Final Project

Brooke O'Donnell

Professor Wang

SOCI 556: Data Analysis

This final project will use a dataset provided by the National Center for Health Statistics about Births in the United States in 2018. This is a large dataset containing information about "live births" in year 2018. The data collected is from information required to be filled out by new mothers for them to receive a "live birth" certificate for their newborns. After becoming familiar with the variables in this dataset I formulated the following hypotheses:

- Weight of the baby might affect the mothers weight gain during pregnancy.
- The number of prenatal visits might affect the mothers weight gain during pregnancy. The mothers that go to more prenatal visits may learn more information from the professional (OBGYN, Midwife) about habits to consider in order to maintain a healthy pregnancy weight.
- The age of the mother might affect the mothers weight gain during pregnancy. Older women might see decreased WTGAIN because they may have learned habits to maintain a healthy pregnancy weight from previous pregnancies.

This dataset contains 1,048,575 observations and 55 variables. The first step is to narrow down these variables by formulating research questions about our units of analysis for this dataset, mothers.

Research Questions:

- What independent variables influence the dependent variable weight gain (WTGAIN)?
- Do the independent variables LB.BABY, PREVIS, and MAGER affect the WTGAIN during the gestation of a mother's pregnancy?

Objective: Loading the dataset into R	**Interpretation:** N/A
Output: N/A	
Code: USBIRTHS.18<-read.csv("USBIRTHS.csv", header=TRUE, sep=",", fileEncoding="UTF-8-BOM")	

Objective: Prepare the variables for the multiple regression model:

- WTGAIN {Weight gain after pregnancy (in pounds)}
  - {DWgt_R – PWgt_R}

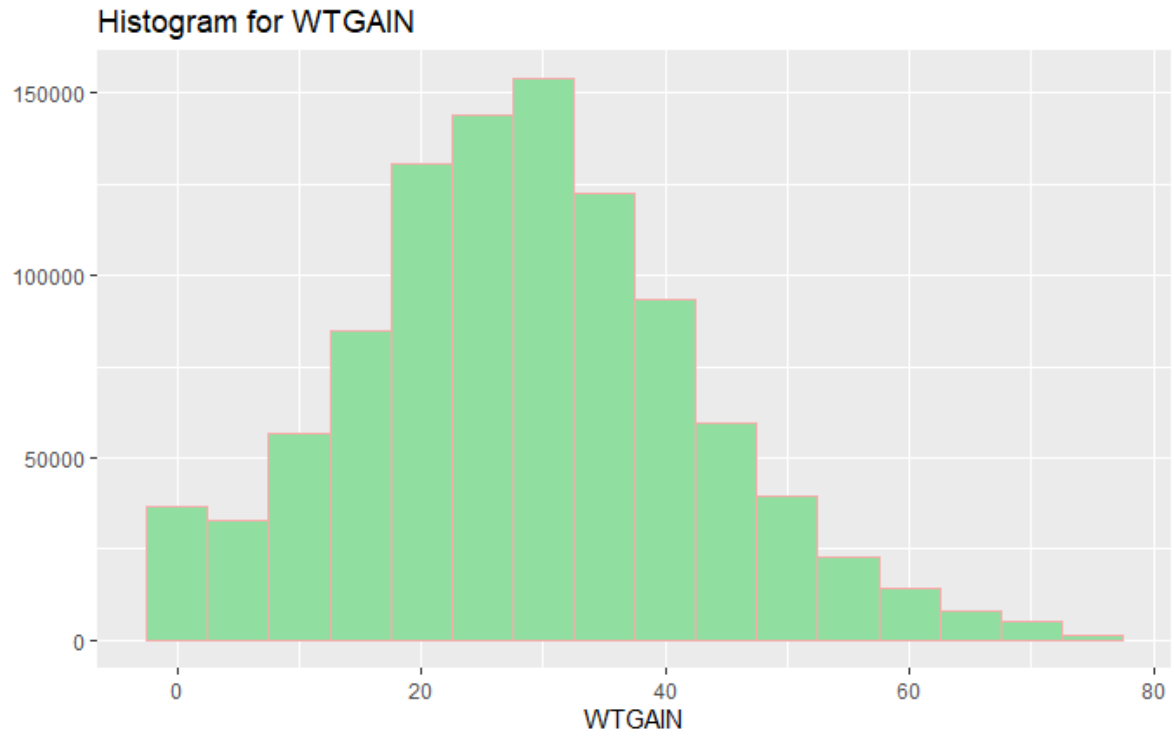
Variables:

- LB.BABY (Weight of baby in pounds)
- PREVIS(# of prenatal visits)
- MAGER (Mother's age)

Obtain a Histogram of WTGAIN

Interpretation: For this assignment the dependent variable will be WTGAIN {Weight gain after pregnancy (in pounds)}. The histogram suggests that my dependent variable is normally distributed.

