HW week 10

w203: Statistics for Data Science

Tako Hisada

11/19/2017

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?

SLR 4 assumes $E(u|x) = 0$. Therefore the expected value of M_i is 0 and therefore adding it should not affect the slope coefficient value.

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")
```

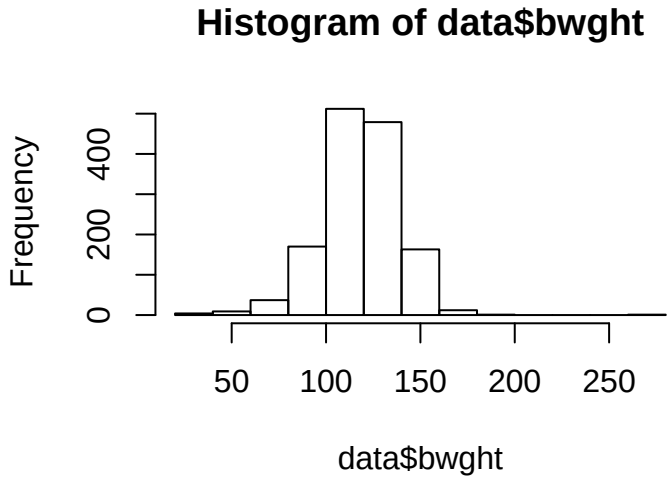
1. Examine the dependent variable, infant birth weight in ounces (bwght) and the independent variable, the number of cigarettes smoked by the mother each day during pregnacy (cigs).

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

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

```
bwght_n = length(data$bwght[!is.null(data$bwght) && data$bwght > 0])
bwght_mu = mean(data$bwght)
bwght_sd = sd(data$bwght)
```

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

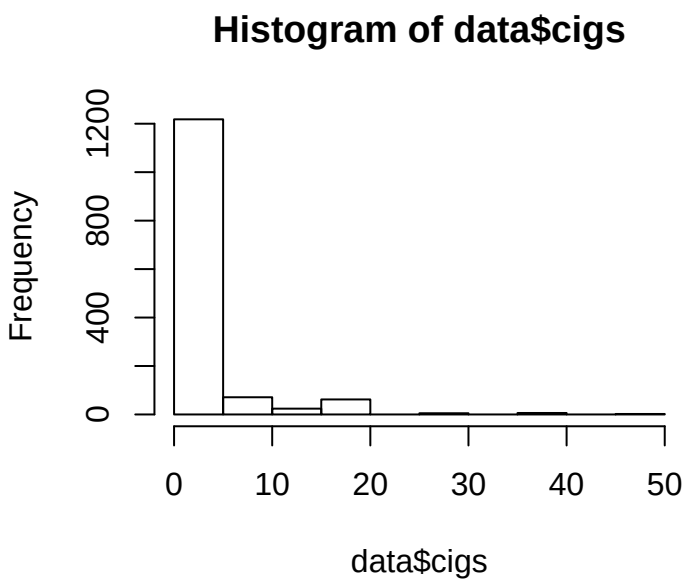


```r
summary(data$cigs)
```

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

```r
cigs_n = length(data$cigs[!is.null(data$cigs) && data$cigs >= 0])
cigs_mu = mean(data$cigs)
cigs_sd = sd(data$cigs)
```

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

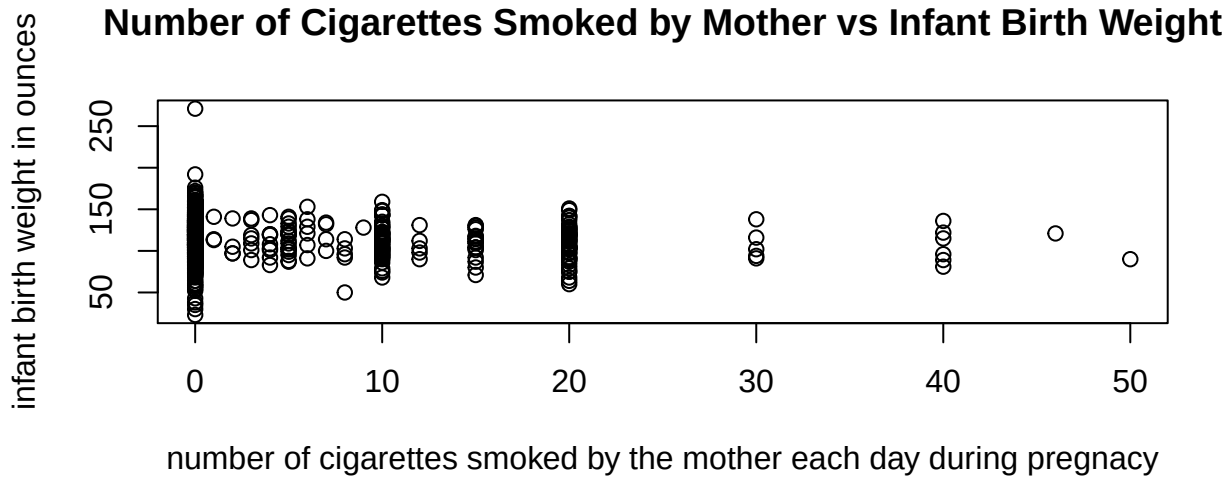


Both bwght and cigs do not contain any null nor invalid values (> 0 for bwght and $>= 0$ for cigs) and n is 1388. bwght has a fairly normal distribution whereas cigs is positively skewed.

	bwght	cigs
n	1388	1388
μ	118.6995677	2.0871758
σ	20.3539643	5.9726879

Generating a scatterplot with cigs and bwght variables

