

TUFTS UNIVERSITY  
Medical School

Prof. M. Bianconi

Course CMBA0264-01; Statistics with Applications; Summer 2009

E-Mail: *marcelo.bianconi@tufts.edu*

**Assignment II - AK**

The data set Dataset-PS1-SF 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=1388.

1. What kind of variable is being analyzed in this case (continuous, discrete etc.)?

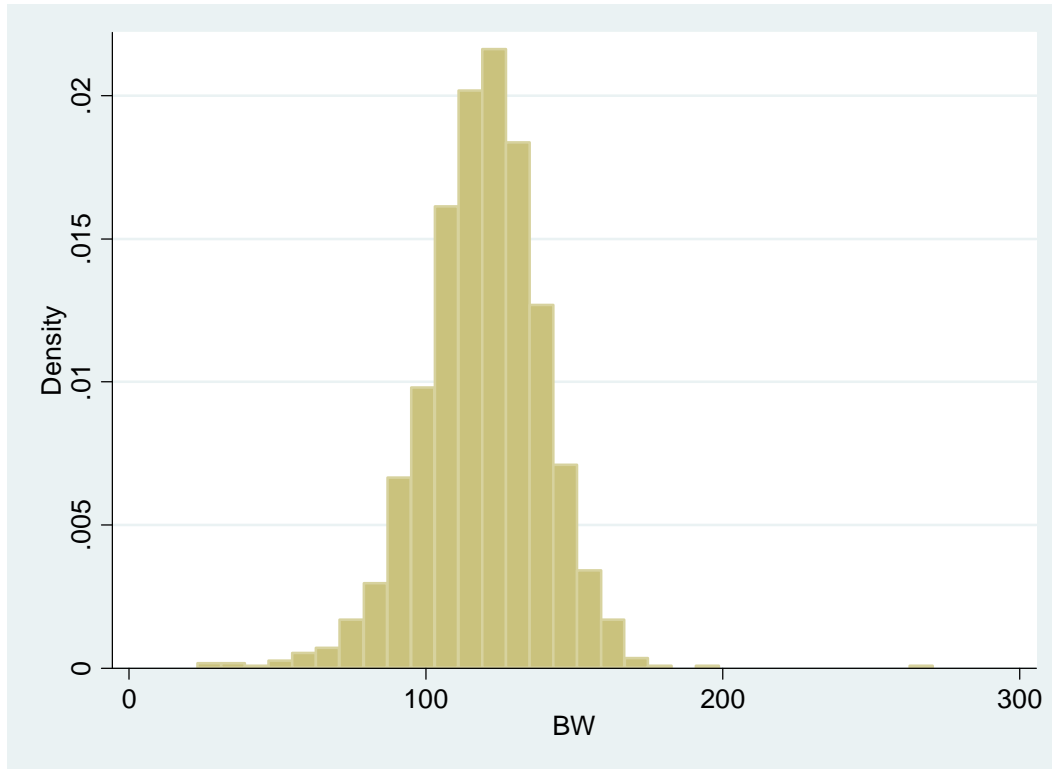
Birth weight: continuous variable (measurement)

Number of cigarettes smoked: discrete variable (counting)

