Birth Analysis

Michael Rose

Spring 2017

Abstract

The following is a data analysis that looks at birth data from the 1995 birth registry at the North Carolina State Center for Health and Environmental Statistics. The Analysis mainly answers 2 questions:

- What Effect does Smoking and Drinking have on newborns?
- What affects variation in newborn weight?

A variety of questions were answered using different techniques.

Methods and Techniques

The following techniques were used in this analysis:

Permutation Test A permutation test is a non-parametric test that uses randomized simulations. This test was used due to the observational nature of the data. Permutation tests do not require that the data follow a normal distribution, but it does require that the samples are taken from similar populations.

Skewness & Kurtosis Tests are a way to measure the normality of a distribution without using qqnorm plots or shapiro-wilks tests.

Skewness is a measure of symmetry or lack of symmetry in a distribution.

Kurtosis is a measure of whether the data has a heavy tail or a light tail, relative to a normal distribution.

Normal Distribution would have a skewness close to 0 and a Kurtosis close to 3.

Step-wise Regression Coefficient Analysis Stepwise Regression is an iterative technique used to identify variables with a lot of effect in a regression model. In this case, backward stepwise regression was used. It starts by creating a multiple regression model with all the variables, then removes them one by one. It then checks permutations of all the explanatory variables and finds the model with the best adjusted R^2 value.

Multiple Regression Multiple Regression seeks to describe the relationship between a response variable and multiple explanatory variables. This technique gives us a regression equation with the form y = b0 + b1x1,i + ... + b(p-2)x(p-2),i + Ei where p represents the number of parameters in the regression model, i is the ith coefficient and n is the total number of observations in the data.

The assumptions for a multiple regression model are the following:

- The Model Coefficients are constant
- Each term in the model is additive
- The error terms are independent and have been distributed from a single population (iid)
- The error terms follow a normal probability distribution centered at zero with a fixed variance

Binary Logistic Regression Binary Logistic Regression allows us to look at the probability (pi) of a binary response variable given a quantitative explanatory variable. It is a case of a generalized linear model.These models consist of a linear predictor, a random component, and a link function. Our Linear predictor in this case is b0 + b1x, and our link function is ln[pi/(1 - pi)]

Assumptions:

- Each data point is independent

- Each response falls into 1 of 2 categories represented by a 0 or 1
- For any given explanatory variable, the probability of success is constant

Possible Biases in Data

Our data is comprised of 1000 babies that were born in one geographic location in North Carolina. In terms of sweeping generalizations, this is much too small a data set. A variety of things could skew our data such as

- local culture
- season (more babies are born during certain months)
- makeup of local population
- human error (recording)

Makeup of Data (Tables)

Pre-Cleaning

First we want to load our data and set all the blank cells to NA. We also want to change our marital system from 1,2 to a binary 0,1 system.

```
births[births == ""] <- NA

# set marital to binary 1 0 system
births$marital[births$marital == 2] <- 0
```

