HW week 10

w203: Statistics for Data Science

Ted Pham

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?

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

Since M_i is uncorrelated with both X_i, Y_i,

$$\hat{cov}(M_i, Y_i) = \hat{cov}(M_i, X_i) = 0$$

$$

_1 = $$

The slope coefficient will decrease.

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")
ls()
```

```
## [1] "data" "desc" "self"
```

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).

```
names(data)
```

```
##  [1] "faminc"   "cigtax"   "cigprice" "bwght"    "fatheduc" "motheduc"
##  [7] "parity"   "male"     "white"    "cigs"     "lbwght"   "bwghtlbs"
## [13] "packs"    "lfaminc"
```

```
names(desc)
```

```
## [1] "variable" "label"
```

```
bwght = data$bwght
cigs = data$cigs
which(is.na(bwght))
```

```
## integer(0)
```

```
which(is.na(cigs))
```

```
## integer(0)
```

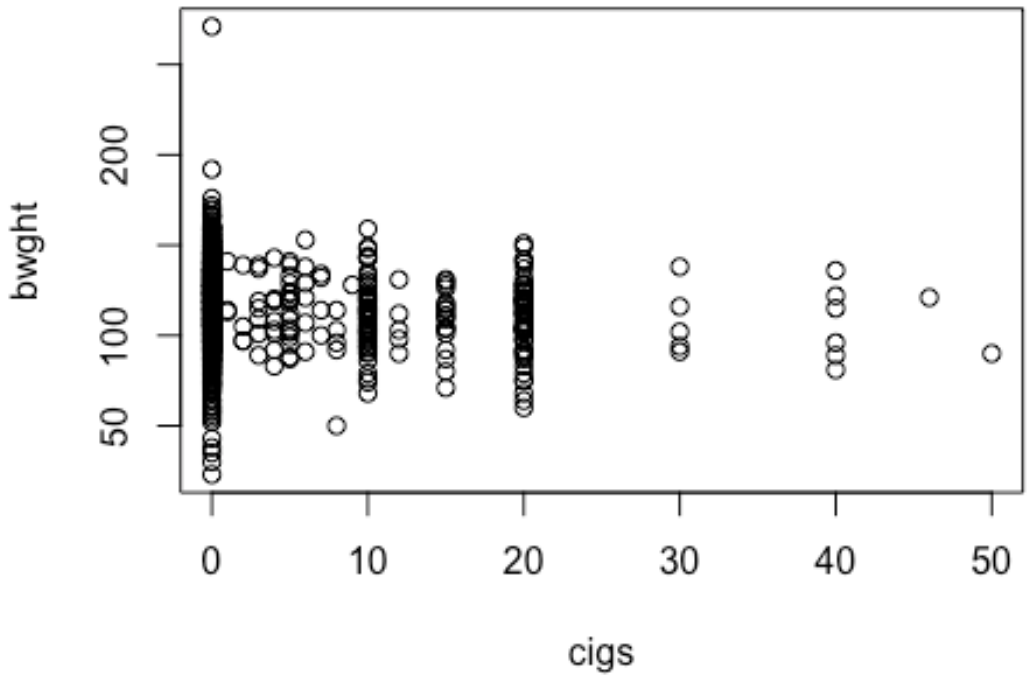
```
summary(bwght)
```

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

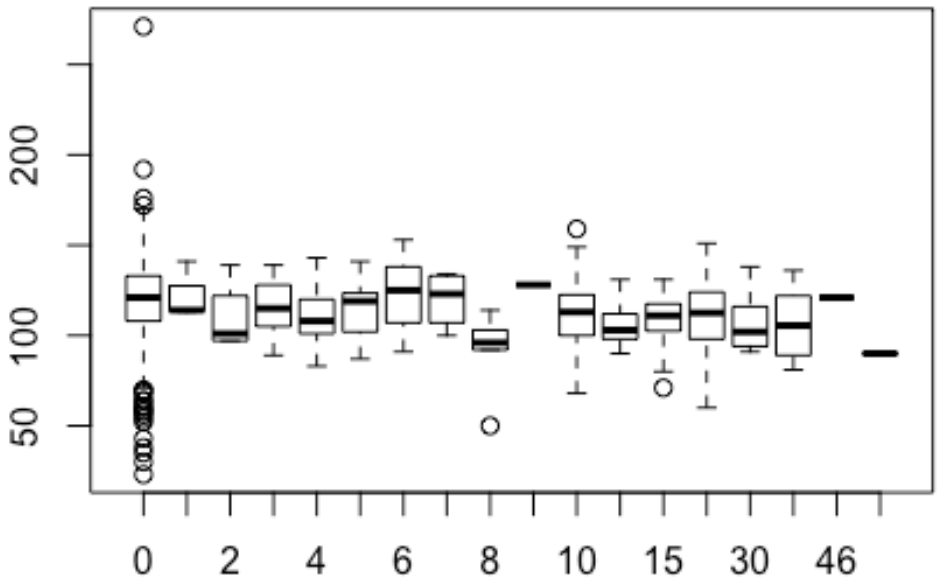
```
summary(cigs)
```

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

