Impact of Smoking during Pregnancy on Infant Birth Weight

Ziyue (Bella) Chen

11/06/2020

# How can mother’s smoking behavior affect infant birth weight?

Smoking is a harmful behavior that can cause birth defects, so it is expected that as the number of cigarettes smoked per day for a mother during pregnancy increases, the weight of her baby at birth will decrease. In this project, the two variables will have their distributions and relationship analyzed, and a LS regression model will be generated and tested to demonstrate whether this relationship can be fitted into a linear model.

# 1. Data Preparation

# Data interested to know: infant birth weight vs. mother's smoking behavior  
data <- read.csv("Bwghtgrams.csv", header = T)  
head(data, n = 2)

## faminc cigtax cigprice fatheduc motheduc parity male white cigs lbwght  
## 1 13.5 16.5 122.3 12 12 1 1 1 0 4.691348  
## 2 7.5 16.5 122.3 6 12 2 1 0 0 4.890349  
## bwghtlbs packs lfaminc bwghtgrams  
## 1 6.8125 0 2.602690 3090.098  
## 2 8.3125 0 2.014903 3770.487

# Dimension of the original data  
dim(data)

## [1] 1388 14

# Structure of the original data  
str(data)

## 'data.frame': 1388 obs. of 14 variables:  
## $ faminc : num 13.5 7.5 0.5 15.5 27.5 7.5 65 27.5 27.5 37.5 ...  
## $ cigtax : num 16.5 16.5 16.5 16.5 16.5 16.5 16.5 16.5 16.5 16.5 ...  
## $ cigprice : num 122 122 122 122 122 ...  
## $ fatheduc : int 12 6 NA 12 14 12 16 12 12 16 ...  
## $ motheduc : int 12 12 12 12 12 14 14 14 17 18 ...  
## $ parity : int 1 2 2 2 2 6 2 2 2 2 ...  
## $ male : int 1 1 0 1 1 1 0 0 0 0 ...  
## $ white : int 1 0 0 0 1 0 1 0 1 1 ...  
## $ cigs : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ lbwght : num 4.69 4.89 4.86 4.84 4.9 ...  
## $ bwghtlbs : num 6.81 8.31 8.06 7.88 8.38 ...  
## $ packs : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ lfaminc : num 2.603 2.015 -0.693 2.741 3.314 ...  
## $ bwghtgrams: num 3090 3770 3657 3572 3799 ...

# Variables interested to know:  
# Independent variable: cigs - number of cigarettes per day that an infant’s mother smoked during pregnancy  
cigs <- data$cigs  
# Dependent variable: bwghtgrams - infant birth weights in grams  
bwghtgrams <- data$bwghtgrams

# 2a. Initial Data Analysis - “cigs”

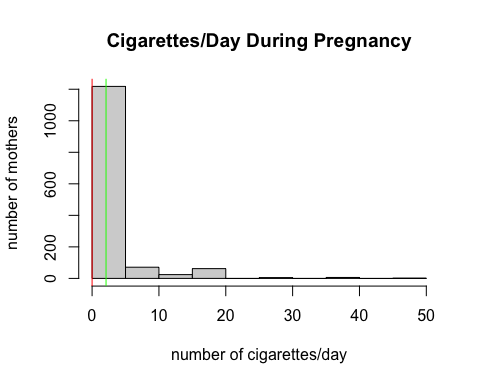
# Structure  
str(cigs)

## int [1:1388] 0 0 0 0 0 0 0 0 0 0 ...

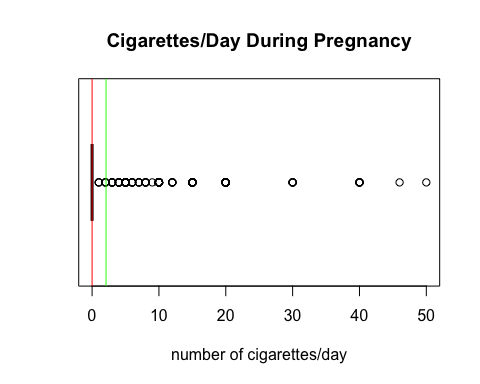
# Statistics  
summary(cigs)

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

# Graphical Summary - Histogram  
hist(cigs, main = "Cigarettes/Day During Pregnancy", xlab = "number of cigarettes/day", ylab = "number of mothers")  
abline(v = mean(cigs), col = "green")  
abline(v = median(cigs), col = "red")



# Graphical Summary - Boxplot  
boxplot(cigs, main = "Cigarettes/Day During Pregnancy", xlab = "number of cigarettes/day", horizontal = T)  
abline(v = mean(cigs), col = "green")  
abline(v = median(cigs), col = "red")



# The mean number of cigarettes/day during pregnancy is 2.087, the median is 0.000, the minimum is 0.000, and the maximum is 50.000.  
# The majority of the mothers don't smoke during pregnancy, but there are also ones who do smoke.  
# The mean is greater than the median because those who smoke are pulling up the mean.  
# The distribution of number of cigarettes/day in this sample is right skewed.