Non-Binary Variables

Mother Race

```
# Non-Bin Vars = Race, Mother Age, Education Level, Gestation
# (Weeks), Apgar Score, # Children, Weight of Child (Ounces)

# Race

Nonwhite_Other <- length(c(births$race[births$race == "0"], na.rm = TRUE))
White <- length(c(births$race[births$race == "1"], na.rm = TRUE))
Black <- length(c(births$race[births$race == "2"], na.rm = TRUE))
American_Indian <- length(c(births$race[births$race == "3"],
    na.rm = TRUE))
Chinese <- length(c(births$race[births$race == "4"], na.rm = TRUE))
Japanese <- length(c(births$race[births$race == "5"], na.rm = TRUE))
Hawaiian <- length(c(births$race[births$race == "6"], na.rm = TRUE))
Fillipino <- length(c(births$race[births$race == "7"], na.rm = TRUE))
Other <- length(c(births$race[births$race == "8"], na.rm = TRUE))
unknown_race <- length(c(births$sex[is.na(births$sex)]))

races_1 <- c(Nonwhite_Other, White, Black, American_Indian, Chinese,
    Japanese, Hawaiian, Fillipino, Other)

races_mean <- mean(races_1)
```

```
races_sd <- sd(races_1)

races <- c(Nonwhite_Other, White, Black, American_Indian, Chinese,
    Japanese, Hawaiian, Fillipino, Other, races_mean, races_sd)

race_names_lat = c("Nonwhite_Other", "White", "Black", "American_Indian",
    "Chinese", "Japanese", "Hawaiian", "Fillipino", "Other",
    "Mean", "Standard Deviation")

DF_NonBin <- data.frame(Race = race_names_lat, Values = races)

format_table(DF_NonBin, formatters = list(Race = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Race	Values
Nonwhite_Other	1
White	701
Black	267
American_Indian	18
Chinese	2
Japanese	1
Hawaiian	1
Fillipino	1
Other	17
Mean	112
Standard Deviation	237

Mother Age

```
lt_20 <- length(births$mothage[births$mothage > 0 & births$mothage <
    21])
lt_30 <- length(births$mothage[births$mothage > 20 & births$mothage <
    31])
lt_40 <- length(births$mothage[births$mothage > 30 & births$mothage <
    41])
lt_50 <- length(births$mothage[births$mothage > 40 & births$mothage <
    51])
gt_50 <- length(births$mothage[births$mothage > 51])

age_mean = mean(births$mothage)
age_sd = sd(births$mothage)

mum_names <- c("Mothers Under 20", "Mothers 21 - 30", "Mothers 31 - 40",
    "Mothers 41 - 50", "Mothers Over 50", "Mean Age of Mothers",
    "Standard Deviation")

mothage <- c(lt_20, lt_30, lt_40, lt_50, gt_50, age_mean, age_sd)

df_mothage <- data.frame(Age = mum_names, Values = mothage)

format_table(df_mothage, formatters = list(Age = color_tile("lightblue",
```

```
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Age	Values
Mothers Under 20	221
Mothers 21 - 30	526
Mothers 31 - 40	251
Mothers 41 - 50	2
Mothers Over 50	0
Mean Age of Mothers	26
Standard Deviation	6

Education Level

```
m_ed_hs <- length(births$mothed[births$mothed < 13])
m_ed_as <- length(births$mothed[births$mothed < 16 & births$mothed >
    12])
m_ed_bh <- length(births$mothed[births$mothed > 15])
m_ed_mean <- mean(births$mothed, na.rm = TRUE)
m_ed_sd <- sd(births$mothed, na.rm = TRUE)

m_ed_names <- c("High School or less", "Some College", "Bachelors + ",
    "Mean Years Completed", "Standard Deviation")

m_ed_vars <- c(m_ed_hs, m_ed_as, m_ed_bh, m_ed_mean, m_ed_sd)

df_mothed <- data.frame(Education = m_ed_names, Values = m_ed_vars)

format_table(df_mothed, formatters = list(Education = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Education	Values
High School or less	549
Some College	242
Bachelors +	211
Mean Years Completed	13
Standard Deviation	3

Gestation Period

```
m_gp_35 <- length(births$gest[births$gest < 36])
m_gp_40 <- length(births$gest[births$gest > 35 & births$gest <
    41])
m_gp_45 <- length(births$gest[births$gest > 40 & births$gest <
    46])
m_gp_45plus <- length(births$gest[births$gest > 45])
m_gp_mean <- mean(births$gest, na.rm = TRUE)
m_gp_sd <- sd(births$gest, na.rm = TRUE)

m_gp_names <- c("Gestation Period < 35 Weeks", "Gestation Period 36 - 40 Weeks",
```