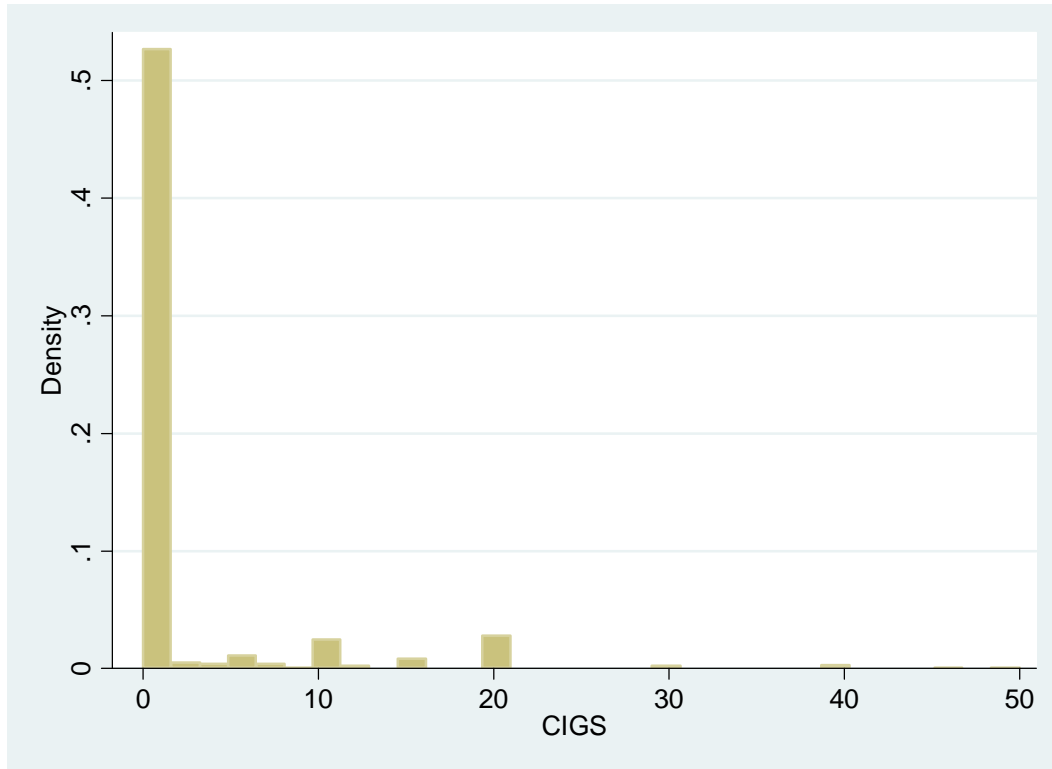
2. Would you describe the experiment with these data as based on a controlled experiment or an observational study?

This is an observational study as opposed to a controlled experiment that would divide treatment and control groups etc.

3. Plot a histogram of the "birth weight of a child." Briefly comment. Plot a histogram of the "number of cigarettes smoked." Briefly comment.



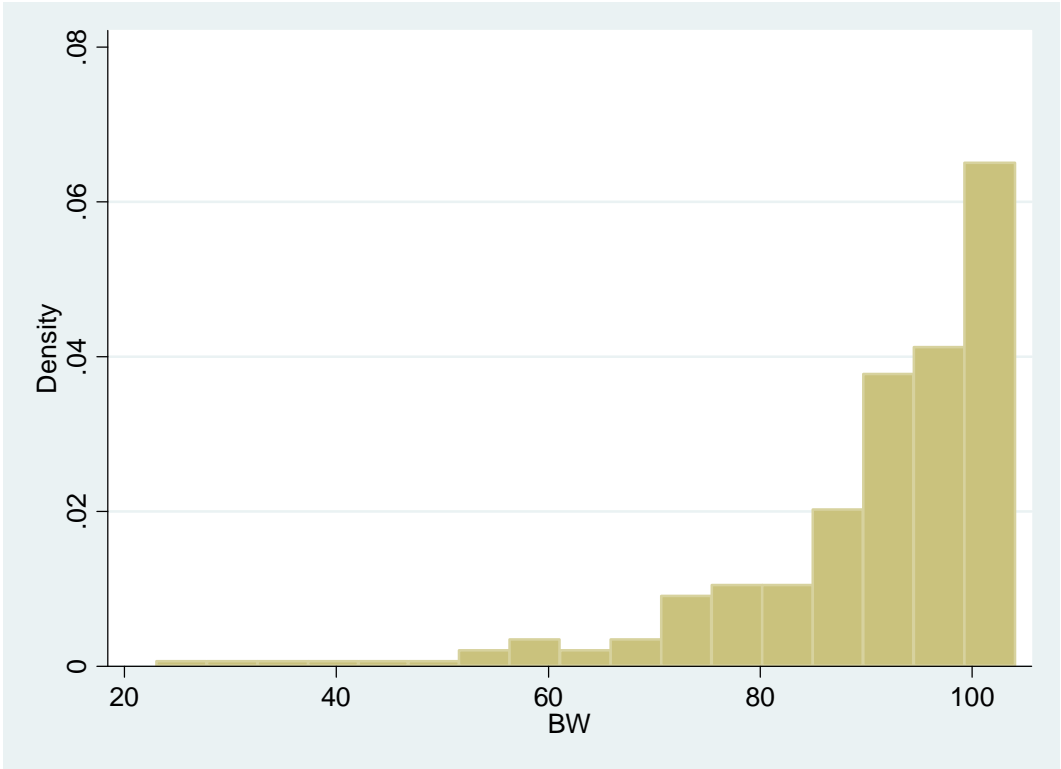
Distribution of birth weight approximately follows the normal distribution. Most observations are in the range of between 90 and 150.



