HW week 10

w203: Statistics for Data Science

Adam Yang

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?

Answer: First we can write out our new slope equation:

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

Then we can simplify the numerator with the property: $cov(x, y + z) = cov(x, y) + cov(x, z)$. We can also simplify the denominator by the property: $var(x + y) = var(x) + var(y) + 2cov(x, y)$

$$\beta_1 = \frac{\hat{cov}(X_i, Y_i) + \hat{cov}(M_i, Y_i)}{\hat{var}(X_i) + \hat{var}(M_i) + 2\hat{cov}(X_i, M_i)}$$

We are assuming there is no correlation between M_i with both X_i and Y_i so $\hat{cov}(M_i, Y_i)$ and $\hat{cov}(X_i, M_i)$ both equal 0. Therefore, we are left with:

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

In the final β_1 value, we see that the denominator is increased by $\hat{var}(M_i)$. That means that the absolute value of the slope coefficient is decreased by the measurement error to be lower than it would've been based on the variance of the measurement error. To be clear, because only the denominator is increased, the slope would approach 0. If the slope was negative, the slope would increase and approach 0.

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 birth weight.

```
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 pregnancy (cigs).

```
X <- data$cigs
Y <- data$bwght
summary(X)
```

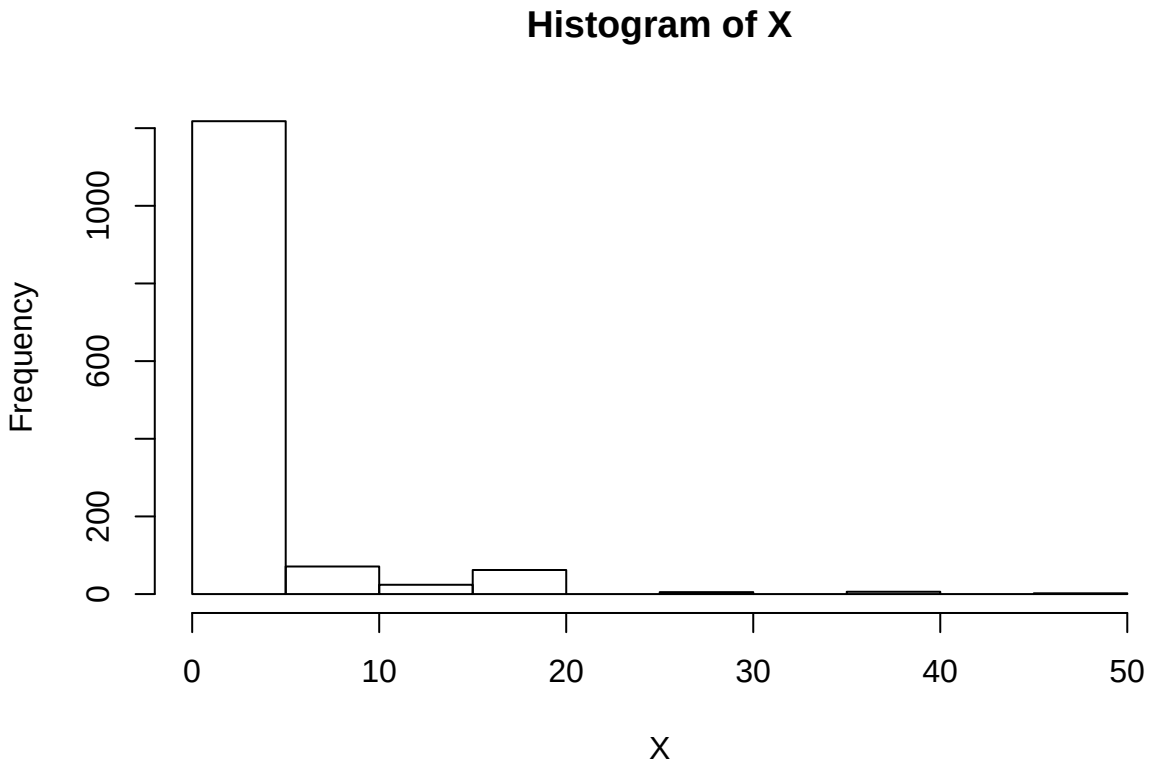
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.087   0.000  50.000
```

```
summary(Y)
```