# 2b. Initial Data Analysis - “bwghtgrams”

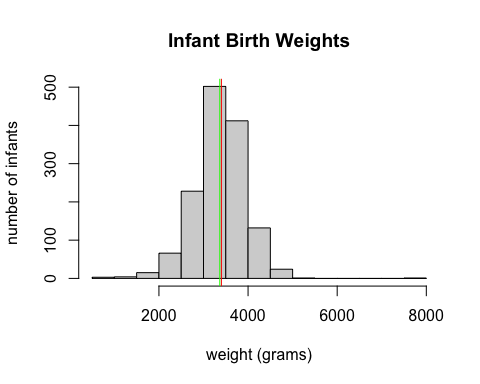
# Structure  
str(bwghtgrams)

## num [1:1388] 3090 3770 3657 3572 3799 ...

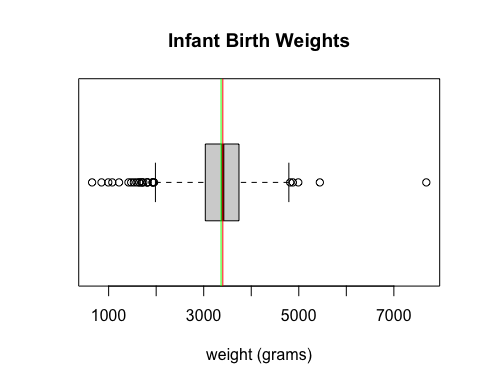
# Statistics  
summary(bwghtgrams)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 652 3033 3402 3365 3742 7683

# Graphical Summary - Histogram  
hist(bwghtgrams, main = "Infant Birth Weights", xlab = "weight (grams)", ylab = "number of infants")  
abline(v = mean(bwghtgrams), col = "green")  
abline(v = median(bwghtgrams), col = "red")



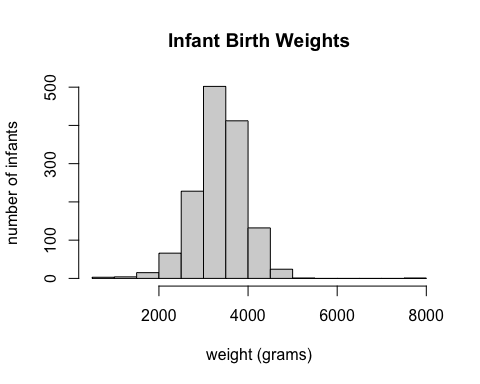
# Graphical Summary - Boxplot  
boxplot(bwghtgrams, main = "Infant Birth Weights", xlab = "weight (grams)", horizontal = T)  
abline(v = mean(bwghtgrams), col = "green")  
abline(v = median(bwghtgrams), col = "red")



# The mean infant birth weight is 3365g, the median is 3402g, the minimum is 652g, and the maximum is 7683g.  
# The mean and the median are similar.  
# The distribution of infant birth weight in this sample is approximately Normal.

# 3. Distribution Analysis

# Distribution of infant birth weights  
hist(bwghtgrams, main = "Infant Birth Weights", xlab = "weight (grams)", ylab = "number of infants")



# Based on the histogram, this distribution is approximately Normal.

# By the Central Limit Theorem, the distribution of mean infant birth weight can also be approximately Normal as the size increases.  
# This sample mean will be replicated and have its distribution analyzed in the following steps.

