**Matthew Bunge**

**Final Paper**

**ECON 484**

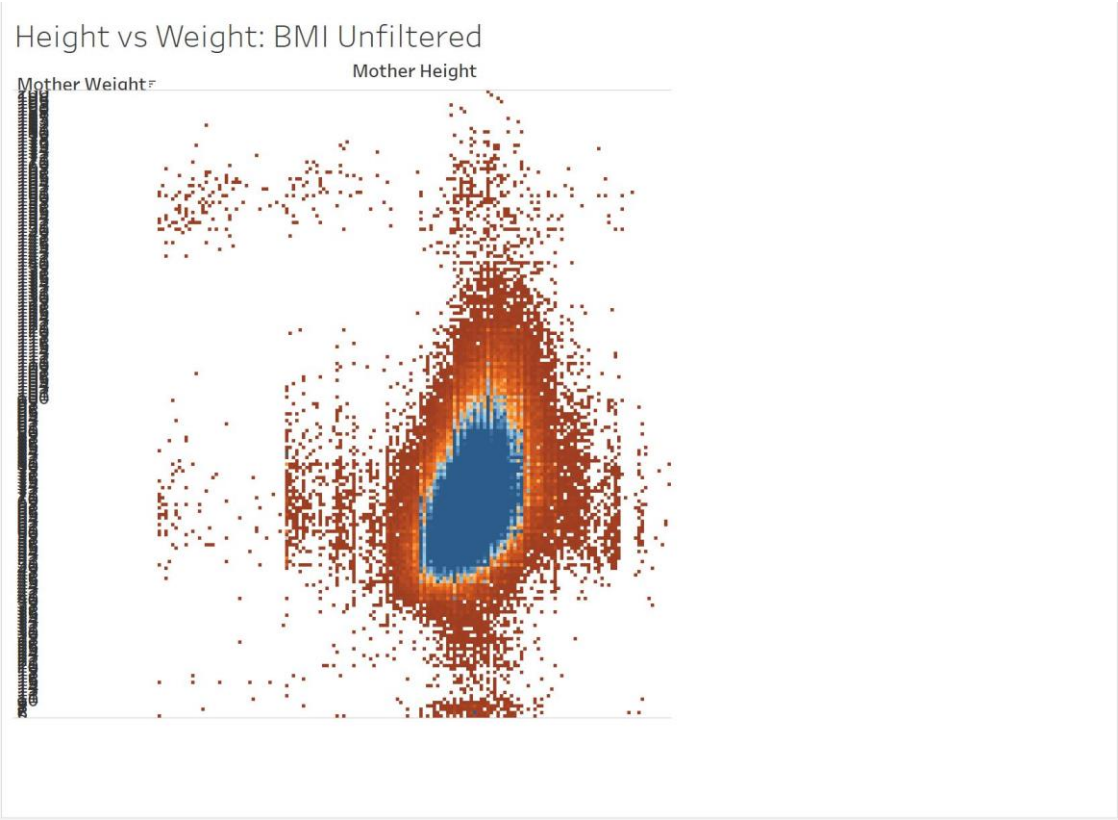
**Introduction:**

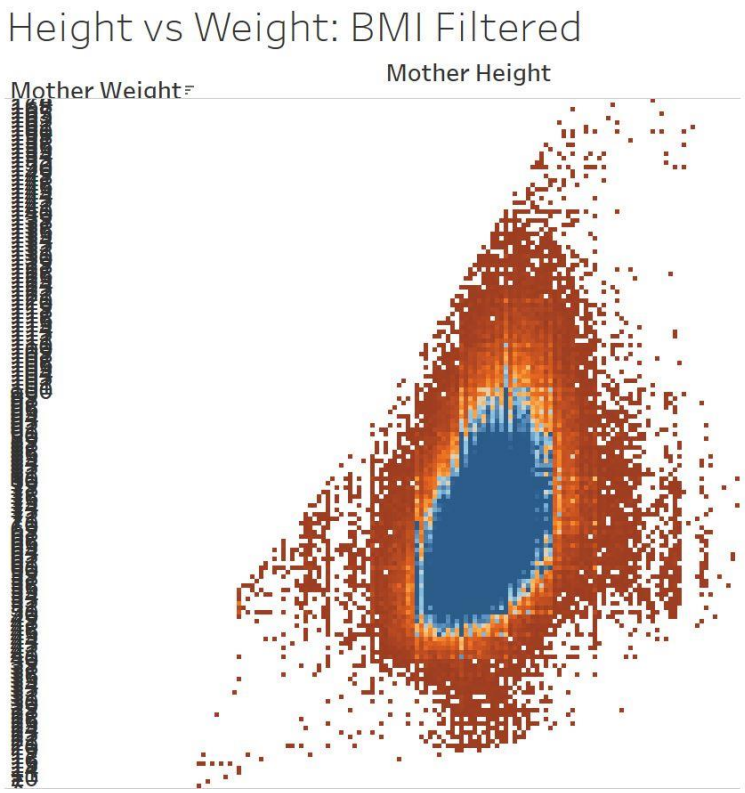
A cesarean section (hereafter referred to as a "C-section") is a medical procedure done during childbirth, where when delivering a baby normally is too difficult, a surgery is performed to remove the baby from the womb. A common case where this would occur would be if the woman is having twins, it is much safer to just remove them than try to deliver them. The goal of this paper is to try and predict whether women attending Mexican hospitals are likely to have C-sections or not, as well as find what factors lead to C-sections being performed rather than common vaginal births.

**The Data and The Cleaning Process:**

The first main file consists of information about women attending an obstetrician appointment. Since the focus of the research is on a method of giving birth, women who either attended with the purpose of getting an abortion or for a general visit are not included. Additionally, birth operations that resulted in multiple births are removed, as C-sections occurred at an 88.877% rate for multiple births versus 35.905% in general. Because multiple births are generally a special condition in comparison to a single birth, the scope of the question was narrowed to single births for the sake of focusing on the less initial clear factors that might cause a need for C-sections. The second main file was a record of the babies born to the obstetric patients from the first file, matched by ID. Because there are some babies who do not have an ID to match to a parent, those records are removed. The third file is the more general hospital records of the mothers, recording information such as age, height, weight, insurance status, etc. All three of these files are merged on the ID matching mother to child to hospital visit. The next step of the cleaning