Table of Contents

Unit 10 Homework

Student Name: Youzhi (Chloe) Wu

Section Number: 05 - Wednesday, 6:30 PM

Cigarettes and Birthweight

Context: 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?

Data: 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.

In [22]:
```
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summ
ary Statistics Tables.
 R package version 5.2.2. https://CRAN.R-project.org/package=stargaz
er
```

In [1]:
```
load("bwght.RData")
```

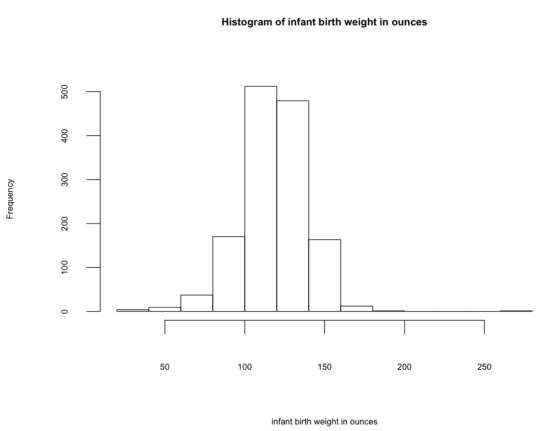
In [2]:
```
summary(data)
```

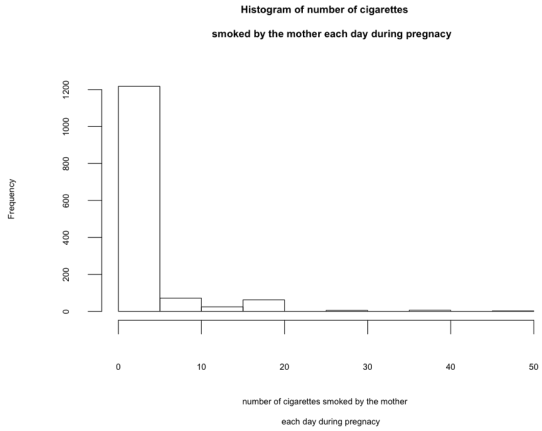
```
     faminc          cigtax         cigprice          bwght      
 Min.   : 0.50   Min.   : 2.00   Min.   :103.8   Min.   : 23.0  
 1st Qu.:14.50   1st Qu.:15.00   1st Qu.:122.8   1st Qu.:107.0  
 Median :27.50   Median :20.00   Median :130.8   Median :120.0  
 Mean   :29.03   Mean   :19.55   Mean   :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  
```

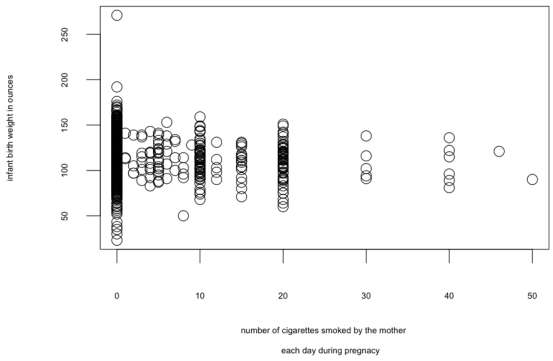
1. Examine the dependent variable, infant birth weight in ounces (bwght) and the independent variable, the number of cigarettes smoked by the mother each day during pregnacy (cigs).

In [3]:
```
# Adjust figure size
options(repr.plot.height = 16, repr.plot.width = 20, repr.plot.pointsi
ze = 32)

par(mfrow = c(2,2))

# Draw histogram to see overall distribution of bwght and cigs in the
dataset
hist(data$bwght, main = "Histogram of infant birth weight in ounces",
xlab = "infant birth weight in ounces")
hist(data$cigs, main = "Histogram of number of cigarettes \n
    smoked by the mother each day during pregnacy",
     xlab = "number of cigarettes smoked by the mother \n
    each day during pregnacy")

# Draw scatterplot to see how bwght changes over cigs
plot(data$cigs, data$bwght,
     xlab = "number of cigarettes smoked by the mother \n
     each day during pregnacy",
     ylab = "infant birth weight in ounces")

# Calculate the correlation bewteen cigs and bwght, we can see there i
s a negative correlation
cor(data$cigs, data$bwght)
```