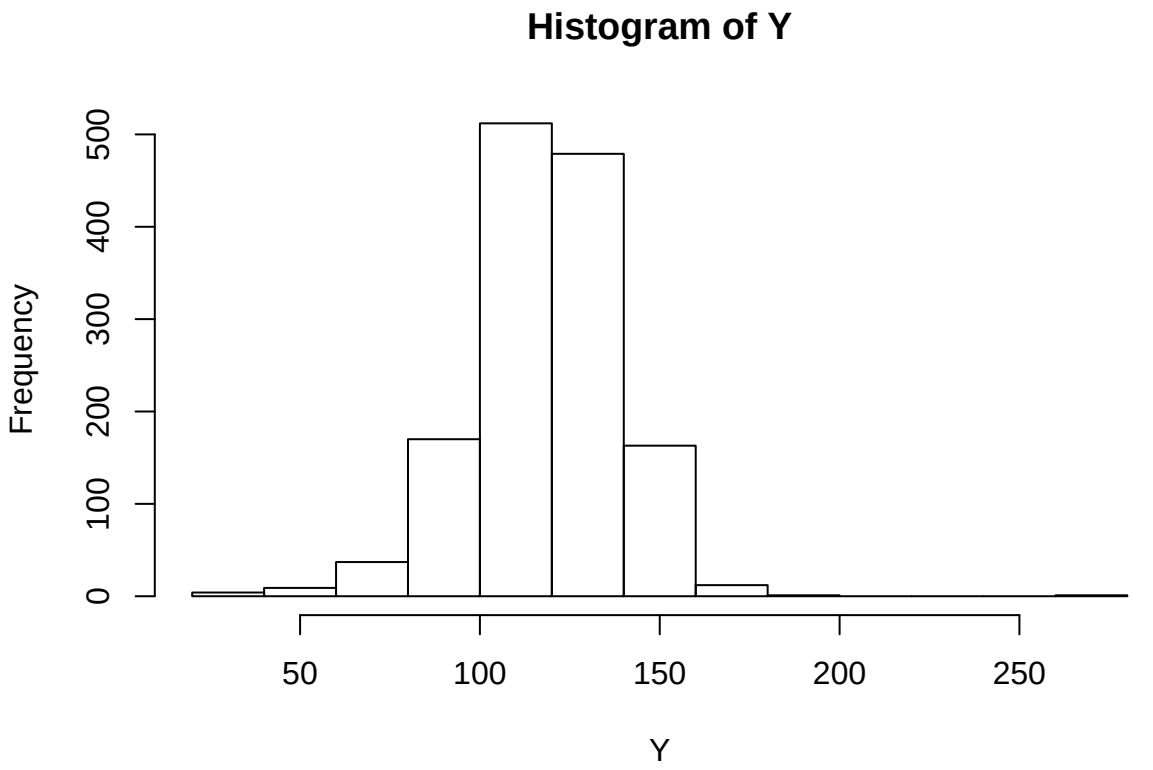
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.0   107.0   120.0   118.7   132.0   271.0
```

It looks like there is an extreme positive skew on the independent variable. This would make sense since I would believe most mothers would not be smoking during pregnancy. The dependent variable seems to be much more normally distributed. To make sure, let's look at the histograms of each variable.