```r
plot(data$cigs, data$bwght,
     xlab="number of cigarettes smoked by the mother each day during pregnacy",
     ylab="infant birth weight in ounces",
     main="Number of Cigarettes Smoked by Mother vs Infant Birth Weight")
```



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

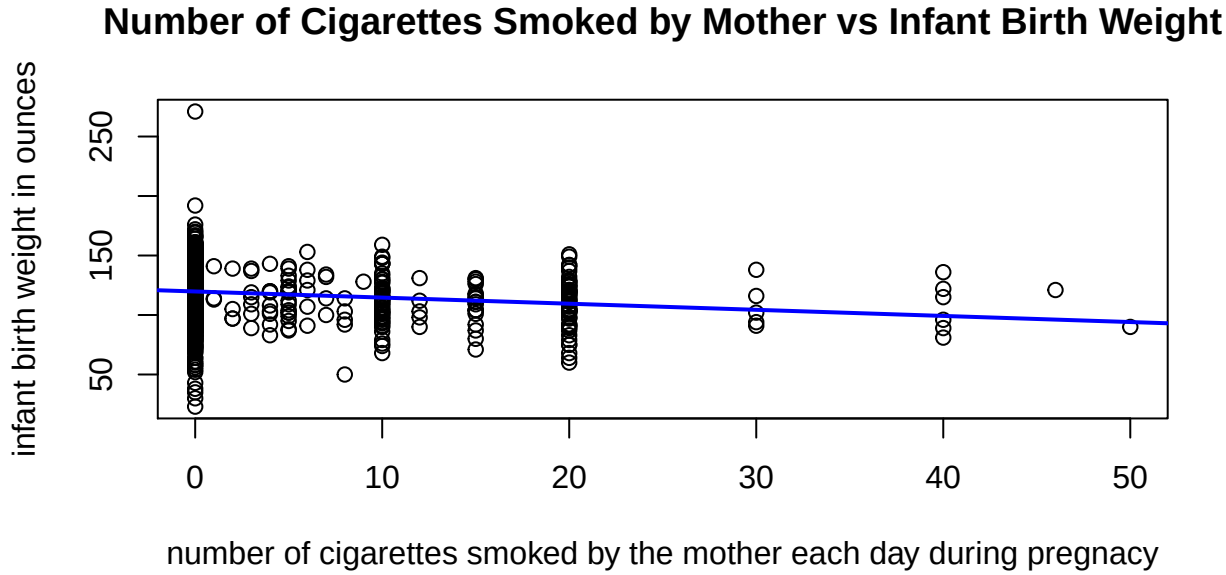
Let's create a simple regression model from data$cigs and data$bwght.

```
cigs = data$cigs
model_smoking <- lm(data$bwght ~ cigs)
model_smoking$coef
```

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

Now we will generate a scatterplot with data$cigs being our x variable and data$bwght being our y variable then superimpose the regression line with the Y-intercept of 119.7719004 and the slope of -0.5137721 over it.

```
plot(data$cigs, data$bwght,
     xlab="number of cigarettes smoked by the mother each day during pregnacy",
     ylab="infant birth weight in ounces",
     main="Number of Cigarettes Smoked by Mother vs Infant Birth Weight")
abline(model_smoking,col="blue",lwd=2)
```



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

```
model_smoking$coef
```

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

The slope coefficient indicates that for each incremental cigarette that the mother smokes, the infant birth weight decreases by 0.5137721 ounces.