Output:



Code:

```
str(USBIRTHS.18)
summary(USBIRTHS.18)
US.FIN<-select(USBIRTHS.18, "WTGAIN", "DBWT", "PREVIS", "MAGER", "MEDUC", "SEX")
US.FIN$LB.BABY<-(US.FIN$DBWT/454)
summary(US.FIN$LB.BABY)
dfdf<-select(US.FIN, "WTGAIN", "LB.BABY", "PREVIS", "MAGER", "MEDUC", "SEX")
```

```
na.omit(dfdf)
dfdf<-subset(US.FIN, US.FIN$WTGAIN<75)
summary(US.FIN$WTGAIN)
qplot(dfdf$WTGAIN,
geom="histogram",binwidth=5, main= "Histogram for WTGAIN",xlab="WTGAIN",
fill=I("#90DFA0"), col=I("#F7ACAC"))
```

Objective: Significant coefficients & R-squared	**Interpretation:** All of the coefficients are significant. The r-squared displays a low value of 0.03. This means that only 3 percent of total variance in WTGAIN can be explained by LB.BABY, PREVIS, and MAGER. The unit of analysis are mothers therefore the variance was expected to be low. If the baby increases by one pound the mother is expected to increase by 1.71 pounds. If the mother's number of prenatal visits increases by one visit her expected weight gain will increase by .04 pounds. When a mothers age increases by one year the mother weight gain is expected to decrease by .08 pounds.

Output:

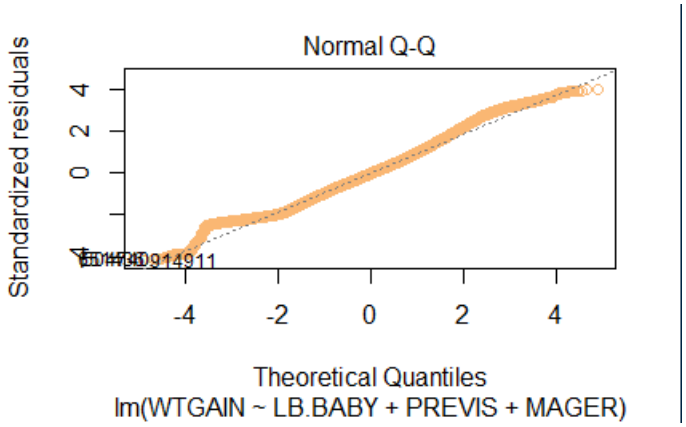
```
Call:
lm(formula = WTGAIN ~ LB.BABY + PREVIS + MAGER, data = dfdf)

Residuals:
    Min      1Q  Median      3Q     Max
-57.741  -8.996  -0.358   8.510  55.729

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.066012   0.102597  176.09   <2e-16 ***
LB.BABY      1.705765   0.010614  160.71   <2e-16 ***
PREVIS       0.035741   0.001109   32.22   <2e-16 ***
MAGER       -0.079506   0.002337  -34.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.79 on 1006315 degrees of freedom
Multiple R-squared:  0.02691,  Adjusted R-squared:  0.02691
F-statistic:  9278 on 3 and 1006315 DF,  p-value: < 2.2e-16
```

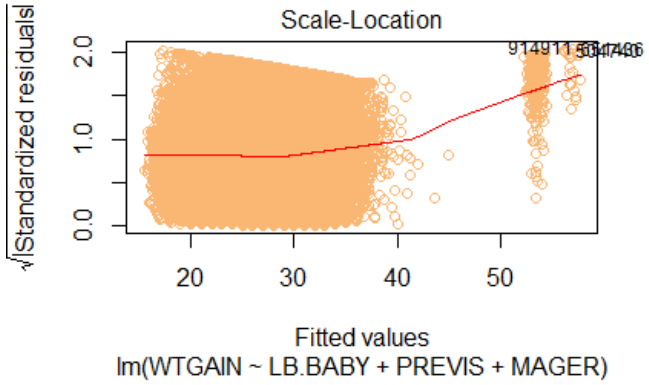
Code: mod.wt<-lm(WTGAIN~LB.BABY+PREVIS+MAGER+MEDUC+SEX, dfdf)
summary(mod.wt)

