

W271-2 – Spring 2016 – HW 1

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 3, 2016

Contents

Data	2
Exercises	2
Question 1	2
Question 2	2
Question 3:	4

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")`

```
# QUESTION 1 -----
# Load the birthweight dataset
load("data/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)`

```
# QUESTION 2 -----
# Examine the basic structure of the data set using desc, str, and summary
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 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)`

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

```
# 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.

```
# Plot the histogram of bwght and comment on the shape of its distribution
# 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")
```

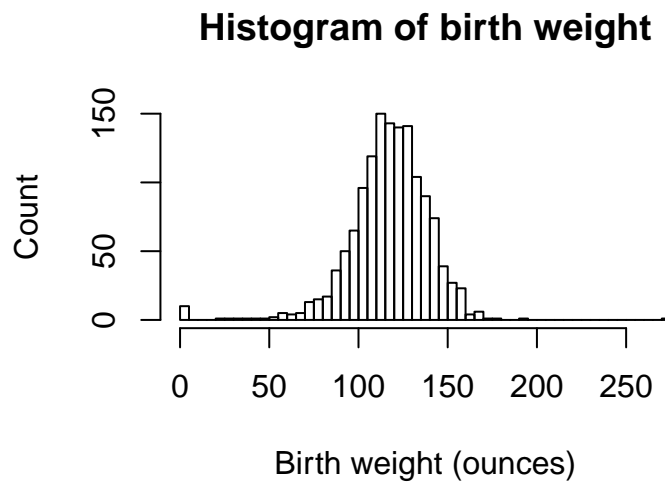


Figure 1: Histogram of birth weight (in ounces), using `hist` and bin width = 5

```
# 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
ggplot(data = data, aes(bwght)) +
```

```
geom_histogram(binwidth = bin_width) +  
labs(x = "Birth weight (ounces)", y = "Count",  
     title = "Histogram of birth weight")
```



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

```
# Use ggplot and bin width = 20  
bin_width = 20  
ggplot(data = data, aes(bwght)) +  
  geom_histogram(binwidth = bin_width) +  
  labs(x = "Birth weight (ounces)", y = "Count",  
       title = "Histogram of birth weight")
```

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.

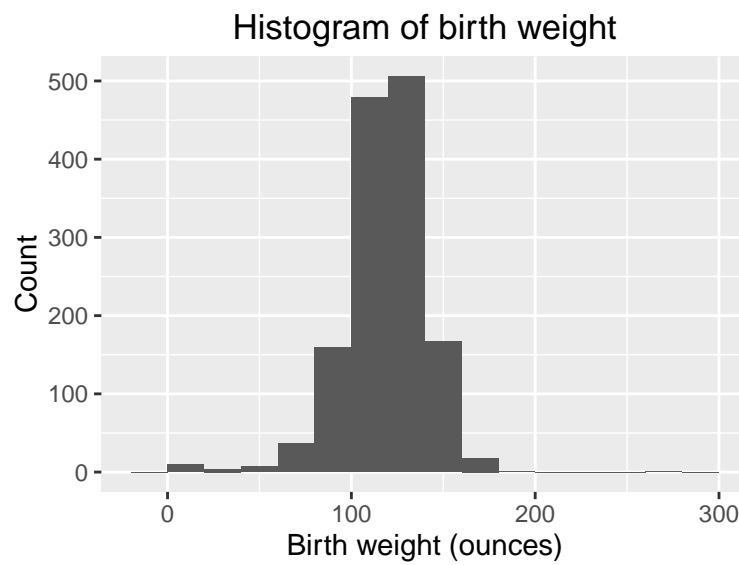


Figure 4: Histogram of birth weight (in ounces), using `ggplot` and `bin width = 20`