involves removing missing values, generally represented in the data by a value of repeating 9s. All records with missing values were generally removed for the sake of creating a compact data set that would not need to undergo any further cleaning if there was a decision in the future to possible drop some of the variables, or add some extra ones in. The type of birth in the data is normally represented as "normal vaginal", "difficult vaginal" or C-section, and since the purpose is testing for C-sections, the two vaginal births were binned into one value,

and C-sections were left as their own value. In this way, any test results in the outcome of either “C-section” or “not C-section”. There is a modification where the original file included the current birth in number of births, so that variable is reduced by one across the board to represent “previous births”. Finally, a BMI variable was generated by taking mother weight, and dividing it by mother height squared. It is then filtered so that all records with BMIs not in the range of [8, 55] are removed. In general, a normal BMI is probably somewhere in the range of 12 to 38, but without filtering, BMIs in this set can range anywhere from 1 to 470. Since many of these are likely mistakes, filtering by BMI removes likely erroneous records. The bounds chosen are chosen based on the two following visualizations, showing the general shape of the height vs weight distributions. The dark blue areas are height/weight combinations with at least 150 records, the closer to red areas are fewer records. Each point is some height/weight combination. The first image shows the general shape of the data with no BMI filtering. The second image shows the general shape of the data after cleaning. In this way, the bottom right and top left areas which are most likely to be erroneous (short fat and tall thin) people are cleaned out, while preserving the important center.