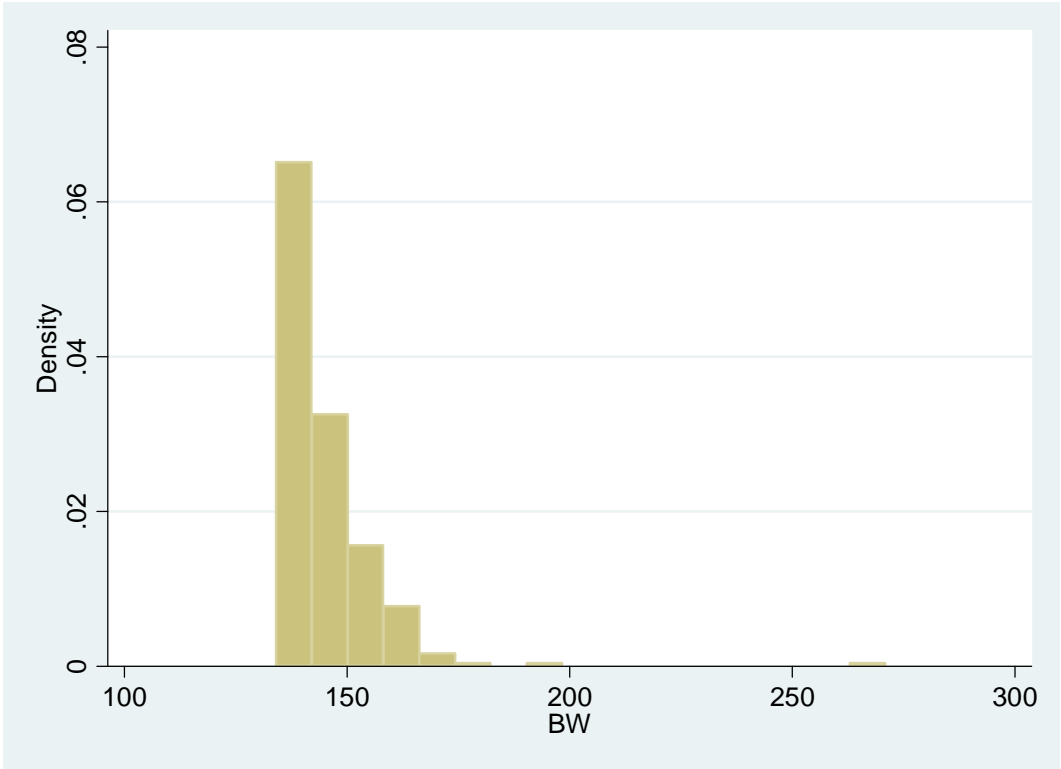
Distribution of number of cigarettes smoked has a long tail to the right. Most mothers did not smoke at all during the pregnancy (mode is zero).