```
plot(cigs,bwght)
```

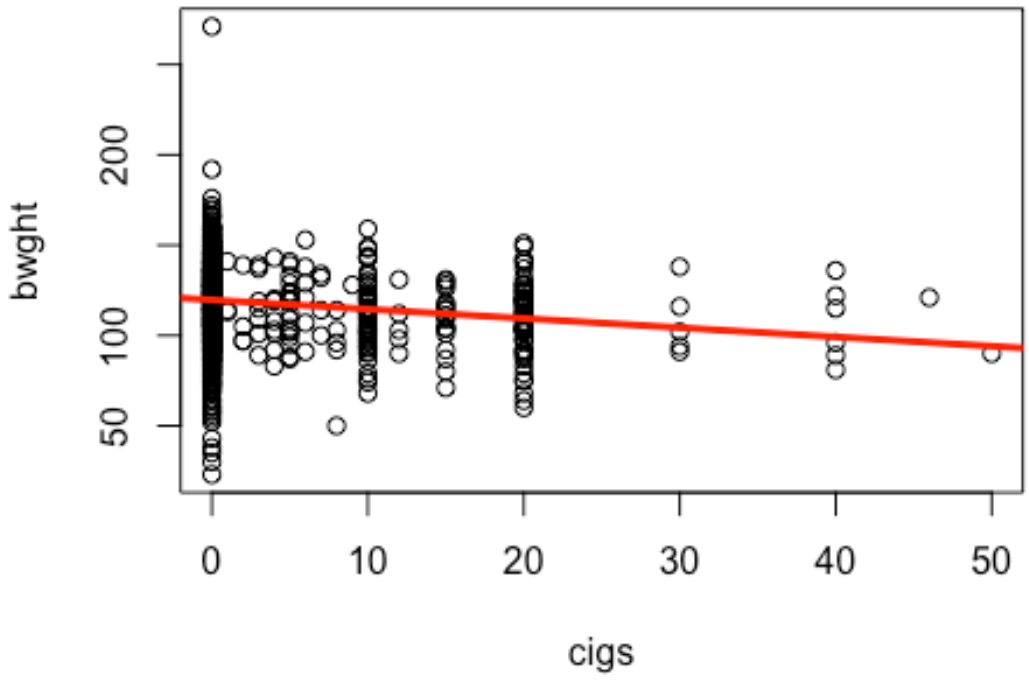



```
plot(as.factor(cigs),bwght)
```



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

```
fit = lm(bwght ~ cigs)
plot(cigs, bwght)
abline(fit,col='red',lwd=3)
```



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

```
coef(fit)

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

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

The var(u) = 0 and sum(X*u)=0

```
bwght_hat = bwght
bwght_hat = -0.5137721*cigs + 119.7719
print(sum(bwght_hat-bwght))

## [1] -0.0005737

print(sum((bwght_hat-bwght)*cigs))

## [1] -0.0015525
```

5. Does this simple regression capture a causal relationship between smoking and birthweight? Explain why or why not. No only 2.2% of variability in birthweight is captured by the variability in

```
summary(fit)$r.squared
```

```
## [1] 0.02272912
```

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? The cigs variables skip values between 20 and 50. The number of cigs increase by 10 for every step. This might be an error in measurement. The birthweight and cigs might have a strong linear relationship without the error.

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

```
y_0 = 119.772
y_1 = -0.514*20 + 119.772
y_0
```

```
## [1] 119.772
```

```
y_1
```

```
## [1] 109.492
```

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

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

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

```
(100-119.772)/(-0.514)
```

```
## [1] 38.46693
```