```
    "Gestation Period 41 - 45 Weeks", "Gestation Period > 46 Weeks",
    "Mean", "Standard Deviation")

m_gp_vars <- c(m_gp_35, m_gp_40, m_gp_45, m_gp_45plus, m_gp_mean,
    m_gp_sd)

df_gest <- data.frame(Gestation = m_gp_names, Values = m_gp_vars)

format_table(df_gest, formatters = list(Education = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Gestation	Values
Gestation Period < 35 Weeks	74
Gestation Period 36 - 40 Weeks	718
Gestation Period 41 - 45 Weeks	210
Gestation Period > 46 Weeks	1
Mean	39
Standard Deviation	3

Kessner Score

```
m_ap_good <- length(births$apgar1[births$apgar1 == 1])
m_ap_okay <- length(births$apgar1[births$apgar1 == 2])
m_ap_bad <- length(births$apgar1[births$apgar1 == 3])
m_ap_uk <- length(births$apgar1[births$apgar1 == 4])
m_ap_mean <- mean(births$apgar1, na.rm = TRUE)
m_ap_sd <- sd(births$apgar1, na.rm = TRUE)

m_ap_names <- c("Adequate Kessner Score", "Intermediate Kessner Score",
    "Inadequate Kessner Score", "Unknown Kessner Score", "Mean",
    "Standard Deviation")

m_ap_vars <- c(m_ap_good, m_ap_okay, m_ap_bad, m_ap_uk, m_ap_mean,
    m_ap_sd)

df_apgar1 <- data.frame(Kessner = m_ap_names, Values = m_ap_vars)

format_table(df_apgar1, formatters = list(Kessner = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Kessner	Values
Adequate Kessner Score	790.0
Intermediate Kessner Score	159.0
Inadequate Kessner Score	50.0
Unknown Kessner Score	1.0
Mean	1.3
Standard Deviation	0.5

Number of Children Born

```
m_pl_1 <- length(births$plural[births$plural == 1])
m_pl_2 <- length(births$plural[births$plural == 2])
m_pl_3plus <- length(births$plural[births$plural > 2])
m_pl_mean <- mean(births$plural, na.rm = TRUE)
m_pl_sd <- sd(births$plural, na.rm = TRUE)

m_pl_names <- c("One Child", "Twins", "Triplets + ", "Mean",
    "Standard Deviation")

m_pl_vars <- c(m_pl_1, m_pl_2, m_pl_3plus, m_pl_mean, m_pl_sd)

df_plural <- data.frame(Plural = m_pl_names, Values = m_pl_vars)

format_table(df_plural, formatters = list(Plural = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Plural	Values
One Child	979.0
Twins	20.0
Triplets +	1.0
Mean	1.0
Standard Deviation	0.2

Weight of Child

```
m_to_100 <- length(births$totounc[births$totounc < 100])
m_to_129 <- length(births$totounc[births$totounc > 99 & births$totounc <
    130])
m_to_160 <- length(births$totounc[births$totounc > 129 & births$totounc <
    161])
m_to_160plus <- length(births$totounc[births$totounc > 160])
m_to_mean <- mean(births$totounc, na.rm = TRUE)
m_to_sd <- sd(births$totounc, na.rm = TRUE)

m_to_names <- c("Weight < 100 Oz", "Weight 100 - 129 Oz", "Weight 130 - 160 Oz",
    "Weight > 160 Oz", "Mean", "Standard Deviation")

m_to_vars <- c(m_to_100, m_to_129, m_to_160, m_to_160plus, m_to_mean,
    m_to_sd)

df_totounc <- data.frame(Total_Ounces = m_to_names, Values = m_to_vars)

format_table(df_totounc, formatters = list(Total_Ounces = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

Total_Ounces	Values
Weight < 100 Oz	174
Weight 100 - 129 Oz	575
Weight 130 - 160 Oz	243

Total_Ounces	Values
Weight > 160 Oz	14
Mean	117
Standard Deviation	21

Binary Variables