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:

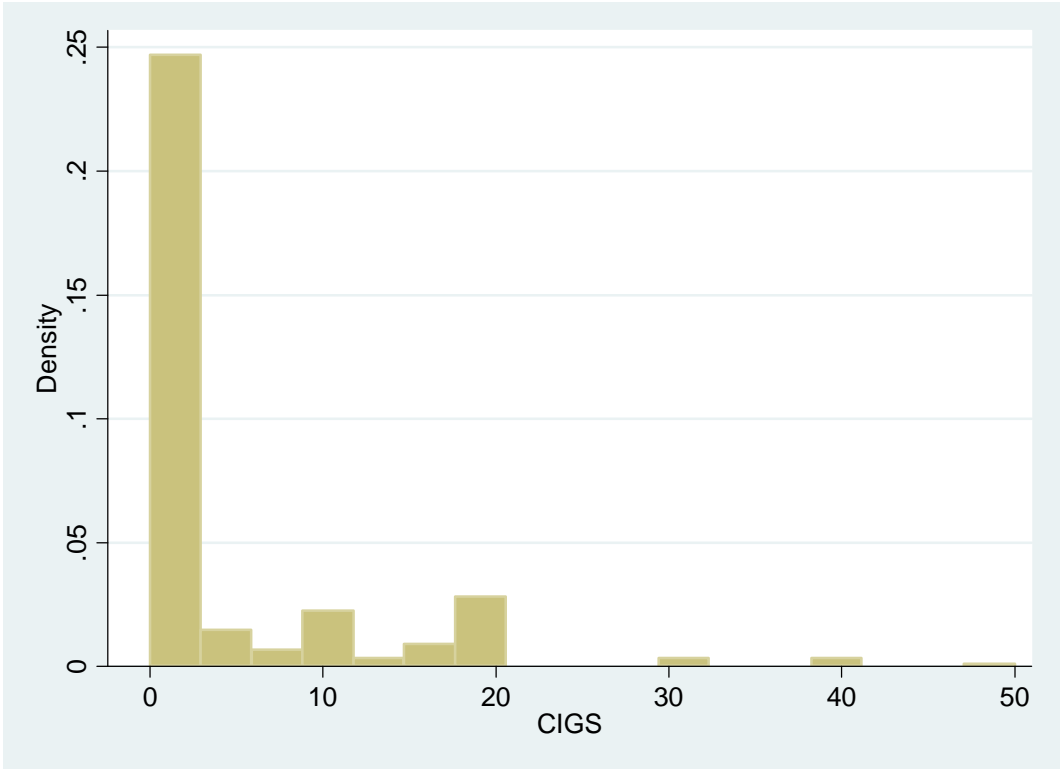
```
hist bw if  $\_n \leq 300$ ;
```



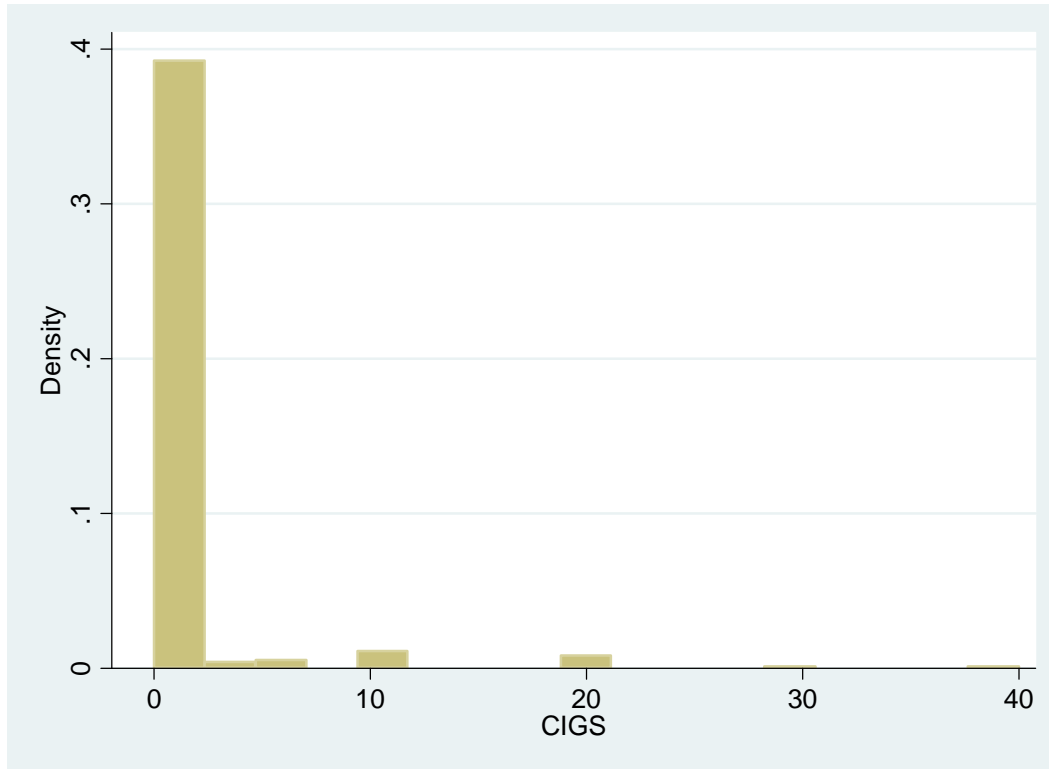
hist bw if \_n>=1088



hist cigs if \_n<=300;



hist cigs if  $_n \geq 1088$



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.”

It appears that there are more significant spikes in the distribution of the number of cigarettes smoked for the low birth weight group. However, a definite