```
model_smoking_r_sq = summary(model_smoking)$r.square
```

The R-square is 0.0227291. This means only 2.2% of the variation in the infant birth weight is explained by the number of cigarettes smoked by the mother. Although the mother's smoking may have other health implications, as far as its impact on the infant birth weight is concerned, it is not practically significant.

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

Below are the 2 moment conditions:
1. $E(u_i) = 0$
2. $COV(u_i, x_i) = 0$

```
(u_mu = mean(model_smoking$residuals))
```

```
## [1] 2.545594e-17
```

```
(x_u_cov = cov(data$cigs, model_smoking$residuals))
```

```
## [1] -4.09992e-14
```

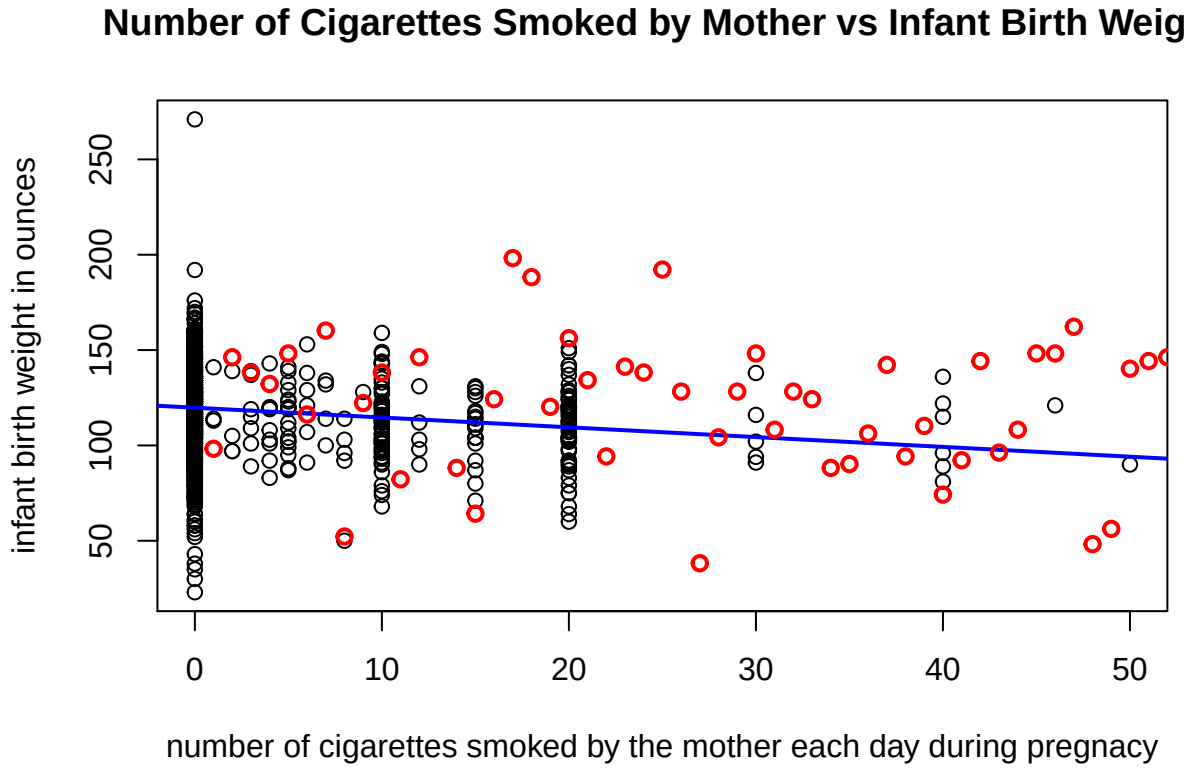
The above values are both very small and are essentially 0.

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

The model does not capture a causal relationship. Regression describe a dependent relationship in which one or more explanatory variables predict an outcome variable and therefore assumes a one-way link between X and Y however does not assume a causality. Rather, it is a more precise description of linearity that allows us to identify the slope of the line that best fits our data.

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?

