

TUFTS UNIVERSITY
Medical School

Prof. M. Bianconi
Course CMBA0264-01; Statistics with Applications; Summer 2009
E-Mail: marcelo.bianconi@tufts.edu

Assignment II

Data sets and other materials posted at the course web page:
<http://www.tufts.edu/~mbiancon/CMBA0264-2009.html>

Due date: Tuesday, July 7, 2009 (by 6:00PM via e-mail)

The data set Data-PSII at the course web page has two columns: Column 1 is the birth weight of a child (in ounces); Column 2 is the number of cigarettes smoked per week by the mother during pregnancy; $n=1,388$.

1. What kind of variable is being analyzed in this case (continuous, discrete etc.)?
2. Would you describe the experiment with these data as based on a controlled experiment or an observational study?
3. Plot a histogram of the "birth weight of a child." Briefly comment. Plot a histogram of the "number of cigarettes smoked." Briefly comment.
4. Sort your data by "birth weight of a child". This should reorder your data from lowest to highest. (sort bw).
Now split your sample into two sub-samples; low birth weight (the first 300 observations after you've sorted the data, $_n \leq 300$) and high birth weight (the last 300 observations, $_n \geq 1088$).
Repeat #3 by making histograms for both sub samples:

```
hist bw if _n <= 300;  
hist bw if _n >= 1088  
hist cigs if _n <= 300;  
hist cigs if _n >= 1088
```
5. Discuss in words whether the evidence from #4 allows you to reject the hypothesis that "mothers who smoke during pregnancy give birth to a low weight child."
6. What other factors could be potentially useful in explaining the child birth weight? Are there potential confounding factors?

7. Give the basic statistical properties of the data, say use the command *summarize (sum)* in STATA:
sum bw cigs

8. Use the twoway scatter command to graph *bw* on *cigs*: *twoway (scatter bw cigs)*. What kind of function do you observe, if any?

9. Use the command *regression (reg)* to run a linear regression of *bw* on *cigs* (*reg bw cigs*) What are the results of the linear regression (i.e. discuss the slope and the intercept)?

10. Use the command *predict* to obtain $E[bw|cigs]$ (*predict bwhat*).
Then, use the command *twoway scatter* to graph the actual and predicted values:
twoway scatter bw bwhat cigs, or *twoway (scatter bw cigs) (lfit bwhat cigs)*
Comment on the results.

11. Estimate the errors by subtracting the average birth weight from the observed birth weight and make a histogram of the errors.

```
gen er1 = bw - bwhat  
sum er1  
hist er1
```

Comment on the average, standard deviation and shape of the distribution of the errors

12. According to the results above:

(i) What would be the average birth weight if a mother smoked 5 cigarettes? Explain.

(ii) What is the change in average birth weight given that a mother smokes one additional cigarette? Explain.

(iii) Do you believe cigarette smoking is the only factor explaining the birth weight? Explain.