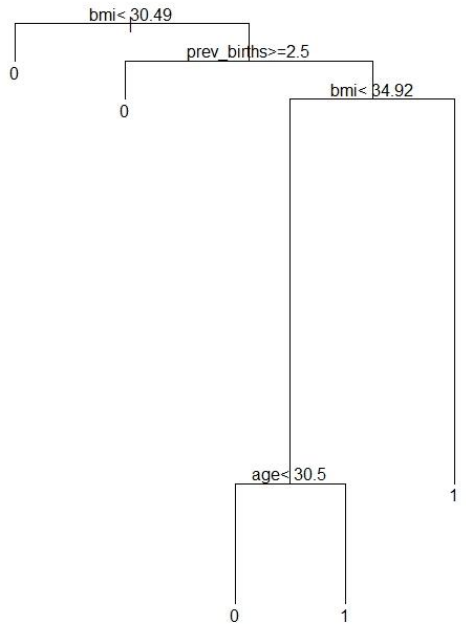





**The Model:**

The selected variables to be used in the final model are number of previous births, number of previous abortions, number of weeks the baby has been in gestation. The weight of the baby upon birth, the sex of the baby, the age of the mother, the weight of the mother, the height of the mother, whether the mother identifies as indigenous, the BMI of the mother, and whether the mother acquired an intrahospital infection. Several variables are left out from the original data set. One such variables is the number of previous pregnancies. The reason this is held out is that the number of births plus the number of abortions is nearly exactly equal to the number of previous pregnancies, save for a small number of miscarriages, so it does not add anything to the regression. The second omitted variable is the type of birth control. This is because, as we found out later, this is actually the type of birth control applied after the procedure, as opposed to used prior. So not only since it is a post birth factor does it not really lead to having a C-section, but also one of the types, tubal ligation, is a procedure well known to occurring after having had a C-section due to the ease of doing it at the same time. The model chosen is a logistic regression. The original plan was to use a decision tree, however on the first

pass of trying a decision tree, it was unable to generate any tree that did better than just guessing that no one had ever had a C-section. This is due to the fact that a tree will not split when the increase in probability from doing so is low. There was an attempt to reduce the threshold for splitting, which in general opens the possibility of more bad splits on factors more likely due to variance, however doing so did result in a simple tree which is printed below.




While the tree is reasonable, it is also sparse in terms of conclusion, since it claims that BMI, number of previous births, and age, are the only real factors at work, and that only mothers with at least 3 previous births, and either a BMI over 34.92 or over the age of 30 have C-sections. In using a log regression, it is possible to better quantify the effects of each individual variable, and realistically it makes sense that C-sections probably occur due to a combination of many values rather than strict cutoffs like are generated in a tree.