-0.150761802543127

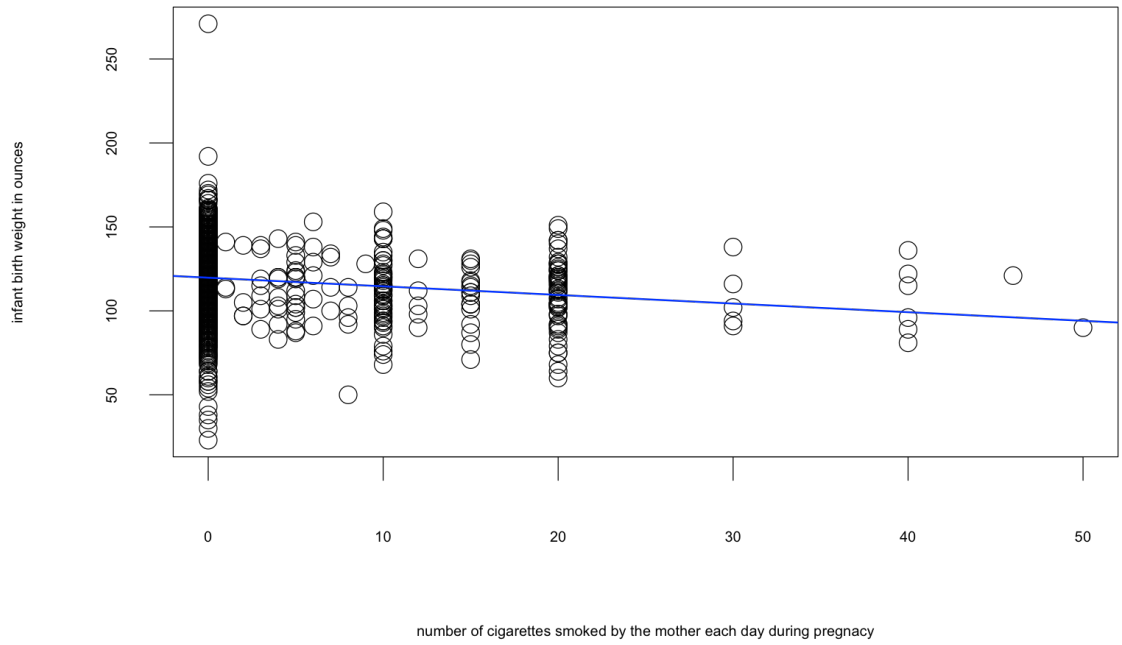






2. Fit a linear model that predicts bwght as a function of cigs. Superimpose your regression line on a scatterplot of your variables.

In [6]:
```
# Adjust figure size
options(repr.plot.height = 10, repr.plot.width = 14, repr.plot.pointsi
ze = 32)

# Fit a linear model to predict bwght as a function of cigs
m1 <- lm(bwght ~ cigs, data = data)

# Draw a regression line on the scatterplot of bwght and cigs
plot(data$cigs, data$bwght,
     xlab = "number of cigarettes smoked by the mother each day during
pregnacy",
     ylab = "infant birth weight in ounces")
abline(m1, 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?

In [7]:
```
# Get the coefficients of fitted model, m1
m1$coefficients
```

(Intercept)	119.77190039835
cigs	-0.513772092823396

- The slope coefficient can be interpreted as the expected change in bwght would be -0.514 ounces given a unit change in cigs, and holding all other factors and error term constant.
- The expected change is not practically significant considering that the median weight of newborn is 7.5 pounds and the mean weight is 7.4 pounds.

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

The two moment conditions are:

$$\hat{\beta_1} = \frac{cov(x_i, y_i)}{var(x_i)}$$
$$\beta_0 = \bar{y} - \hat{\beta_1} \cdot \bar{x}$$

In [8]:
```
# Use R to calculate beta_1 and beta_0 based on the two moment conditi
ons
(beta_1 <- cov(data$cigs, data$bwght)/var(data$cigs))
(beta_0 <- mean(data$bwght) - beta_1*mean(data$cigs))