conclusion cannot be reached because we are missing the whole middle part of the distribution.

6. What other factors could be potentially useful in explaining the child birth weight? Are there potential confounding factors?

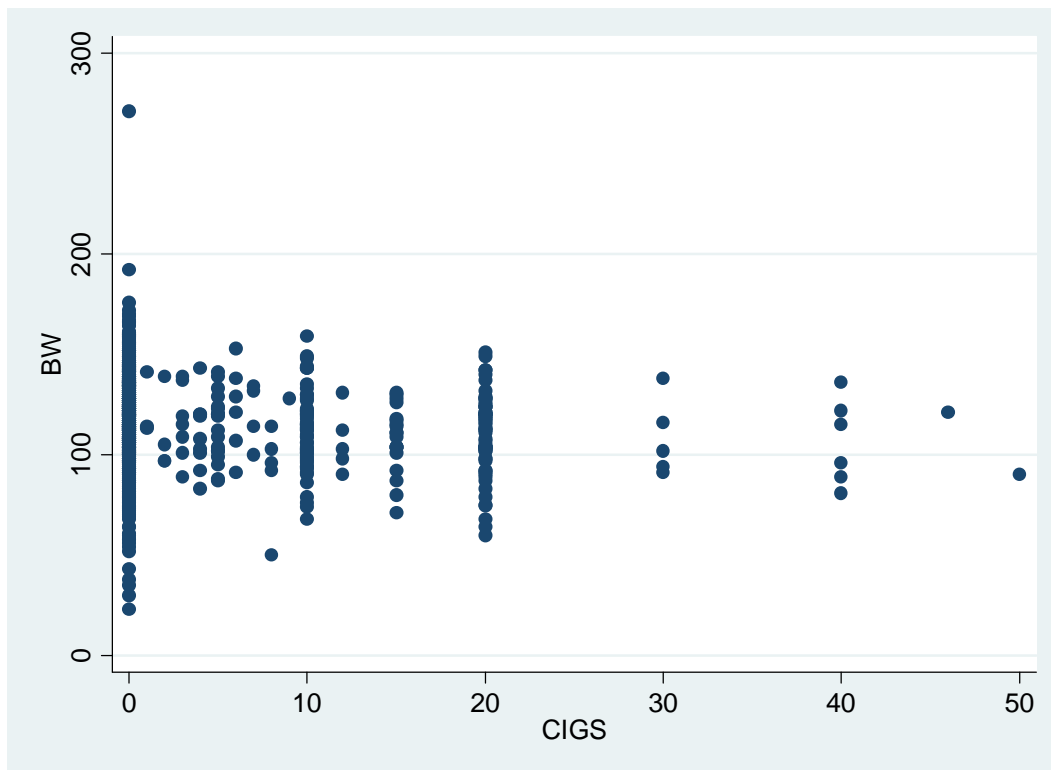
Mother’s overall health and nutrition, prenatal care, the size of mother’s body, etc. These could be potential confounding factors such as single parent vs. two-parent households; socio-economic status.