## The Results:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3822  -0.9331  -0.7647   1.2624   2.7299

Coefficients:
                 Estimate Std. Error  z value Pr(>|z|)
(Intercept)     1.205e+00  4.576e-01    2.633 0.008473 **
prev_births    -3.408e-01  3.275e-03 -104.045  < 2e-16 ***
prev_abortions  7.567e-02  7.414e-03   10.208  < 2e-16 ***
gest_weeks     -9.027e-02  1.827e-03  -49.407  < 2e-16 ***
baby_weight     2.229e-04  7.114e-06   31.337  < 2e-16 ***
baby_sex2      -4.882e-02  6.169e-03   -7.914 2.50e-15 ***
age             5.056e-02  6.241e-04   81.020  < 2e-16 ***
mother_weight   7.087e-03  2.567e-03    2.761 0.005761 **
mother_height  -4.482e-03  2.340e-03   -1.915 0.055466 .
health_ins1    -2.149e-01  1.144e-01   -1.879 0.060291 .
health_ins2     1.912e-01  9.788e-02    1.953 0.050767 .
health_ins3     6.471e-01  2.831e-01    2.286 0.022248 *
health_ins4    -1.177e+00  5.498e-01   -2.140 0.032350 *
health_ins5    -3.994e-01  5.348e-01   -0.747 0.455168
health_ins6     6.959e-01  1.299e-01    5.355 8.53e-08 ***
health_ins7    -1.761e-01  1.288e-01   -1.367 0.171584
health_ins8     3.869e-02  1.238e-02    3.125 0.001775 **
health_ins9    -6.782e-02  1.767e-02   -3.838 0.000124 ***
health_insG     1.843e-01  1.266e-01    1.455 0.145578
health_insP    -4.864e-01  6.018e-01   -0.808 0.418941
state2         -3.778e-01  2.133e-01   -1.771 0.076494 .
state3          9.585e-02  2.144e-01    0.447 0.654793
state4         -1.791e-01  2.146e-01   -0.835 0.403843
state5         -2.035e-01  2.130e-01   -0.955 0.339437
state6         -4.533e-02  2.199e-01   -0.206 0.836665
state7          8.174e-02  2.124e-01    0.385 0.700412
state8         -3.476e-01  2.128e-01   -1.634 0.102314
state9         -2.185e-01  2.133e-01   -1.024 0.305789
state10        -2.701e-01  2.135e-01   -1.265 0.205963
state11        -1.341e-01  2.122e-01   -0.632 0.527360
state12        -8.970e-03  2.124e-01   -0.042 0.966307
state13         8.871e-02  2.126e-01    0.417 0.676519
state14        -2.211e-01  2.125e-01   -1.040 0.298164
state15        -1.475e-01  2.121e-01   -0.695 0.486858
state16        -5.964e-02  2.124e-01   -0.281 0.778885
state17        -1.593e-02  2.128e-01   -0.075 0.940316
state18        -3.891e-01  2.142e-01   -1.816 0.069297 .
state19        -9.065e-02  2.163e-01   -0.419 0.675102
state20         3.911e-01  2.124e-01    1.841 0.065565 .
state21        -1.521e-02  2.123e-01   -0.072 0.942882
state22        -3.231e-01  2.129e-01   -1.518 0.129127
state23        -7.777e-02  2.133e-01   -0.365 0.715463
state24        -3.992e-01  2.133e-01   -1.871 0.061287 .
state25        -5.487e-02  2.131e-01   -0.257 0.796803
state26        -2.357e-01  2.135e-01   -1.104 0.269392
state27        -2.066e-01  2.125e-01   -0.972 0.330901
state28        -1.804e-01  2.130e-01   -0.847 0.397091
state29         8.172e-02  2.129e-01    0.384 0.701103
state30        -3.292e-01  2.125e-01   -1.549 0.121392
state31         8.483e-02  2.140e-01    0.396 0.691779
state32        -6.061e-01  2.134e-01   -2.840 0.004508 **
indigenous2     4.185e-01  1.905e-02   21.964  < 2e-16 ***
indigenous3     4.858e-01  4.692e-02   10.352  < 2e-16 ***
indigenous4     4.742e-02  4.203e-02    1.128 0.259209
infection      -6.105e-01  8.018e-02   -7.613 2.67e-14 ***
bmi             4.476e-02  6.277e-03    7.130 1.00e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At the basic level, the variables with highly significant impacts, specifically those at a 99% chance of being significant, are number of previous births, number of previous abortions, weeks the baby was in gestation, the weight of the baby, the sex of the baby, the age of the mother, the weight of the mother, whether they are on state government insurance or a program

called “Seguro Popular” or if their insurance is unrecorded, whether they live in the state of Zacatecas, their BMI, whether they contracted an intra hospital infection, and whether they do not consider themselves indigenous or did not respond to the question. The most significant factors are number of previous births, wherein having more births leads to a greater chance of C-section, age where older mothers are more likely to have C-sections, weeks in gestation where earlier born babies are more likely to be delivered by C-section, and the weight of the baby, where heavier babies are more likely to be delivered by C-section. This model has a 24.152% True Positive rate, meaning it correctly predicts 24.152% of C-sections. It also has a 7.594% False Positive rate, meaning that 7.954% of non C-sections are predicted to be C-sections. The overall correct prediction rate is 66.556%. In general this indicates that while this model is a start, it is not optimal for the purposes of predicting C-sections. The low true positive rate indicates that there are likely variables missing that would help to explain whether or not women get C-sections. I think that some of the stranger results can help lead to future work that can be done on this topic.

**Conclusion:**

While this model is not a strong predictor of C-sections, the significant variables do help to identify some of the possibly paths of further exploration that are outside the scope of this data. In particular, the insurance variables that are significant are both public insurance, which possibly indicates that there is a wealth factor at play. Since this data is all medical, there is not a particularly good way to factor in wealth other than checking insurance, which certainly is not entirely tied to wealth. However, the significance of those variables warrants further investigation into the matter. Furthermore, at least one state was highly significant and there are several more with some low levels of significance. This begs the question as to whether there might be preferences between hospitals or even individual attending doctors. The hospital data was not included due to the wide range of values in terms of visits to hospitals (some were extremely small in comparison and it seemed like a bad idea to remove them in case geographic bias was accidentally introduced), but a possible continuation using that data instead of state might reveal some more trends.