```
hist(X)
```



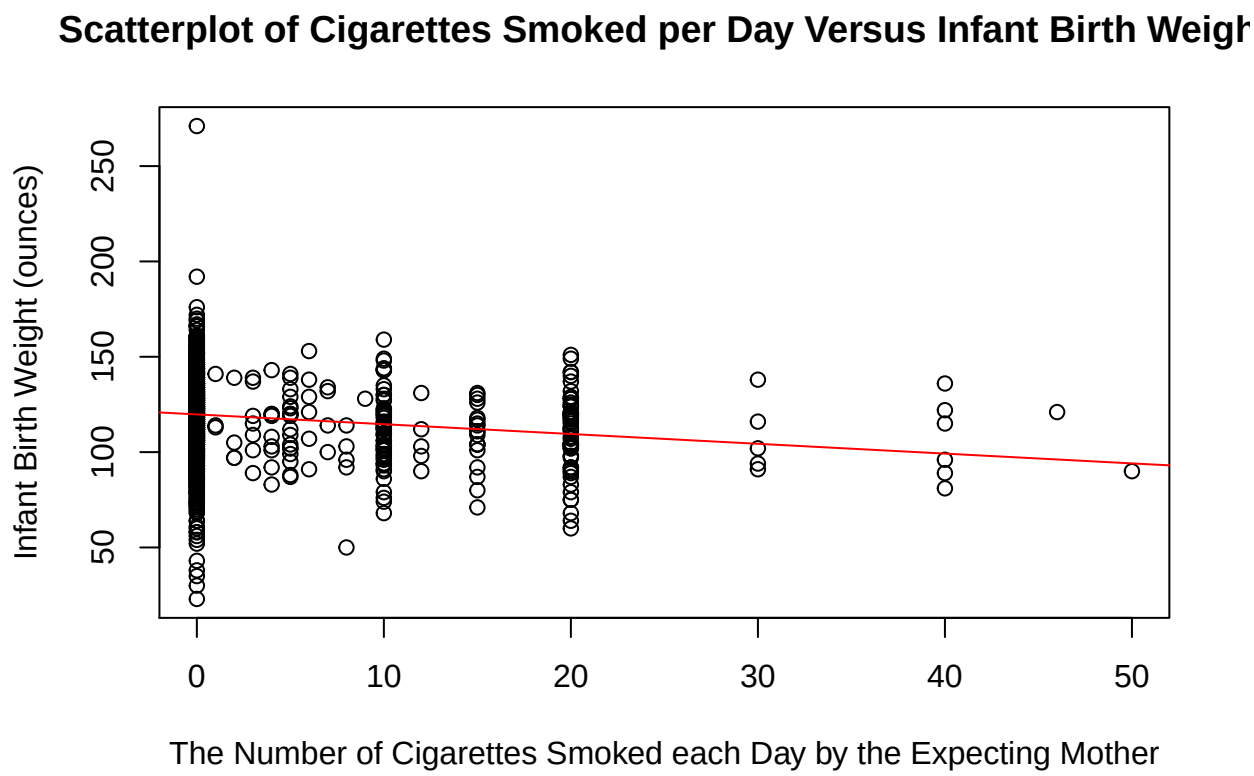

```
hist(Y)
```



As suspected, the cigs variable has a very strong positive skew, while the birth weight variable looks close to a normal distribution.

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

```
model <- lm(Y ~ X)
plot(X, Y, main="Scatterplot of Cigarettes Smoked per Day Versus Infant Birth Weight",
     xlab="The Number of Cigarettes Smoked each Day by the Expecting Mother",ylab="Infant Birth Weight
abline(model, col = "red")
```



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

```
model$coefficients
```

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

The slope coefficient on cigs is -0.51. That means if the mother smokes 1 more cigarette per day, that accounts for the baby being 0.5 ounces ligher. The practical significance of this very unclear to me because I have never had a baby of my own. I would think that if there is a negative slope at all, it would be significant to the mother because she would want the baby to be as healthy as possible. A friend of mine told me that a normal feeding amount for a baby is 4 ounces so 0.5 ounces is 12.5% of a baby's normal feeding amount which sounds kind of significant (especially if you're the baby's parent). However, there are so many variables that are not accounted for in this data set. This includes the ethnicity, sex and whether the baby is a twin or not. I would imagine those variables would cause the baby to be born with a wide spread of different weights. Lets look at the birth weights of babies whos mothers don't smoke during pregnancy:

```
summary(data[data$cigs == 0,]$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.0   108.0   121.0   120.1   133.0   271.0
```

```
sd(data[data$cigs == 0,]$bwght)
```

```
## [1] 20.26849
```

This shows that the range of all baby weights go from 23 to 271 ounces for babies whos mothers did not smoke during pregnancy. The standard deviation is 20.3 ounces. That means 0.5 ounces is only around 2.5% of the standard deviation. That seems very small. So maybe the slope is so small that there is not much practical significance to it. However, not being a parent or an expert on pregnancy and babies at birth, I do not know how significant the slope is.

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

Assume $Y_i = \beta_0 + \beta_1 X_i + u_i$ and that the CLT holds for each random variable. The two conditions for the method of moments are: $E(u_i) = 0$ and $cov(u_i, x_i) = 0$. To confirm this, we can look at the residuals of our model.

```
u = model$residuals
paste("mean(u) = ", mean(u))
```

```
## [1] "mean(u) =  -1.2928689099128e-16"
```

```
paste("cov(u,X) = ", cov(u, X))
```

```
## [1] "cov(u,X) =  7.03837848173681e-15"
```

As shown above, both $E(u_i) = 0$ and $cov(u_i, x_i) \approx 0$.

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

A simpel regression is not meant to capture causality but rather to see if the two variables are positively or negatively correlated. Therefore, I don't think the simple regression is evidence enough to assume causality between smoking and birth weight. On top of this, there are many variables that were not kept constant for the babies such as the baby's sex and ethnicity.

6. Does your scatter plot show evidence of measurement error in cigs? If so, what does this say about the true relationship between cigarettes and birth weight?

Yeah I think the cigs variable is an estimated variable by the mother rather than a recorded variable. This can be seen by the numbers being mostly clean numbers such as 0, 10, or 20. It is most likely approximated by the mother rather than a precise measurement. Therefore, there is likely measurement error involved with the cigs variable. If this is indeed the case, the slope coefficient is closer to zero than it would've been. It is possible that the true relationship between cigarettes and birth weight has a larger (more negative) slope.

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

```
paste("The predicted birth weight when cigs is 0 is", model$coefficients[1] + model$coefficients[2]*0)
```

```
## [1] "The predicted birth weight when cigs is 0 is 119.77190039835"
```

```
paste("The predicted birth weight when cigs is 20 is", model$coefficients[1] + model$coefficients[2]*20)
```

```
## [1] "The predicted birth weight when cigs is 20 is 109.496458541882"
```

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

```
predict(model , data.frame(X = c(0,20)))
```

```
##        1        2
## 119.7719 109.4965
```

The values obtained from the predict fucntion matches my calculated values.

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

```r
paste("To predict a birth weight of 100 ounces, cigs would have to be", (100 - model$coefficients[1])/m
```

```
## [1] "To predict a birth weight of 100 ounces, cigs would have to be 38.4837959759448"
```

It would seem the mother would have to smoke 38.5 cigarettes a day to predict a birth weight of 100 ounces.