```
# lets isolate our first variables by value, Sex

male <- length(c(births$sex[births$sex == 1], na.rm = TRUE)) -
    1
female <- length(c(births$sex[births$sex == 0], na.rm = TRUE)) -
    1
unknown_sex <- length(c(births$sex[is.na(births$sex)]))

sexes <- c(male, female, unknown_sex)

# Marital

not_married <- length(c(births$marital[births$marital == 0],
    na.rm = TRUE)) - 1
married <- length(c(births$marital[births$marital == 1], na.rm = TRUE)) -
    1
unknown_marital <- length(c(births$marital[is.na(births$marital)]))

marital <- c(not_married, married, unknown_marital)

# Smoker

non_smoker <- length(c(births$cigs[births$cigs == 0], na.rm = TRUE)) -
    1
smoker <- length(c(births$cigs[births$cigs == 1], na.rm = TRUE)) -
    1
unknown_smoke <- length(c(births$cigs[is.na(births$cigs)]))

cigs <- c(non_smoker, smoker, unknown_smoke)

# Drinker

non_drinker <- length(c(births$drinks[births$drinks == 0], na.rm = TRUE)) -
    1
drinker <- length(c(births$drinks[births$drinks == 1], na.rm = TRUE)) -
    1
unknown_drinker <- length(c(births$drinks[is.na(births$drinks)]))

drinks <- c(non_drinker, drinker, unknown_drinker)

# Fetal Alcohol Syndrome

non_fac <- length(c(births$fac[births$fac == 0], na.rm = TRUE)) -
    1
is_fac <- length(c(births$cigs[births$fac == 1], na.rm = TRUE)) -
```

```
    1
unknown_fac <- length(c(births$fac[is.na(births$fac)]))

fac <- c(non_fac, is_fac, unknown_fac)

# Birth Weight

non_babw <- length(c(births$btotounc[births$btotounc == 0], na.rm = TRUE)) -
    1
is_babw <- length(c(births$btotounc[births$btotounc == 1], na.rm = TRUE)) -
    1
unknown_babw <- length(c(births$btotounc[is.na(births$btotounc)]))

babw <- c(non_babw, is_babw, unknown_babw)

names_lat <- c("Male", "Female", "Unknown Sex", "Not Married",
    "Married", "Unknown Marital", "Non-Smoker", "Smoker", "Unknown Smoking",
    "Non-Drinker", "Drinker", "Unknown if Drinker", "No Fetal Alcohol Syndrome",
    "Fetal Alcohol Syndrome", "Fetal Alcohol Syndrome not Recorded",
    "Normal Birth Weight", "Below Normal Birth Weight", "Birth Weight not Recorded")
vals_lat <- c(sexes, marital, cigs, drinks, fac, babw)


DF_Bin <- data.frame(Categories = names_lat, Values = vals_lat)