## Appendix:

```
data_births<-read.csv("OBSTET.csv")
data_babies<-read.csv("PRODUCTOS.csv")
data_main<-read.csv("EGRESO.csv")

# initial birth event cleaning
births <- subset(data_births, data_births$HAYPROD == 1) # only birth events
births <- subset(births, births$TIPATEN == 2) # only births
births <- subset(births, births$PRODUCTO == 1) # singleton births (multiple births extremely
likely to have c-section; not as interesting of a question)
births <- births[,-c(5, 6, 8)] # d
on't need this variable anymore
colnames(births) <- c("ID", "prev_preg", "prev_births", "prev_abortions", "gest_weeks",
"birth_type", "bc_type")

# baby details
babies <- subset(data_babies, !is.na(data_babies$ID)) # drop non-ID'd babies
baby_details <- babies[,c(1, 3, 4)] # grab ID, weight, sex
colnames(baby_details) <- c("ID", "baby_weight", "baby_sex") # adjust column names

births <- merge(births, baby_details, by = "ID") # add baby details to birth events

# mother details
mother_details <- data_main[, c(1, 10, 11, 14, 15, 16, 17, 20, 41)] # grab mother details
colnames(mother_details) <- c("ID", "age_code", "age", "mother_weight", "mother_height",
"health_ins", "state", "indigenous", "infection")
births <- merge(births, mother_details, by = "ID") # add the mother details to birth events
births <- births[, -10] # age code not needed; all ages in years for mothers


# clean and construct LHS variable of birth method
births <- subset(births, birth_type != 9) # drop undefined
births[births$birth_type != 3, "birth_type"] <- 0 # block all vaginal births
#births[births$birth_type == 1, "birth_type"] <- 0 # vaginal birth = 0
births[births$birth_type == 3, "birth_type"] <- 1 # c-section = 1
births$birth_type <- as.factor(births$birth_type)

# cleaning/preparing RHS. If no line for a variable, no missing values.
births <- subset(births, prev_preg != 99) # drop undefined previous pregnancies, fixes next two
variables too

## Previous births
births <- subset(births, prev_births != 0) # impossible value
births$prev_births <- births$prev_births - 1 # make it into previous births

## Gestation weeks
births <- subset(births, gest_weeks != 99) # drop missing gestation weeks

# Birth control type
births <- subset(births, bc_type != 9) # drop missing bc method
births$bc_type <- as.factor(births$bc_type)
```

```
# Baby weight
births <- subset(births, baby_weight != 9999)

# Baby sex
births <- subset(births, baby_sex != 9)
births$baby_sex <- as.factor(births$baby_sex)

# Mother weight, may need cleaning
births <- subset(births, mother_weight != 999)

# Mother height, may need cleaning
births <- subset(births, mother_height != 999)

# Health Insurance, needs binning
births$health_ins <- as.factor(births$health_ins)

# State
births <- subset(births, state != 99)
births$state <- as.factor(births$state)

# Indigenous, needs binning
births$indigenous <- as.factor(births$indigenous)

#BMI
births$bmi<-births$mother_weight/((births$mother_height / 100) ^ 2)
births<-births[births$bmi >= 8,]
births<-births[births$bmi <= 55,]
write.csv(births, "cleanedData.csv")

# Validation set
set.seed(1)
train_rows <- sample(nrow(births), .75 * nrow(births))
train <- births[train_rows,]
test <- births[-train_rows,]

# Logit
fit1 <- glm(birth_type ~ . -ID - bc_type - prev_preg, subset = train_rows, data = births, family
= binomial)
fit1_probs <- predict.glm(fit1, newdata = test, type = "response")
fit1_preds <- rep(0, length(fit1_probs))
fit1_preds[fit1_probs > .5] <- 1
table(fit1_preds, test$birth_type)
mean(fit1_preds == test$birth_type)
summary(fit1)


library(rpart)
tree.birth = rpart(birth_type ~ prev_preg + prev_births + prev_abortions + gest_weeks +
baby_weight + baby_sex + age + mother_weight + mother_height + indigenous + infection + bmi, data
= births, subset = train_rows, control=rpart.control(minsplit=1, minbucket=1, cp=0.005))
summary(tree.birth)
plot(tree.birth)
text(tree.birth,pretty=0)
```

```
tree.pred = predict(tree.birth, newdata = test, type = "class")
mean(tree.pred == test$birth_type) # just can't predict trues
```