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,

$$\beta_1 = \frac{\hat{cov}(X_i, Y_i)}{\hat{var}(X_i)}$$

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?

Since we know that M_i is uncorrelated to X_i:

$$\beta_1' = \frac{\hat{cov}(X_i + M_i, Y_i)}{\hat{var}(X_i + M_i)} = \frac{\hat{cov}(X_i, Y_i) + \hat{cov}(M_i, Y_i)}{\hat{var}(X_i) + \hat{var}(M_i)}$$

And that M_i is uncorrelated to Y_i

$$\beta_1' = \frac{\hat{cov}(X_i, Y_i)}{\hat{var}(X_i) + \hat{var}(M_i)}$$

Which, as long as M_i is not always the same, decreases the slope coefficient β_1

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.

```
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).

```
length(data$bwght)
```

```
## [1] 1388
```

```
summary(data$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.0   107.0   120.0   118.7   132.0   271.0
```

```
length(data$cigs)
```

```
## [1] 1388
```

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

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

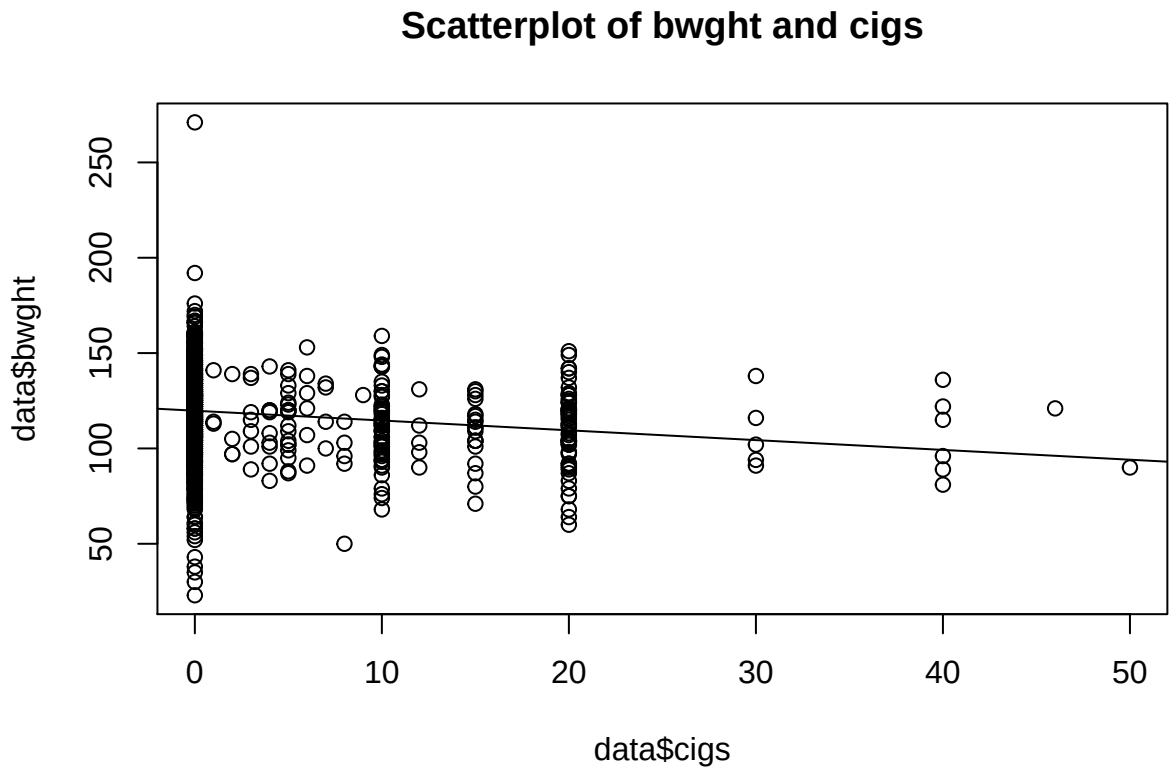
The mean and range for both variables look normal.

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

```
m1 <- lm(data$bwght ~ data$cigs)
m1
```

```
##
## Call:
## lm(formula = data$bwght ~ data$cigs)
##
## Coefficients:
## (Intercept)    data$cigs
##    119.7719      -0.5138
```

```
plot(data$cigs, data$bwght, main = "Scatterplot of bwght and cigs")
abline(m1)
```



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

```
m1$coefficients
```

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

Our estimator for the slope coefficient is -0.51 and the estimator for the intercept is 119.77. The estimated slope coefficient means that every additional cigarette/day that the preganant mom smokes during pregancy is associated with a 0.5 ounce decrease in the infant birth weight, which seems practially significant.

4. Write down the two moment conditions for this regression. Use R to verify that they hold for your fitted model.

1) $\mathrm{E}(u_i) = 0$
2) $\mathrm{cov}(u_i, x_i) = 0$

```r
u_i = data$bwght - m1$fitted.values
mean(u_i)
```

```
## [1] 4.260881e-15
```

```r
cov(u_i, data$cigs)
```

```
## [1] -5.013781e-14
```

The two moment conditions hold for our fitted model

5. Does this simple regression capture a causal relationship between smoking and birthweight? Explain why or why not.

No; Pregenant mothers who smoke different amount of cigarettes have many differences than just the cigarette consumption. Hence, it doesn't mean that if we manipulate the daily cigarette consumption for a pregnant mother to increase by 1/day, we are going to see a decrease of 0.5 ounce for her infant birth weight.

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?

Yes, we can see that there are clustered observations at 20, 30, and 40 but no where between 20-30 or 30-40. This could indicate that there is a lack of precision in the surevey data. However, according the potentially rounded cigs, we still see a negative correlation between cigs and bwght.

7. Using your coefficients, what is the predicted birthweight when cigs is 0? When cigs is 20?

```r
#when cigs = 0
y_0 = as.numeric(m1$coefficients[1])
y_0
```

```
## [1] 119.7719
```

```r
#when cigs = 20
y_20 = as.numeric(m1$coefficients[1] + 20*m1$coefficients[2])
y_20
```

```
## [1] 109.4965
```

8. Use R's predict function to verify your previous answers. You may insert your linear model object into the command below.

```r
y = data$bwght
x = data$cigs
predict(lm(y ~ x), data.frame(x = c(0, 20) ) )
```

This verifies the previous answers

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

```r
cigs_100 = (100 - as.numeric(m1$coefficients[1])) / as.numeric(m1$coefficients[2])
cigs_100
```

```
## [1] 38.4838
```

Cigs would have to be 38.48