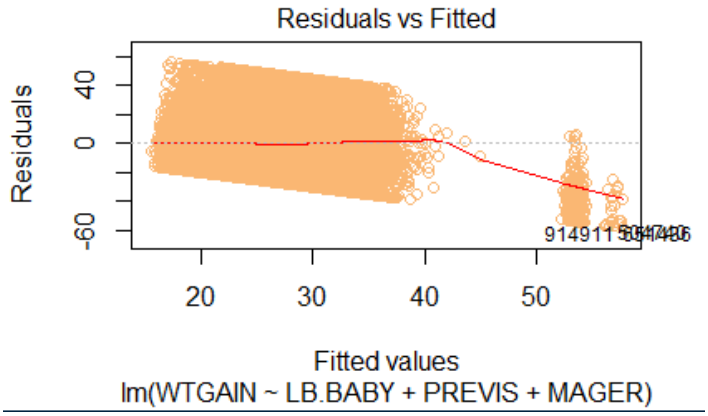
Objective: Examine normality and	**Interpretation:** To check the assumptions first

homoscedasticity	I look at the Normal Q-Q plot that appears to be acceptable. The residuals vs fitted plot is not ideal. The variance of the residuals is skewed to the left.

Output:







Code: plot(mod.wt, col="#FAB773")

Objective: Checking multicollinearity	**Interpretation:** The variance inflation factor is acceptable because the values are close to one.

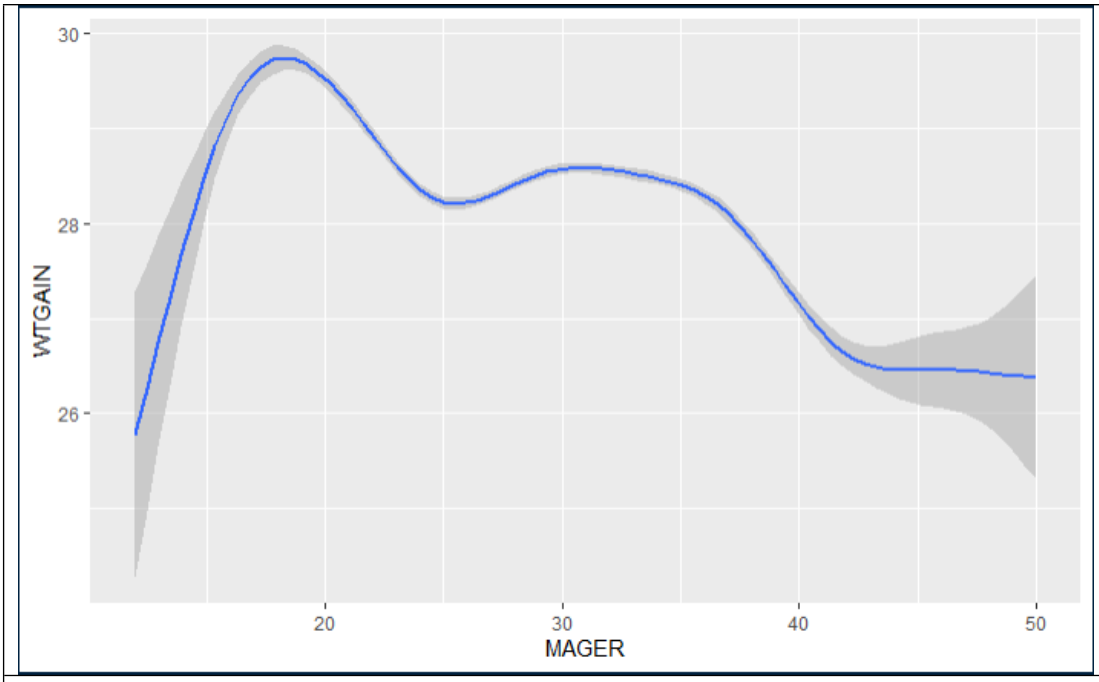
Output:
 	
