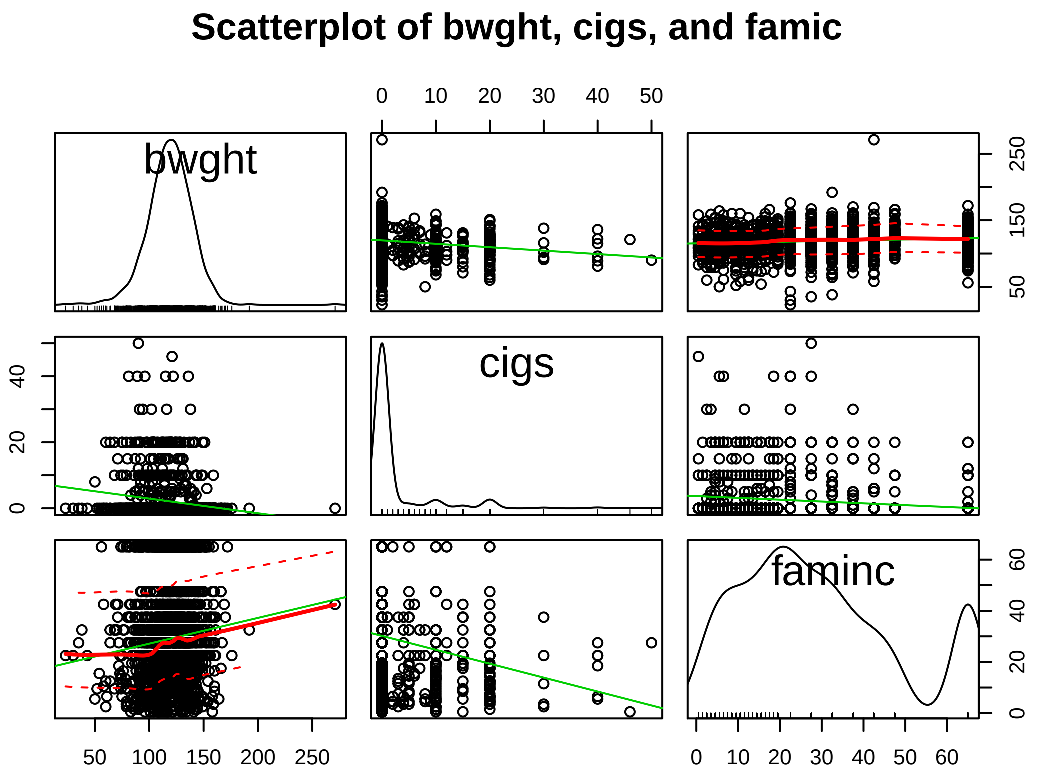

Figure 5: Scatterplot matrix of bwght, cigs and faminc

Question 8:

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

The estimated coefficient and its standard error are listed in the belowed coefficient matrix:

```
bwght_cigs_faminc.lm = lm(bwght ~ cigs+faminc, data=newdata)
summary(bwght_cigs_faminc.lm)$coefficients
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 116.97933227 1.05363032 111.025025 0.000000e+00
## cigs         -0.46406827 0.09182343  -5.053920 4.910700e-07
## faminc        0.09314249 0.02928298   3.180772 1.501637e-03
```

```
# QUESTION 9 ---------------------------------------------------------------
```

First, the intercept 116.98 is the predicted birth weight of babies if the lady do not smoke cigarettes during the pregnancy and their family income in 1988 are zero. there is a partial negative relationship between cigs and bwght. If Holding faminc fixed, one more cigarettes per day during the pregnancy would lead to a decrease of 0.464 ounce regarding the new born's birth weight. on the other hand, there is a partial positive relationship between faminc and bwght. If cigs is fixed, everyone one thousand dollar increase of the family income would contribute to 0.093 ounce gain of new born's weight.

Question 9:

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

The coefficient on cigs derived from the multiple regresion is a little bit larger (less negative) than its coefficient derived from the simple regression. However, unlike the simple regression, now we can utilize this equation to differentiate the birth weight prediction for those women with the same cigs value but different family incomes. i

Question 10:

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