# HW week 10

## w203: Statistics for Data Science

1. Recall that the slope coefficient in a simple regression of $Y_i$ on $X_i$ can be expressed as,

$$\beta_1 = \frac{c\hat{o}v(X_i, Y_i)}{v\hat{a}r(X_i)}$$

Suppose that you were to add a random variable, $M_i$, representing measurement error, to each $X_i$. You may assume that $M_i$ is uncorrelated with both $X_i$ and $Y_i$. You then run a regression of $Y_i$ on $X_i + M_i$ instead of on $X_i$. Does the measurement error increase or decrease your slope coefficient?

---

The file bwght.RData contains data from the 1988 National Health Interview Survey. It was used by J Mullahy for a 1997 paper ("Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," Review of Economics and Statistics 79, 596-593.) and provide by Wooldridge. You will use this data to examine the relationship between cigarette smoking and a child's birthweight.

```r
load("bwght.RData")
```

1. Examine the dependent variable, infant birth weight in ounces (bwght) and the independent variable, the number of cigarettes smoked by the mother each day during pregnacy (cigs).

```r
summary(data)
```

```
##      faminc          cigtax         cigprice         bwght      
##  Min.   : 0.50   Min.   : 2.00   Min.   :103.8   Min.   : 23.0  
##  1st Qu.:14.50   1st Qu.:15.00   1st Qu.:122.8   1st Qu.:107.0  
##  Median :27.50   Median :20.00   Median :130.8   Median :120.0  
##  Mean   :29.03   Mean   :19.55   Mean   :130.6   Mean   :118.7  
##  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.   :3.135   Min.   : 1.438  
##  1st Qu.:1.0000   1st Qu.: 0.000   1st Qu.:4.673   1st Qu.: 6.688  
##  Median :1.0000   Median : 0.000   Median :4.787   Median : 7.500  
##  Mean   :0.7846   Mean   : 2.087   Mean   :4.760   Mean   : 7.419  
##  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
##
```

```
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)
```