`LB.BABY PREVIS MAGER`
`1.001204 1.000804 1.001510`	

Code: vif(mod.wt)	

Objective: Plot MAGER against WTGAIN Interpret this relationship	**Interpretation:** MAGER or Mother's age suggest that quadratic and cubic terms should be called because there are 2 bends. This explains that there is a decrease in WTGAIN shown in young mothers. Then there is a slight increase in WTGAIN mothers ages 25-30. Then there is another decrease in WTGAIN for women over age 32 (roughly).

Output:

```
Call:
lm(formula = WTGAIN ~ MAGER + SQ + CU, data = df.age)

Residuals:
    Min      1Q  Median      3Q     Max
-36.236 -11.333  -2.255   7.649  72.469

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.213e+01  1.658e-01 193.727  < 2e-16 ***
MAGER       -3.633e-02  5.501e-03  -6.605 3.99e-11 ***
SQ           4.129e-03  4.547e-04   9.083  < 2e-16 ***
CU          -6.303e-04  5.177e-05 -12.174  < 2e-16 ***
```



Code:

```
ggplot(dfdf, mapping= aes(x=MAGER, y=WTGAIN))+geom_smooth()df.age$SQ<-
(df.age$MAGER-mean(df.age$MAGER))^2
df.age$CU<-(df.age$MAGER-mean(df.age$MAGER))^3model.fin<-
lm(WTGAIN~MAGER+SQ+CU, data=df.age)
summary(model.fin)
```

Practicum Fall 2020 my research proposal topic is the knowledge of childbirth options shown among a sample of females. In my literature review I will explain how the medicalization of childbirth in the United States has led to an increase in cesarean sections. My first research question might be; Is the rate of cesarean sections significant in the United States? According to the sample of United States birth certificates in 2018 the amount of cesarean sections was significant. 322,534/1,006,175 This means about 32% of births were performed by cesarean section. What variables are significant for mothers who choose a cesarean section as the method of delivery? For my next section of this final I will provide an interpretation of what independent variables might affect my dependent variable method of delivery (RDMETH_REC).

```
  vaginal c-Section
   683641    322534
```

Hypotheses:

- Midwives and Certified Nurse Midwives will have a negative relationship with cesarean sections as the method of delivery. Why? Midwives and Certified Nurse Midwives promote and implement holistic alternatives for high-risk pregnancies before considering cesarean section as the method of delivery.

- Higher levels of mother's education will associate negatively with cesarean sections. Why? Mothers with high levels of education might be in a better financial position to hire a Midwife or a Certified Nurse Midwife.
- Minorities such Black mothers will have a positive relationship with cesarean sections as the method of delivery. Why? This might be because in American history black Americans were often taken advantage of by medical professionals. A brief example was shown in the Tuskegee Study where doctors/researchers injected Black men with syphilis to measure the impacts of the disease. That experiment did not end until 1972. Black Americans might not trust medical professionals as much as White Americans. The Black mothers might not know how to communicate with the medical professional to create a relationship with them as easily as a white mother. This introduces a barrier between the Black mother and the medical professional. They will not be able to establish pregnancy plans about the method of delivery the Black mother prefers and what alternative methods to turn to if the pregnancy evolves into a high-risk pregnancy.

Objective: Log(odds) Create a glm model for the variable method of delivery (RDMETH_REC). What variables have a significant effect on cesarean sections for mothers? Relevel: Make no c-section the reference category. Interpret the coefficients: What is interesting?	**Interpretation:** By superficially interpreting the coefficients there are a few relationships that stand out. If the sex of the baby is female the odds of having a c-section reduces. If the attendant of delivery is a Doctor of Osteopathy the odds of having a c-section reduces. However, this negative relationship is not as significant when comparing the odds of having a c-section when a Midwife or Certified Nurse Midwife is the attendant during delivery. Shown in the coefficients as education increases the odds of the mother having a c-section reduces. Finally, Black mothers have a positive relationship with having c-sections.