# The results are consistent with the fitted model, m1.
```

-0.513772092823396

119.77190039835

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

No, it does not. The model only explains the expected change in bwght given a unit change in cigs under the circumstances that all other factors are hold constant. There are many other omitted variables that may contribute to low birthweight, such as mother's health conditions before and during, or her diet during pregnancy, etc. As such, we cannot conclude causality between the two variables based on this model.

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?

From the scatterplot, we can see that majority of cigs value at 0, 10, 20, 30, 40 and 50. It is suspected that when providing the number of cigs, interviewees may tend to round up or round down into tens. As such it may present certain inaccuracy in measuring cigs. If interviewees were able to provide accurate counts of cigarettes consumed each day during pregnancy, the distribution of cigs may have been more dispersed, instead of concentrating in the tens. In this case, the scatterplot between cigs and bwght may be more dispersed which may shift the regression line.

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

In [9]:
```
(bwght_0 = -0.513772092823396*0 + 119.77190039835)
(bwght_20 = -0.513772092823396*20 + 119.77190039835)
```

119.77190039835

109.496458541882

- The predicted birthweight when cigs is 0 would be 119.772 ounces;
- The predicted birthweight when cigs is 20 would be 109.496 ounces.

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

In [10]:
```
predict(m1 , data.frame(cigs = c(0, 20)) )
```

1	119.77190039835
2	109.496458541882

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

In [11]:
```
(cigs_100 = (100 - 119.77190039835)/-0.513772092823396)
```

38.4837959759453

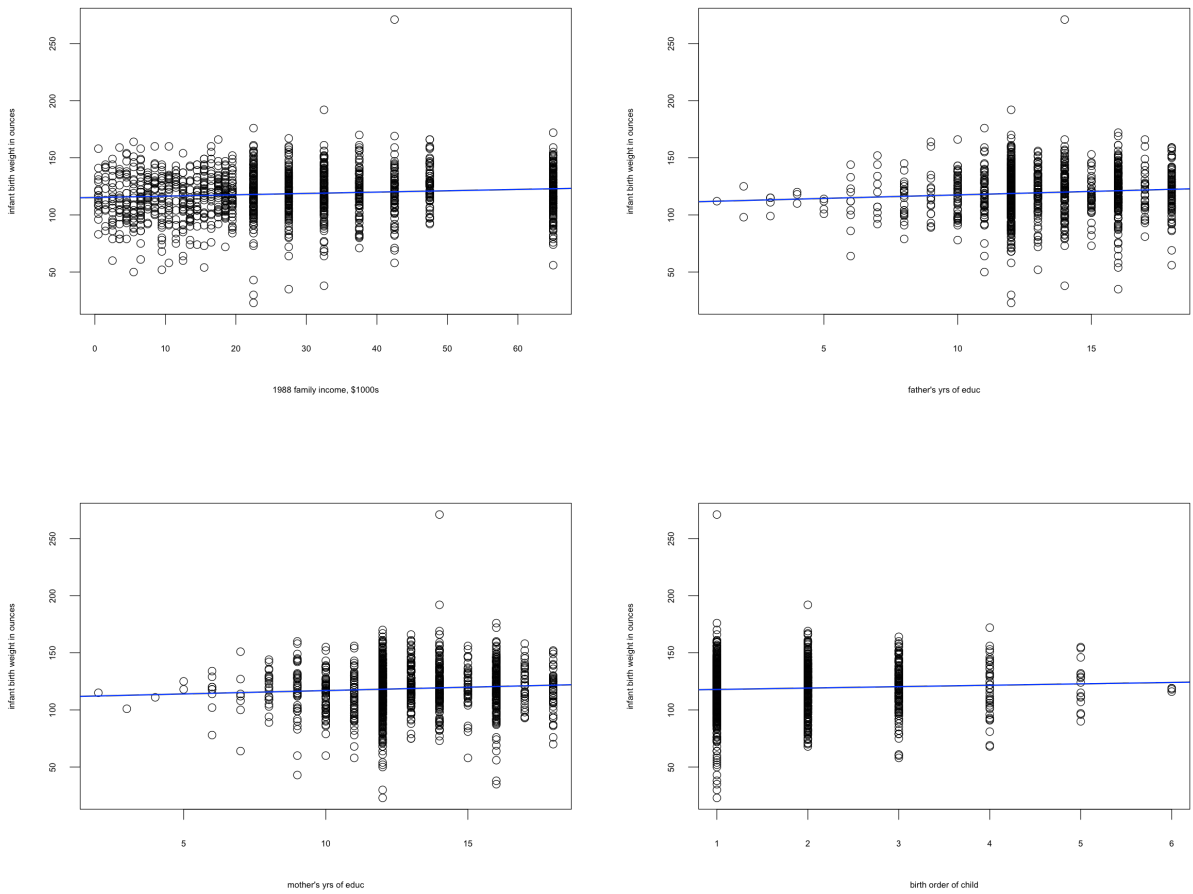
> To predict a birthweight of 100 ounces, cigs would have to be 38.4837959759453.

10. Based on all available variables, select a model that best explains the birthweight. Interpret your finding

In [21]:
```
# First examine all available variables
# Select several variables that may affect birthweight and at the same time meet the CLM assumptions
# Run correlation test to check the correlation between potential predictors and outcome variable
cor.test(data$faminc, data$bwght)
cor.test(data$fatheduc, data$bwght)
cor.test(data$motheduc, data$bwght)
cor.test(data$parity, data$bwght)
```

```
        Pearson's product-moment correlation

data:  data$faminc and data$bwght
t = 4.0799, df = 1386, p-value = 4.762e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05664503 0.16063261
sample estimates:
      cor
0.1089368

        Pearson's product-moment correlation

data:  data$fatheduc and data$bwght
t = 2.909, df = 1190, p-value = 0.003693
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02738105 0.14014032
sample estimates:
       cor
0.08402968

        Pearson's product-moment correlation

data:  data$motheduc and data$bwght
t = 2.5788, df = 1385, p-value = 0.01002
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01655179 0.12132107
sample estimates:
       cor
0.06912704