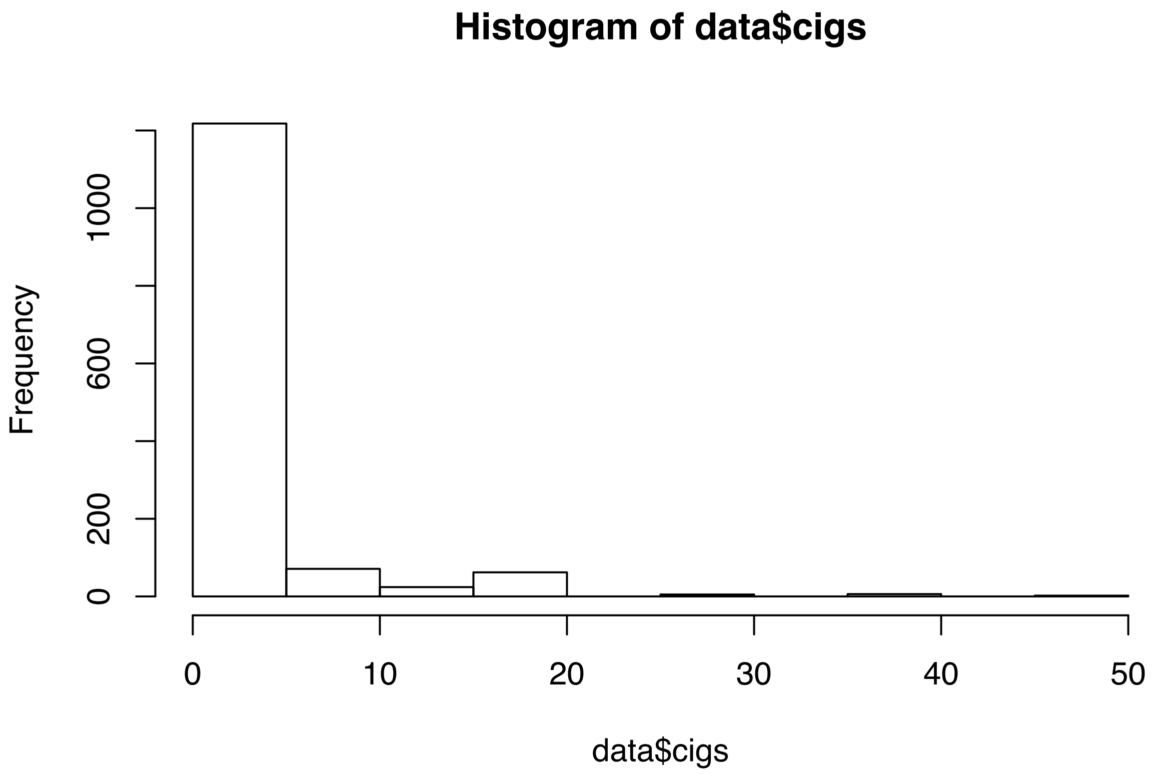
```
summary(data$cigs)
```

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

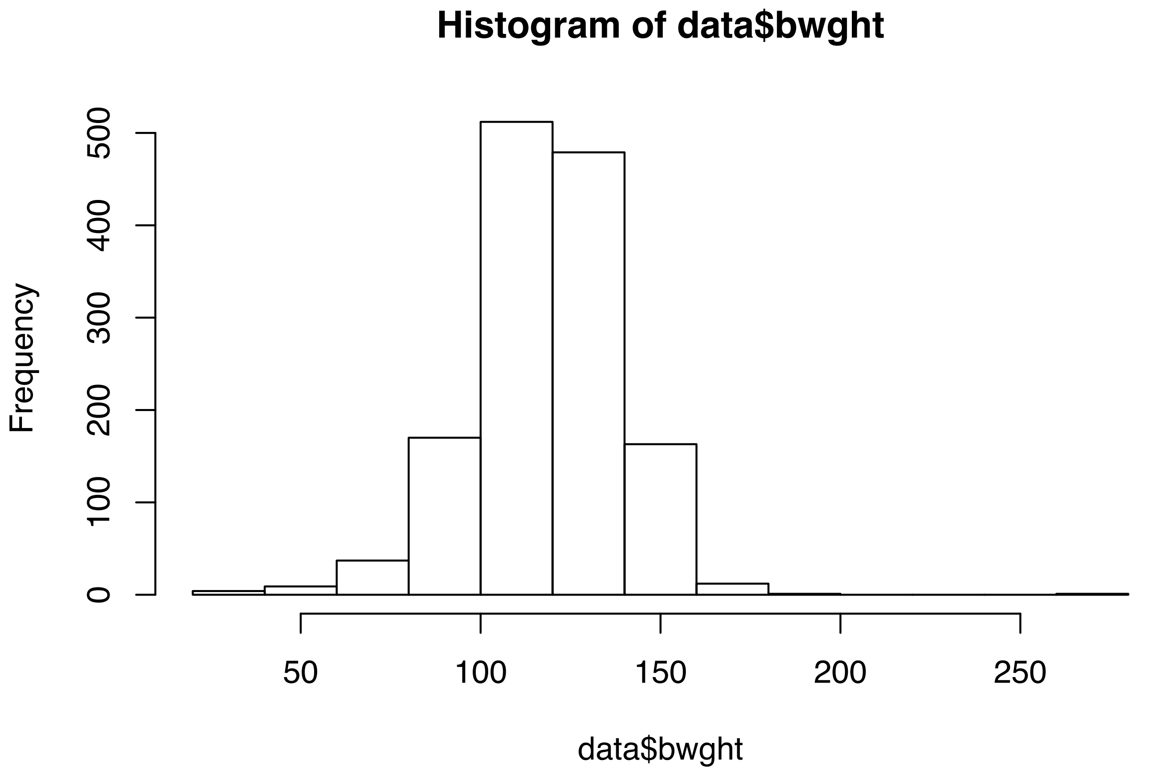
```
summary(data$bwghtlbs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.438   6.688   7.500   7.419   8.250  16.940
```

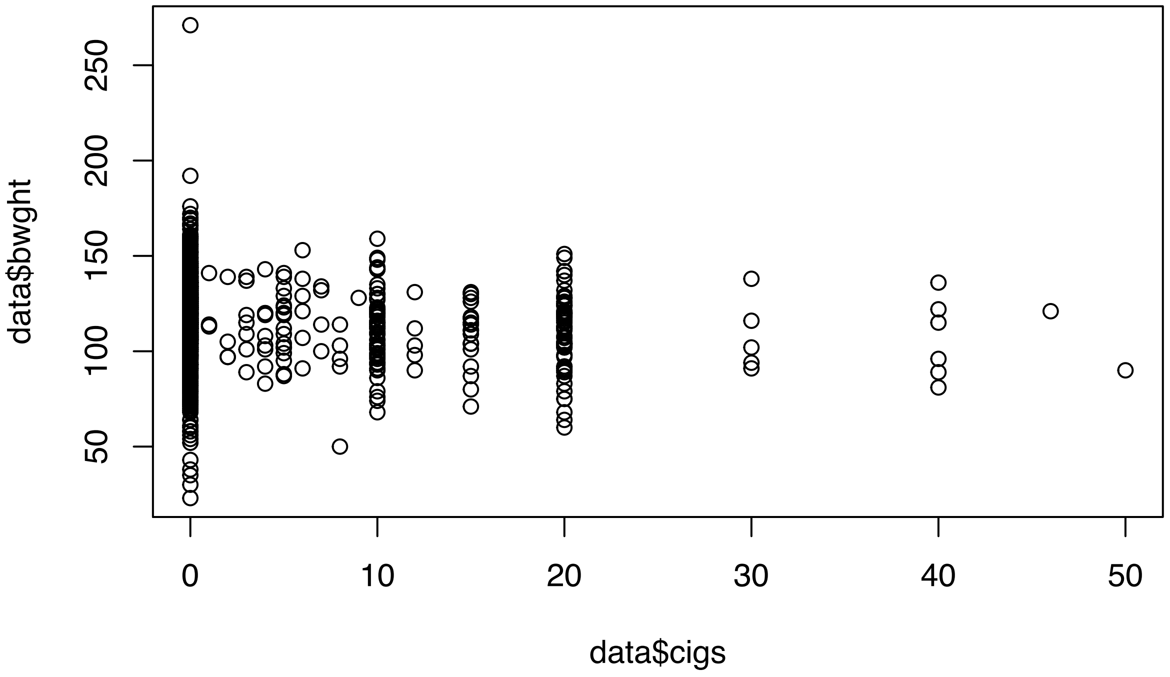
```
hist(data$cigs)
```



```r
hist(data$bwght)
```