Output:

```
Call:
glm(formula = RDMETH_REC ~ SEX + MEDUC + ATTEND + MAGER + PREVIS +
    LB.BABY + WTGAIN, family = binomial, data = D.METH, na.action =
na.exclude)
Coefficients:
                                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)                     -1.8266517  0.0184857  -98.814  < 2e-16
***
SEXF                            -0.1128637  0.0044938  -25.115  < 2e-16
***

ATTENDCNM(Certified Nurse Midwife) -3.6741023  0.0258294 -142.245  < 2e-16
***
```

```
ATTENDOther Midwife                -5.7602108  0.2673266  -21.547  < 2e-16
***
ATTENDDO(Doctor of Osteopathy      -0.0625602  0.0088648   -7.057 1.70e-12
***

MEDUCHigh school or GED             0.0933808  0.0078945   11.829  < 2e-16
***
MEDUCsome college credit            0.0743841  0.0082779    8.986  < 2e-16
***
MEDUCassociate degree               0.0674118  0.0103103    6.538 6.22e-11
***
MEDUCbachlor's degree              -0.1027399  0.0084646  -12.138  < 2e-16
***
MEDUCbeyond bachlor's              -0.1889072  0.0095432  -19.795  < 2e-16
***

MRACE15Black(only)                  0.2508881  0.0065539   38.281  < 2e-16
***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 1150803
```

Code: D.METH<-select(dfdf, "WTGAIN", "LB.BABY", "PREVIS", "MAGER", "MEDUC", "SEX", "RDMETH_REC", "MRACE15", "ATTEND")
D.METH<-D.METH%>%mutate(SEX=recode_factor(SEX, "1"="M", "2"="F"))
D.METH<-D.METH%>%mutate(RDMETH_REC=recode_factor(RDMETH_REC, "1"="no c-section", "2"="no c-section","5"="no c-section", "3"="c-Section","4"="c-Section","6"="c-Section", "9"= "NA"))
D.METH<-D.METH%>%mutate(MEDUC=recode_factor(MEDUC, "1"="no hs diploma", "2"="no hs diploma","3"="High school or GED", "4"="some college credit","5"="associate degree","6"="bachlor's degree", "7"= "beyond bachlor's", "8"="beyond bachlor's", "9"="NA"))
D.METH<-D.METH%>%mutate(MRACE15=recode_factor(MRACE15, "01"="white(only)", "02"="Black(only)", "03"="AIAN(only)","04"="asian(only)", "05"="Chinese(only)","06"="Filipino(only)","07"="japanese(only)","08"="Korean(only)", "09"="Vietamese (only)", "10"="Other asian(only)", "11"="Hawaiian (only)","12"="Guamanian(only)", "13"="Samoan", "14"="Other Pacific Islander", "15"="More than one race"))
D.METH<-D.METH%>%mutate(ATTEND=recode_factor(ATTEND, "1"="MD(Doctor of Medicine)", "2"="DO(Doctor of Osteopathy)", "3"="CNM(Certified Nurse Midwife)","4"="Other Midwife", "5"="other","9"="unknown"))
na.omit(D.METH)
D.METH$RDMETH_REC<-relevel(D.METH$RDMETH_REC, ref="no c-section")
model1<-glm(RDMETH_REC~ SEX+MEDUC+ATTEND+MRACE15+MAGER+PREVIS+LB.BABY+WTGAIN, data=D.METH, family=binomial, na.action=na.exclude)
summary(model1)

Objective: Exponentially transform log(odds)	**Interpretation:** If the attendant during delivery is a Doctor of Osteopathy the odds of

into odds. Interpret the coefficients:	having a c-section reduces by 5%. This is not as drastic as the odds of having a c-section if the attendant is a Certified Nurse Midwife or a Midwife. These attendants reduce the odds of the mother having a c-section by 98-99%. Now I will compare levels of education with the likelihood of having a c-section. If the mother has a high school diploma or GED the odds of having a c-section is increased by 8%. If the mother achieved beyond a bachelor's degree, the odds of her having a c-section reduces by 15%. Finally, if the mother is black the odds of her having a c-section increases by 29%.
Output: ATTENDDO(Doctor of Ostheopathy)= -0.052517200 ATTENDCNM(Certified Nurse Midwife)= -0.974719573 ATTENDOther Midwife= -0.996872303 MEDUCHigh school or GED= 0.084341308 MEDUCsome college credit= 0.068898638 MEDUCassociate degree= 0.069564718 MEDUCbachlor's degree= -0.077280518 MEDUCbeyond bachlor's= -0.154123956 MRACE15Black(only)= 0.285166312	
Code: exp(model1$coefficients) exp(model1$coefficients)-1	