format_table(DF_Bin, formatters = list(Categories = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l")
```

Categories	Values
Male	489
Female	511
Unknown Sex	0
Not Married	332
Married	669
Unknown Marital	1
Non-Smoker	827
Smoker	173
Unknown Smoking	0
Non-Drinker	984
Drinker	16
Unknown if Drinker	0
No Fetal Alcohol Syndrome	999
Fetal Alcohol Syndrome	1
Fetal Alcohol Syndrome not Recorded	0
Normal Birth Weight	920
Below Normal Birth Weight	80
Birth Weight not Recorded	0

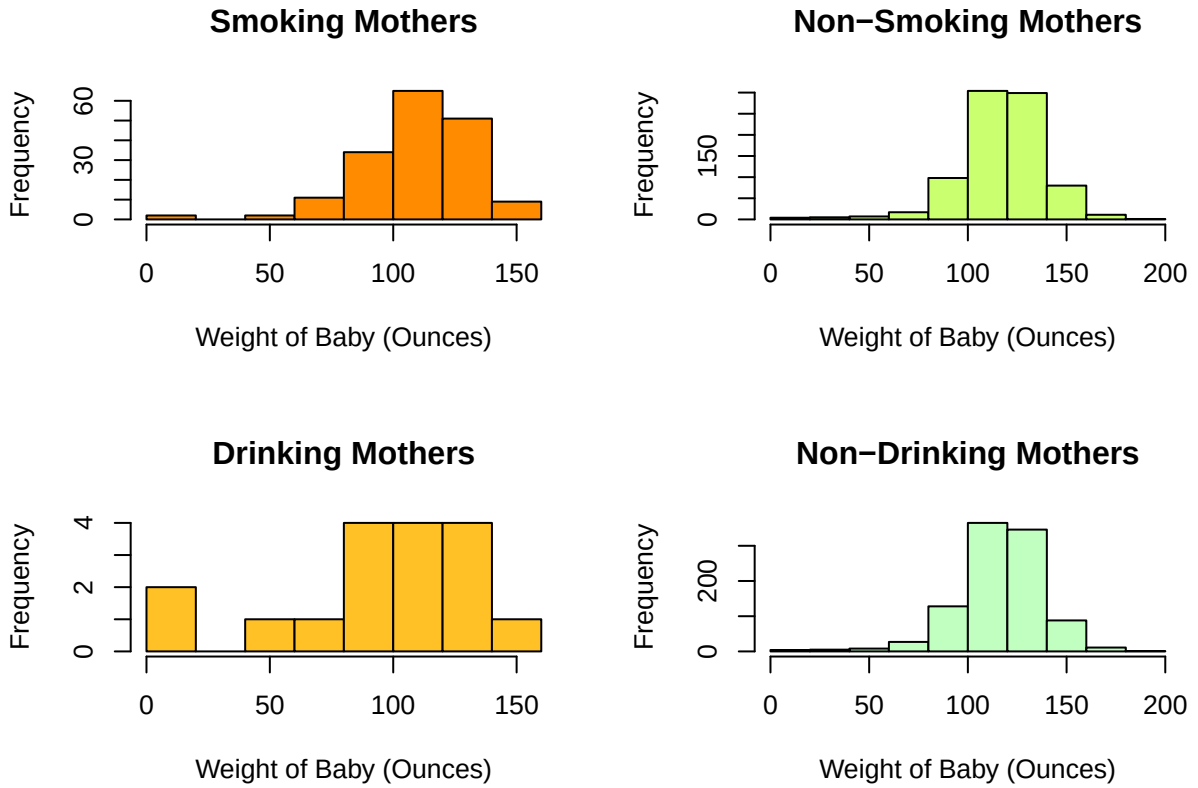
What effect does Smoking and Drinking have on the Baby?

First, lets seperate the variables by habit / non-habit mothers:

```r
# birth weights of non-smoker mothers
non_smoker_mothers <- c(births$totounc[births$cigs == 0], na.rm = TRUE)

# birth weights of smoker mothers
smoker_mothers <- c(births$totounc[births$cigs == 1], na.rm = TRUE)

# birth weights of non-drinking mothers
non_drinking_mothers <- c(births$totounc[births$drinks == 0],
    na.rm = TRUE)

# birth weights of drinking mothers
drinking_mothers <- c(births$totounc[births$drinks == 1], na.rm = TRUE)
```

Visualization:

```r
par(mfrow = c(2, 2))
hist(smoker_mothers, col = "darkorange", main = "Smoking Mothers",
    xlab = "Weight of Baby (Ounces)")
hist(non_smoker_mothers, col = "darkolivegreen1", main = "Non-Smoking Mothers",
    xlab = "Weight of Baby (Ounces)")
hist(drinking_mothers, col = "goldenrod1", main = "Drinking Mothers",
    xlab = "Weight of Baby (Ounces)")
hist(non_drinking_mothers, col = "darkseagreen1", main = "Non-Drinking Mothers",
    xlab = "Weight of Baby (Ounces)")
```



We can see from the plots above that the Smoking Mothers have a distribution that is slightly left-skewed compared to the non-smoking mothers which seem to have a more normal distribution. That being said, it is a little unclear. The drinking mothers have a very small sample size (n = 17), so it is tough to reason about normality. The non-drinking mothers have a similar distribution to non-smoking mothers.

Let's test to see if our subsets are normally distributed with skewness and kurtosis values. I would have used the Shapiro Wilks test or QQNorm visualization, but from recent reading it seems that they are unreliable for smaller sample sizes or for small deviations with larger sample sizes. If they are normal, we can use a 2 sample t-test, which requires normality. If not, we can use a non-parametric test.