7. Return to the whole sample and give the basic statistical properties of the data, say use the command *summarize (sum)* in STATA: `sum bw cigs`

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bw	1388	118.6996	20.35396	23	271
cigs	1388	2.087176	5.972688	0	50

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



The slope is slightly trending downwards indicating that more cigs are associated with low birth weight.

9. Use the command *regression* (*reg*) to run a linear regression of *bw* on *cigs* (*reg bw cigs*) What are the results?

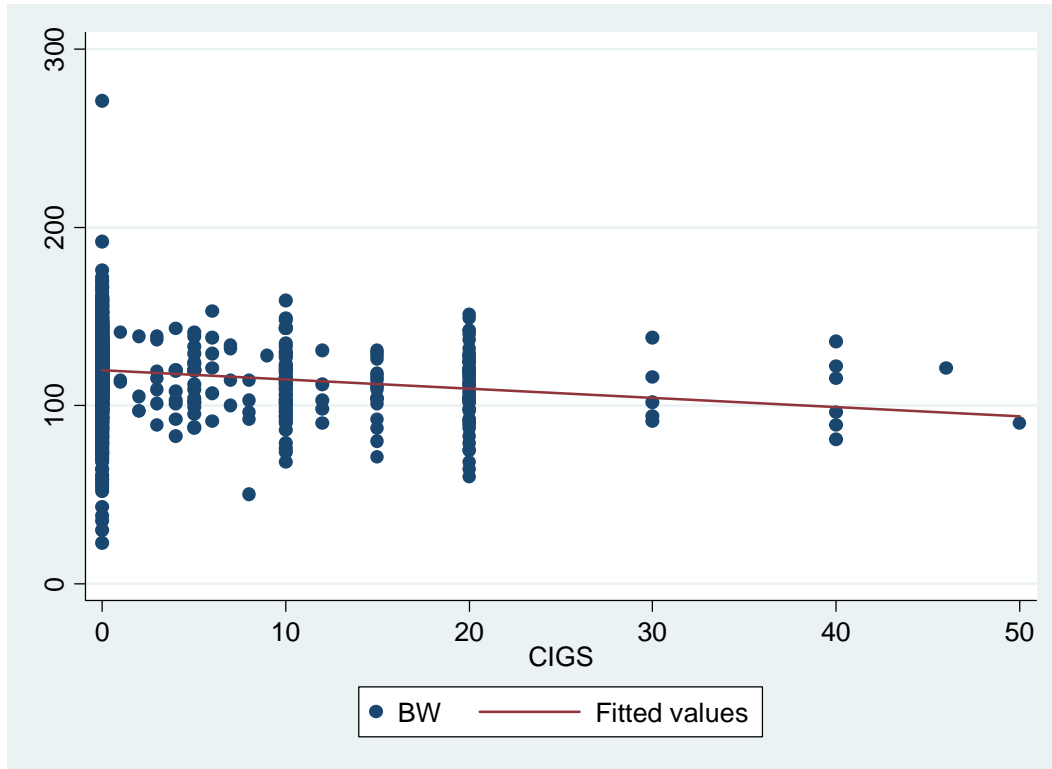
```
. reg bw cigs, r

Linear regression                               Number of obs =    1388
                                                F( 1, 1386) =    34.29
                                                Prob > F      =    0.0000
                                                R-squared    =    0.0227
                                                Root MSE    =    20.129

-----+-----
           |               Robust
           |               Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
    cigs   |   -.5137721   .0877334    -5.86   0.000   - .6858767   - .3416675
    _cons  |   119.7719   .5745494   208.46   0.000    118.6448    120.899
-----+-----
```

The 95% CI for the slope is [-0.6858, -0.3416] indicating a downward slope: more cigarettes imply lower birth weight. The 95% CI for the constant is [118.6448, 120.899] indicating the average birth weight for zero cigarettes smoked. (Note that I used the qualifier , r after the reg command; we'll see in class the advantages of using the qualifier).