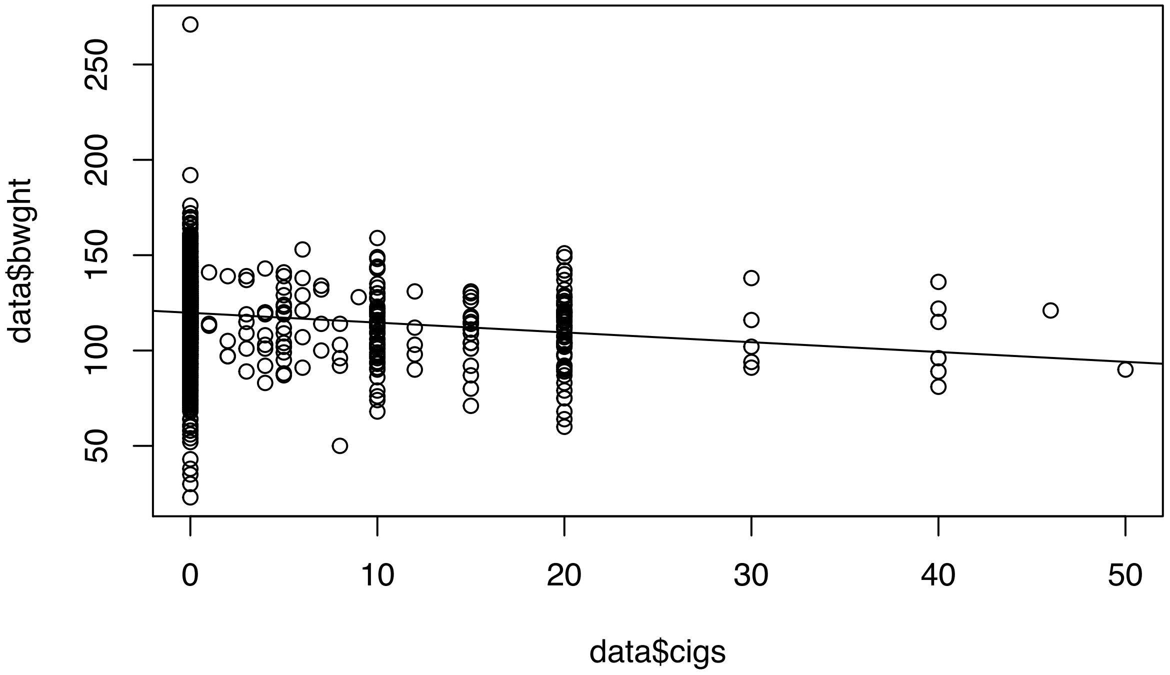


```r
plot(data$cigs, data$bwght)
```



2. Fit a linear model that predicts bwght as a function of cigs. Superimpose your regression line on a scatterplot of your variables.

```
m = lm(bwght ~ cigs, data = data)
plot(data$cigs, data$bwght)
abline(m)
```



3. Examine the coefficients of your fitted model. Explain, in particular, how to interpret the slope coefficient on cigs. Is it practically significant?

```
coef(m)
```

```
## (Intercept)        cigs
## 119.7719004  -0.5137721
```

```
# each cigarette smoked during pregnancy is associated with about half an ounce lower birthweight.
```

4. Write down the two moment conditions for this regression. Use R to verify that they hold for your fitted model.
5. Does this simple regression capture a causal relationship between smoking and birthweight? Explain why or why not.
6. Does your scatterplot show evidence of measurement error in cigs? If so, what does this say about the true relationship between cigarettes and birthweight?
7. Using your coefficients, what is the predicted birthweight when cigs is 0? When cigs is 20?
8. Use R's predict function to verify your previous answers. You may insert your linear model object into the command below.

```
predict(your_lm_object , data.frame(cigs = c(0, 20) ) )
```

9. To predict a birthweight of 100 ounces, what would cigs have to be?