```
# Smoker Mothers
skewness(smoker_mothers)
```

```
## [1] -1.163797
```

```
kurtosis(smoker_mothers)
```

```
## [1] 6.458658
```

```
# Non-Smoker Mothers
skewness(non_smoker_mothers, na.rm = TRUE)
```

```
## [1] -1.039219
```

```
kurtosis(non_smoker_mothers, na.rm = TRUE)
```

```
## [1] 6.859538
```

```
# Drinking Mothers
skewness(drinking_mothers)
```

```
## [1] -0.9275618
```

```
kurtosis(drinking_mothers)
```

```
## [1] 3.125236
```

```
# Non-Drinking Mothers
skewness(non_drinking_mothers, na.rm = TRUE)
```

```
## [1] -0.9076834
```

```
kurtosis(non_drinking_mothers, na.rm = TRUE)
```

```
## [1] 6.214338
```

For the values above, the first value is a measure of the skewness of our subset, and the second value is a measure of the kurtosis of our subset.

Skewness is a measure of symmetry or lack of symmetry in a distribution.

Kurtosis is a measure of whether the data has a heavy tail or a light tail, relative to a normal distribution.

A **Normal Distribution** would have a skewness close to 0 and a Kurtosis close to 3

From our skewness values above, we find that all of the data sets are slightly skewed to the left (indicated by our negative value).

From our kurtosis values above, we see that all of our data sets except Drinking Mothers have a strongly positive value slightly above our normal goal of 3. This indicates we have a leptokurtic distribution with a fat tail.

As we do not have values within a reasonable range, we will say that these are not normal and move ahead with a non-parametric test.

First we compare the means:

```
# Mean Smoker Mothers
print(mean(smoker_mothers))
```

```
## [1] 109.7586
```

```
# Mean Non-Smoker Mothers
print(mean(non_smoker_mothers, na.rm = TRUE))
```

```
## [1] 118.0339
```

```
# Mean Drinking Mothers
print(mean(drinking_mothers, na.rm = TRUE))
```

```
## [1] 95
```

```
# Mean Non-Drinking Mothers
print(mean(non_drinking_mothers, na.rm = TRUE))
```

```
## [1] 116.9674
```

From the two smoking means above, we can see there is a difference of 8.2753 ounces or roughly 1/2 lb.

From the two drinking means above, we can see there is a difference of 21.9674 ounces or 1.373 lbs.

Let's let

```
H0: mean(smoker) = mean(non-smoker)
Ha: mean(smoker) < mean(non-smoker)

and

H0b: mean(drinker) = mean(non-drinker)
Hab: mean(drinker) < mean(non-drinker)
```

Since this is an observational study, we do not have a random allocation to groups. We will use a permutation test. The permutation test does not require that the data follow a normal distribution, but we must choose from 2 populations that are similar. Since we can assume this single hospital has similar enough mothers, we can move forward with the test:

Permutation Test

Smoker

```
reps <- 10000
results <- numeric(reps)
combo_smokers <- c(smoker_mothers, non_smoker_mothers)

for (i in 1:reps) {
    temp <- sample(combo_smokers)
    results[i] <- mean(temp[1:5] - temp[6:10], na.rm = TRUE)
}

p.value_smokers <- sum(results >= 8.2753)/reps
p.value_smokers
```

```
## [1] 0.2782
```

From our test above, we can see that our p-value is not statistically significant. As a result, we fail to reject our null hypothesis. We can not say that smoking causes a lower mean birth rate.

Approximately 1 in every 4 groups of 1000 babies born may result in a mean difference equal or greater than 8.275 between the 2 groups.

Drinker