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 e1 = bw - bwhat
```

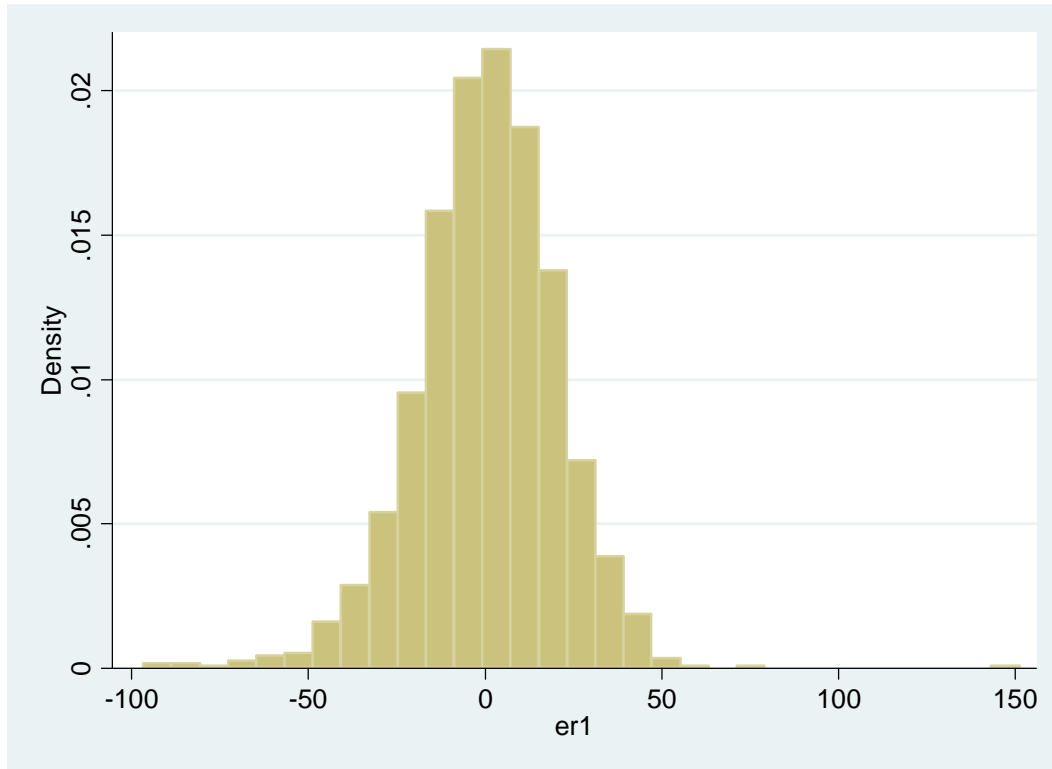
```
hist b1
```

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

```
. sum e1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
e1	1388	-3.00e-06	20.12132	-96.7719	151.2281

Mean is zero, standard deviation is very close to standard deviation of birth weight.



Almost (approximately) normal with mean zero.

12. According to the results:

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

$$E[bw|cigs=5] = 119.7719 - .5137721 * 5 = 117.2030$$

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

$$\frac{\text{Change } E[bw | \cdot]}{\text{Change } cigs} = -0.5138, \text{ approximate } \frac{1}{2} \text{ pound less birth weight.}$$

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

No, there are several omitted variables (As mentioned in 7: Mother's overall health and nutrition, prenatal care, the size of mother's body, single parent vs. two-parent households; socio-economic status). or Blood glucose level before and after ingestion of glucose load, age, body mass index, % of change in weight during pregnancy, height, gestational age, parity, fetal sex,

---

Bianconi, July 2009