```
plot(data$cigs, data$bwght,
     xlab="number of cigarettes smoked by the mother each day during pregnacy",
     ylab="infant birth weight in ounces",
     main="Number of Cigarettes Smoked by Mother vs Infant Birth Weight")
abline(model_smoking,col="blue",lwd=2)
points(data$bwght + model_smoking$residuals, col="red", lwd=2)
```



As can be seen in the scatterplot, we are seeing residuals for the infant birth weight values, indicating the sample values are off from the their fitted values. This may imply that there are some measurement error in cigs that are possibly accountable and therefore the true population model does not exactly match the linear model we have come up with.

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

```
x = 0
(cig_0 = model_smoking$coef[1] + model_smoking$coef[2] * x)
```

```
## (Intercept)
##    119.7719
```

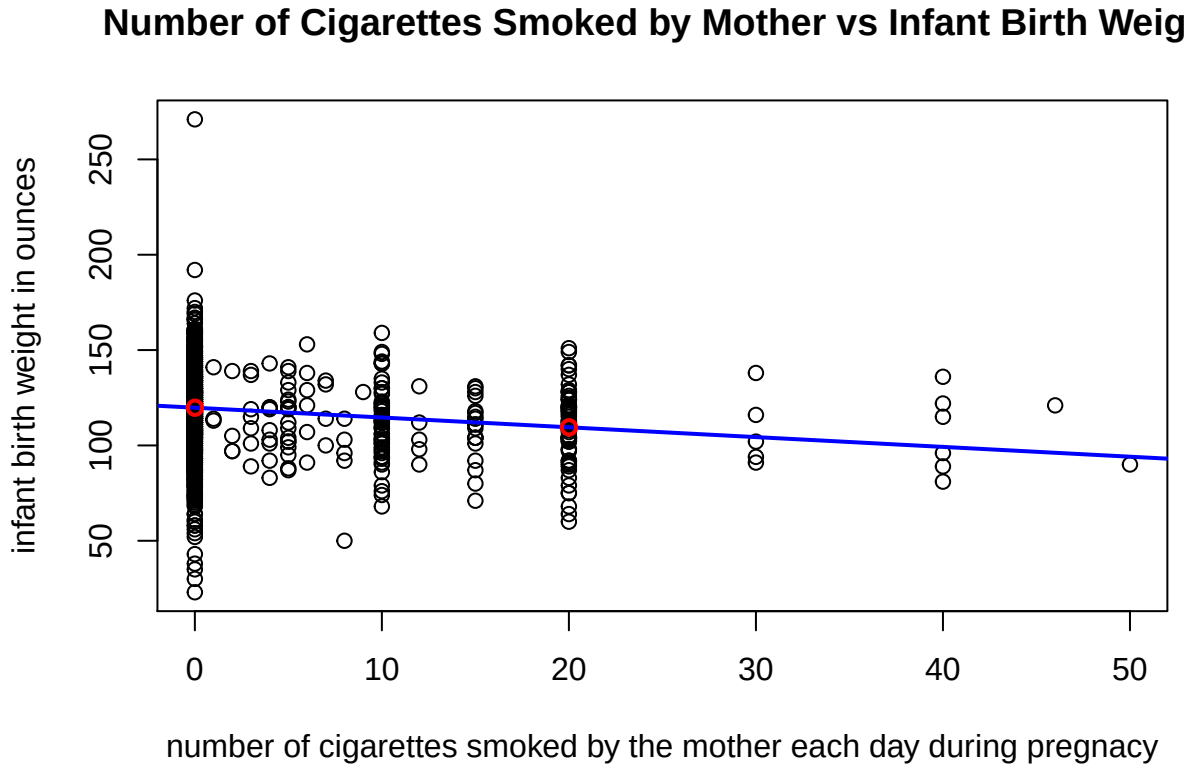
The predicted birthweight when cigs is 0 is 119.7719004.

```
x = 20
(cig_20 = model_smoking$coef[1] + model_smoking$coef[2] * x)
```

```
## (Intercept)
##    109.4965
```

The predicted birthweight when cigs is 20 is 109.4964585.

```
plot(data$cigs, data$bwght,
     xlab="number of cigarettes smoked by the mother each day during pregnacy",
     ylab="infant birth weight in ounces",
     main="Number of Cigarettes Smoked by Mother vs Infant Birth Weight")
abline(model_smoking,col="blue",lwd=2)
points(c(0, 20), c(cig_0, cig_20), col="red", lwd=2)
```



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

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

The predicted values 119.7719004, 109.4964585 match what we got in Q7 (119.7719004 and 109.4964585).

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

$$100 = \beta_0 + \beta_1 * x_i = 119.7719004 + -0.5137721 * x_i$$

$$-0.5137721 * x_i = 100 - 119.7719004$$

```
q9 = (100 - model_smoking$coef[1])/model_smoking$coef[2]
```

$$x_i = 38.5$$

The mother would have to be smoking 39 cigarrets each day for the predicted birth hweight to be 100 ounces.