        Pearson's product-moment correlation

data:  data$parity and data$bwght
t = 2.0324, df = 1386, p-value = 0.04231
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.001898496 0.106819678
sample estimates:
       cor
0.05450955
```

In [27]:
```
# Adjust figure size
options(repr.plot.height = 16, repr.plot.width = 20, repr.plot.pointsi
ze = 24)

# Plot scatterplots between potential predictors and outcome variable
# Fit a reguression line for each plot to check their linearity
par(mfrow = c(2,2),font = 2)

plot(data$faminc, data$bwght,
     xlab = "1988 family income, $1000s",
     ylab = "infant birth weight in ounces")
mfb <- lm(bwght ~ faminc, data = data)
abline(mfb, col="blue", lwd=2)

plot(data$fatheduc, data$bwght,
     xlab = "father's yrs of educ",
     ylab = "infant birth weight in ounces")
mfeb <- lm(bwght ~ fatheduc, data = data)
abline(mfeb, col="blue", lwd=2)

plot(data$motheduc, data$bwght,
     xlab = "mother's yrs of educ",
     ylab = "infant birth weight in ounces")
mmeb <- lm(bwght ~ motheduc, data = data)
abline(mmeb, col="blue", lwd=2)

plot(data$parity, data$bwght,
     xlab = "birth order of child",
     ylab = "infant birth weight in ounces")
mpb <- lm(bwght ~ parity, data = data)
abline(mpb, col="blue", lwd=2)
```



In [23]:
```
# Build a regression model
model2 <- lm(bwght ~ faminc + fatheduc + motheduc + parity + cigs, dat
a = data)

# Compare model specifications between m1 and model2
stargazer(m1, model2,
          type="text", keep.stat=c("n", "adj.rsq"))
```

```
==========================================
                Dependent variable:
              ---------------------------
                        bwght
                  (1)            (2)
------------------------------------------
faminc                            0.056
                                 (0.037)

fatheduc                         0.472*
                                 (0.283)

motheduc                         -0.370
                                 (0.320)

parity                          1.788***
                                 (0.659)

cigs            -0.514***       -0.596***
                 (0.090)         (0.110)

Constant        119.772***     114.524***
                 (0.572)         (3.728)

------------------------------------------
Observations      1,388           1,191
Adjusted R2       0.022           0.035
==========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

In [24]:
```
# Compare the AIC for m1 and model2
AIC(m1)
AIC(model2)
```

12276.9142763355

10498.4408603578

1. Model Building

- I first examined all available variables in the dataset, and run scatterplot matrix for all. There are some variables that are in perfect linearity with the dependent variable, bwght, such as lbwght and bwghtlbs. These variables will not be included in the model.
- There is one variable, pack, that is in linear relation with cigs. This variable will not be included in the model.
- lfaminc and faminc are essentially evaluating the same factor on bwght. By comparing the scatterplots between lfaminc and bwght, with faminc and bwght, I did not see a strong argument to prefer one variable over the other. In this case, I used faminc in the new model.
- male (=1 if male child) and white (=1 if white child) variables are considered as outcomes of birth, and will not likely to determine the infant birth weight. Therefore, they are not included in the new model.
- cigtax and cigprice would affect the cigs and pack variables, but may impose less direct affect on bwght. As such, they are not included in the new model.

2. Model Interpretation

- The new model (model2) includes faminc, fatheduc, motheduc, parity and cigs variables as the predictors for bwght.
- Under model 2, it is expected that infant birth weight will
  - increase by 0.056 ounces with unit increase in family income (in thousand), holidng all other factors constant.
  - increase by 0.472 ounces with unit increase in years of father education, holidng all other factors constant.
  - decrease by 0.370 ounces with unit increase in years of mother education, holidng all other factors constant.
  - increase by 1.788 ounces with unit increase in the birth order of the infant (parity), holidng all other factors constant.
  - decrease by 0.596 ounces with unit increase in the number of cigarets smoked during pregnancy, holidng all other factors constant.
- The adjusted R square (0.035) of model 2 is higher than model 1 (0.022). In addition, the AIC for model 2 (10498.441) is lower than the AIC for model 1 (12276.914). Both indicate that model 2 is a better fit to explain birthweight than model 1.

11. Using this model, did your estimate of *cigs* change? What does it mean for your interpretation of *cigs* if it changed/did not change in your preferred model?

Yes, the estimate of cigs changed in Model 2.

- The original coefficient of cigs in Model 1 is -0.514; whereas, the new coefficient of cigs in Model 2 is -0.596.
- This means that the expected decrease in bwght with a unit increase in cigs is greater in Model 2 than in Model 1.
- With less omitted variables in Model 2, the effect of cigs appears stronger in bwght. It indicates that the omitted variables have negative bias in Model 1, driving the expected effect of cigs downwards.

In [ ]: