

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:	2
Question 4:	4
Question 5:	5
Question 6:	5
Question 7:	5
Question 8:	5
Question 9:	5
Question 10:	5

Data

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 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")`

```
#### 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 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 interest 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(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 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.

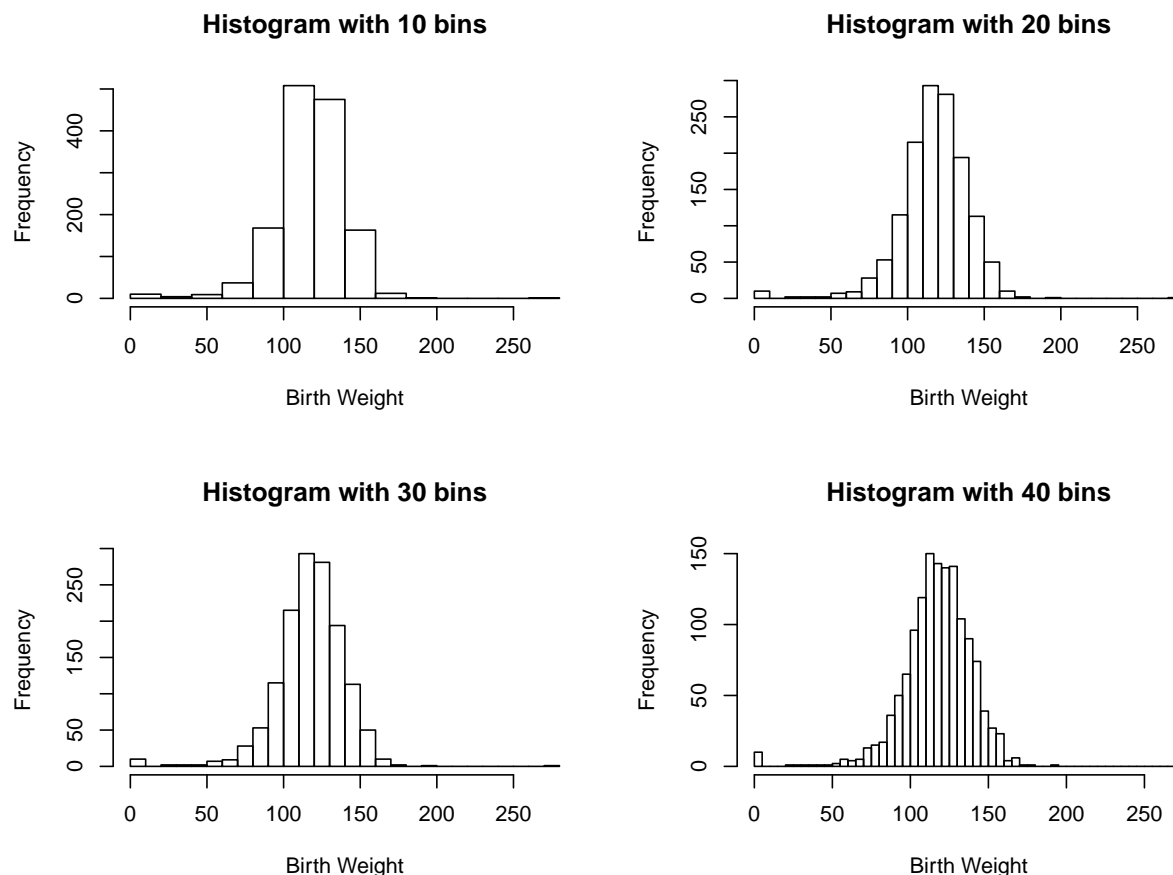


Figure 1: Histogram plots of Birth Weight with varied bins

As shown in the above histogram 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 identified.

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 `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.

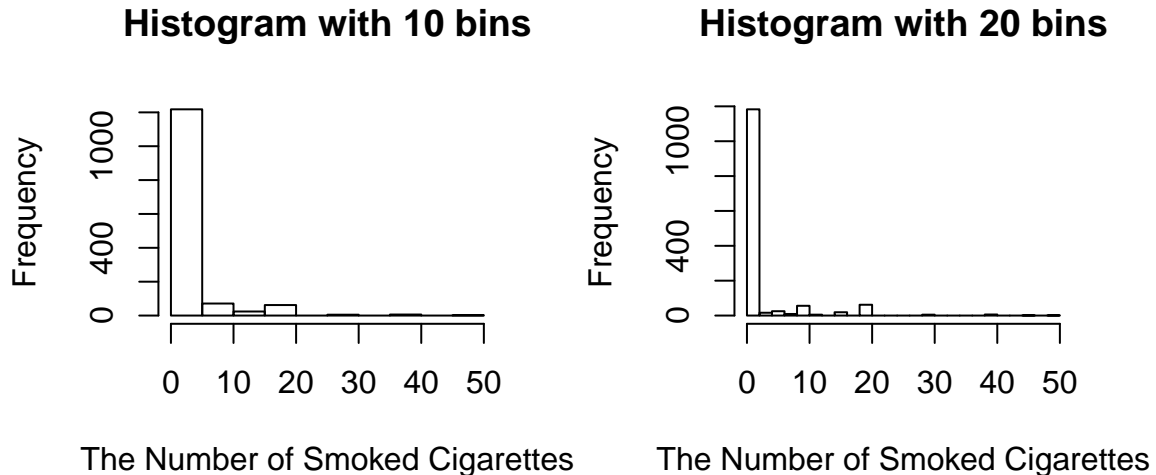


Figure 2: Varied Histogram plots for the number of cigarettes smoked each day

As shown in the above histogram plots, their histogram 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 `bwght` variable and the shape of histogram this variable? If so, please elaborate on your observations and investigate any issues you have identified.

???

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?

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.

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.

Question 8:

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

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.

Question 10:

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