# Random sample generated based on the bwghtgrams data  
sample(bwghtgrams, 100, replace = TRUE)

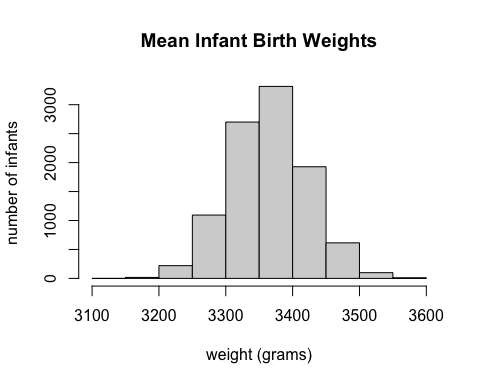
## [1] 3940.584 3997.283 3827.186 4082.331 3061.749 3260.195 2863.302 2891.651  
## [9] 2126.214 4252.428 4337.477 3316.894 3118.448 3770.487 3883.885 4110.681  
## [17] 2664.855 3543.690 3118.448 3628.739 3288.545 3345.244 4422.525 3061.749  
## [25] 3543.690 3260.195 3005.050 3912.234 2749.904 3486.991 3401.943 2778.253  
## [33] 3855.535 3203.496 4110.681 4224.079 3430.292 2126.214 2579.807 3628.739  
## [41] 3005.050 3401.943 3742.137 3855.535 2834.952 2324.661 3401.943 2353.010  
## [49] 3005.050 3798.836 3657.088 2693.205 3118.448 3685.438 3997.283 3175.146  
## [57] 3798.836 3515.341 3515.341 3572.040 3515.341 3855.535 2863.302 3175.146  
## [65] 3316.894 3628.739 3515.341 4053.982 3458.642 2523.108 3827.186 2126.214  
## [73] 3657.088 3090.098 2636.506 3628.739 2636.506 3798.836 3061.749 3288.545  
## [81] 3458.642 3600.389 3997.283 3543.690 3005.050 3628.739 3572.040 3146.797  
## [89] 3118.448 3288.545 3742.137 3231.846 3940.584 3401.943 3486.991 3316.894  
## [97] 2948.350 4167.380 3486.991 3770.487

# Mean of this random sample  
mean(sample(bwghtgrams, 100, replace = TRUE))

## [1] 3263.03

# Replicate this mean 10000 times  
set.seed(1)  
mean\_weight <- replicate(10000, mean(sample(bwghtgrams, 100, replace = TRUE)))

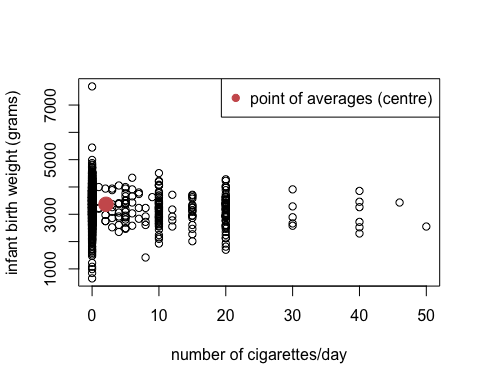
# Distribution of replicated sample mean  
hist(mean\_weight, main = "Mean Infant Birth Weights", xlab = "weight (grams)", ylab = "number of infants")



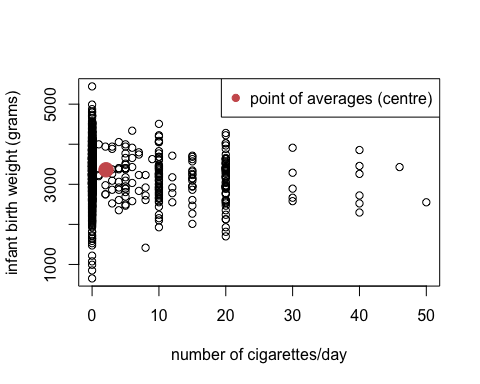
# The distribution of the mean also shows a Normal distribution because of the Central Limit Theorem.

# 4. Regression Model Estimation

# Scatter plot of bwghtgrams vs. cigs  
plot(cigs, bwghtgrams, xlab = "number of cigarettes/day", ylab = "infant birth weight (grams)")  
points(mean(cigs), mean(bwghtgrams), col = "indianred", pch = 19, cex = 2)  
legend("topright", c("point of averages (centre)"), col = "indianred", pch = 19)



# Data cleaning: there seems to be an outlier with an infant birth weight greater than 7000 grams  
data2 <- data[data$bwghtgrams<7000,]  
cigs <- data2$cigs  
bwghtgrams <- data2$bwghtgrams  
plot(cigs, bwghtgrams, xlab = "number of cigarettes/day", ylab = "infant birth weight (grams)")  
points(mean(cigs), mean(bwghtgrams), col = "indianred", pch = 19, cex = 2)  
legend("topright", c("point of averages (centre)"), col = "indianred", pch = 19)



# Correlation coefficient  
cor(cigs, bwghtgrams)

## [1] -0.1519832

# The correlation coefficient is -0.15, so there is a negative correlation between infant birth weight and number of cigarettes during pregnancy.  
# However, this value is close to 0, which means the the correlation is not strong.