Objective: Transform odds into probability.	**Interpretation:** My practicum will attempt to provide evidence that females do not have the appropriate knowledge needed when choosing birthing options. Pregnant females are not informed about alternative attendants for delivery such as Certified Nurse Midwives or Midwives. Pregnant females who are of a minority group or have lower levels of education might not have access to these alternative attendants. These pregnant women often use a Doctor of Osteopathy as their

	attendant during delivery. When a Black mother has a Doctor of Osteopathy as their attendant, they will be 14% likely to have a C-section. As compared to a Black mother who has a Certified Nurse Midwife as their attendant, they will be .5% likely to have a C-section. In order to reduce c-sections United States should make alternative attendants such as Certified Nurse Midwives and Midwives more accessible to mothers of minority groups and decreased levels of education.

Output:

Probability of a Black mother with a Certified Nurse Midwife as the attendant during delivery having a c-section.

```
[1] 0.004583504
[1] 0.004583504
               Log Odds                     Odds Probability of C-section
           -5.380697500              0.004604609              0.004583504
```

Probability of a Black mother with a Doctor of Osteopathy as the attendant during delivery having a c-section.

```
[1] 0.1471766
[1] 0.1471766
               Log Odds                     Odds Probability of C-section
             -1.7569192                0.1725757                0.1471766
```

Code:
```
logodds.DO<- -1.9538537-3.6777248+0.250881
odds1<-exp(logodds.DO)
odds1
prob1<-exp(logodds.DO)/(1+exp(logodds.DO))
prob2<-odds1/(1+odds1)
prob1
prob2
predict1<-c(logodds.DO, odds1, prob1)
names(predict1)<- c("Log Odds", "Odds", "Probability of C-section")
predict1
logodds.MI<- -1.9538537-0.0539465+0.250881
odds2<-exp(logodds.MI)
odds2
prob3<-exp(logodds.MI)/(1+exp(logodds.MI))
prob4<-odds2/(1+odds2)
prob3
prob4
predict2<-c(logodds.MI, odds2, prob3)
```

```
names(predict2)<- c("Log Odds", "Odds", "Probability of C-section")
predict2
```

Objective: Psuedo R-squared	**Interpretation:** The McFadden R-squared is not ideal because it holds a value of .09. Researchers usually accept the model when the value is in between .2-.4. The Cox and Snell should not be accepted because it does not range from .7-.4. Since the Mcfadden does not function well the Nagelkerke should not be considered.

Output:

```
McFadden                          0.0907096
Cox and Snell (ML)                0.1075770
Nagelkerke (Cragg and Uhler)      0.1504900
```

Code:
```
modelnull<-glm(RDMETH_REC~1, data=D.METH, family=binomial,
na.action=na.exclude)
McFadden<-1-logLik(model1)/logLik(modelnull)
McFadden
```

Objective: Logistic Regression Model Prediction	**Interpretation:** 34937+655766/ 1006319 = 69% 27875+287597/1006319= 31% This model prediction is slightly better than the null model prediction. The null model will allow you to predict 68%. The Logistic Regression Model Prediction improves by 1%. The probability of a mother having "no c-section" is 69%.

Output:

```
 no c-section    c-Section
0.6793481987 0.3205087055
```

```
c.pred      no c-section c-section
  c-section    34937       27875
  no c-section 287597      655766
[1] 1006319
```

Code:
```
c.prob=predict(model1, D.METH, type="response")
c.pred=rep("no c-section", dim(D.METH)[1])
```

```
c.pred[c.prob>.5]= "c-section"
table(c.pred, D.METH$RDMETH_REC)
length(D.METH$RDMETH_REC)
```

In Conclusion, when working on my practicum the most important variables to consider that influence a female's knowledge of childbirth options might be level of education and race. I am now considering measuring the knowledge of childbirth options among different levels of education. Here I might find that a sample of females who have a high school diploma have much less knowledge on childbirth options than compared to a sample of female college students. This suggests that high school education in the United States is not providing females with the appropriate knowledge about the childbirth options available to women.