# Linear model  
lm(bwghtgrams~cigs)

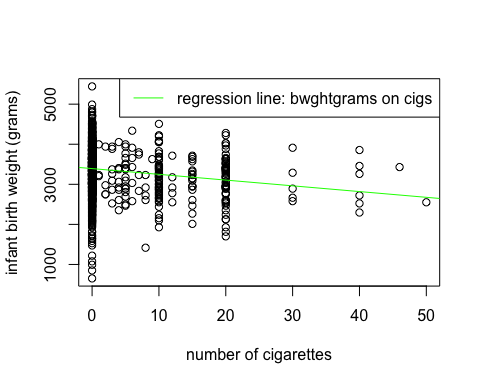
##   
## Call:  
## lm(formula = bwghtgrams ~ cigs)  
##   
## Coefficients:  
## (Intercept) cigs   
## 3392.01 -14.38

# Linear model regression line  
model <- lm(bwghtgrams~cigs)  
model$coeff

## (Intercept) cigs   
## 3392.0072 -14.3842

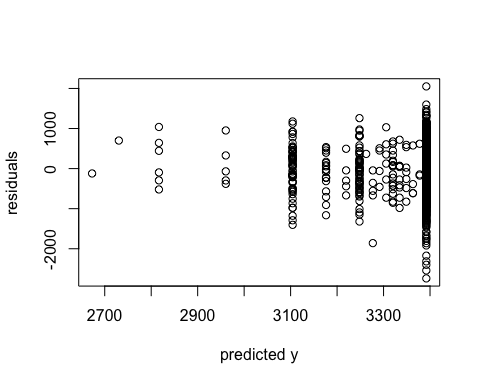
# The estimated linear model should have a regression line of bwghtgrams = -14.3842 \* cigs + 3392.0072.

# Plot with regression line  
plot(cigs, bwghtgrams, xlab = "number of cigarettes", ylab = "infant birth weight (grams)")  
abline(lm(bwghtgrams~cigs), col = "green")  
legend("topright", c("regression line: bwghtgrams on cigs"), col = "green", lty = 1)

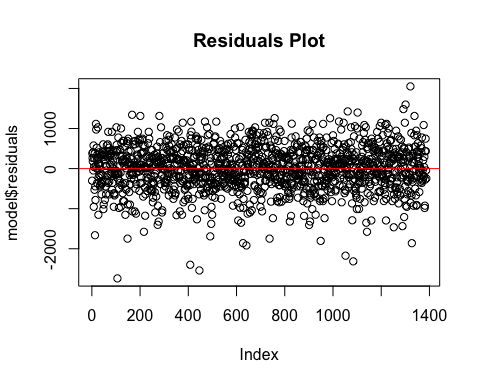


# 5. Regression Model Testing

# Residuals vs. predicted y  
preds <- predict(model)  
resids <- residuals(model)  
plot(preds, resids, xlab = "predicted y", ylab = "residuals")



# Residual plot  
plot(model$residuals, main = "Residuals Plot")  
abline(h = 0, col = "red")



# Classic linear regression model assumptions:  
# 1) Linear in parameters: bwghtgrams is linearly related to cigs by bwghtgrams = -14.3842 \* cigs + 3392.0072.  
# 2) Random sampling: It is assumed that the mothers and the infants are randomly chosen from the population.  
# 3) Sample variation: The independent variable, which is the number of cigarettes smoked during pregnancy, is not a constant value.  
# 4) Zero condition mean: There seems to be a relationship between residuals and predicted y such that as the predicted y gets larger, the variance of the residuals also gets larger.  
# 5) Homoscedasticity: The residual plot shows no pattern, so the variance of the model is constant.

# The classical linear regression model assumptions are not all satisfied because as the predicted y gets larger, the variance of the residuals also gets larger.  
# Therefore, a linear regression model is not an appropriate model for the relationship between infant birth weight and number of cigarettes smoked during pregnancy.

# 6. Conclusions and Interpretations

# Solely based on the data provided, several conclusions can be made:  
# 1) Infant birth weight is negatively correlated with number of cigarettes/day during pregnancy, but the correlation is not strong. This means that smoking can make some contributions to a decreased birth weight.  
# 2) The linear model is not an appropriate model for the samples data.  
# 3) The infant birth weights is more Normally distributed.

# These conclusion are expected because of several reasons:  
# 1) In the data provided, the number of mothers who smoke is much smaller that of mothers who do not smoke, so there is not enough data to effectively measure the effect of smoking.  
# 2) Smoking is not the only factor that can affect infant birth weight, some other factors such as diet, genetic characteristics, emotions, etc. all have the potential to affect infant birth weight.

# Some adjustments can be made for future estimations:  
# 1) More samples can be drawn, especially samples from mothers who smoke during pregnancy.  
# 2) Try to minimize the effects of other factors, which means making them be controls instead of variables.

Even though this sample does not effectively demonstrate the harmful effect of smoking, mothers should still be aware that smoking is harmful and should avoid this behavior during pregnancy. Smoking is just bad :(