```
reps2 <- 10000
results2 <- numeric(reps)
combo_drinkers <- c(drinking_mothers, non_drinking_mothers)

for (i in 1:reps) {
    temp2 <- sample(combo_drinkers)
    results2[i] <- mean(temp2[1:5] - temp2[6:10], na.rm = TRUE)
}

p.value_drinkers <- sum(results >= 21.9674)/reps
p.value_drinkers
```

```
## [1] 0.0625
```

From our test above, we can see that our p-value is not statistically significant. It is very much on the line. Perhaps if we had a larger sample size for mothers who have drank then we could be more certain. In this case, we fail to reject our null hypothesis that our means are equal.

Approximately 1 in every 20 groups of 1000 babies born may result in a mean difference greater than or equal to 21.9674 ounces between the 2 groups.

From our results above, we can say that Drinking or Smoking by the Mother does not have a statistically significant effect on birth weight. More studies will need to be done to be more conclusive.

What about Fetal Alcohol Syndrome?

In our data we only had 1 incidence of Fetal Alcohol Syndrome. This particular mother also had no alcohol claimed. In this incidence, our data would report a 0% correlation with Mothers using alcohol. That being said, it is very likely that the mother lied about her alcohol use.

What about Gestation Periods?

```
# Non-Drinking
non_drink_gest <- births$gest[births$drinks == 0]
print(mean(non_drink_gest, na.rm = TRUE))
```

```
## [1] 38.96948
```

```
# Drinking
drink_gest <- births$gest[births$drinks == 1]
print(mean(drink_gest, na.rm = TRUE))
```

```
## [1] 37
```

```
# Non-Smoking
non_smoke_gest <- births$gest[births$cigs == 0]
print(mean(non_smoke_gest, na.rm = TRUE))
```

```
## [1] 38.99153
```

```
# Smoking
smoke_gest <- births$gest[births$cigs == 1]
print(mean(smoke_gest, na.rm = TRUE))
```

```
## [1] 38.68208
```

From our mean values above, we can see that there is roughly a 2 week difference in mean gestation periods between non-drinking and drinking mothers. For smoking mothers, there is less than 1 week difference in gestation period.

What makes the variation in birth weight?

First lets take a look at each variable's effect on birth weight:

```
# remove all rows with NA
births2 <- births[complete.cases(births), ]

# combine all variables expect ones related to birth weight
full_data <- cbind(births2$sex, births2$race, births2$mothage,
    births2$mothed, births2$gest, births2$marital, births2$cigs,
    births2$drinks, births2$apgar1, births2$fac, births2$plural)

# identify best models
best_model <- leaps(full_data, births2$totounc, method = "adjr2",
    names = c("Sex", "Race", "Mother Age", "Mother Education",
        "Gestation Period", "Marital Status", "Cigarettes", "Alcohol",
        "Kessner Score", "Fetal Alcohol Syndrome", "# of Babies"))

data_out <- cbind(best_model$which, best_model$adjr2)

# Find max R^2 value for permutations of model
which.max(data_out[, 12])
```

```
##  9
## 81
```

```
print(data_out[81, ])
```

```
##                   Sex                  Race            Mother Age
##             1.0000000             1.0000000             1.0000000
##       Mother Education       Gestation Period         Marital Status
##             1.0000000             1.0000000             1.0000000
##             Cigarettes               Alcohol          Kessner Score
##             1.0000000             0.0000000             1.0000000
## Fetal Alcohol Syndrome           # of Babies
##             0.0000000             1.0000000             0.3608963
```

From our best subsets model above, we see that the best parameters for determining our birth weight would be

```
Sex, Race, Mother Age, Mother Education, Gestation Period,
Marital Status, Cigarettes, Kessner Score, # of Babies
```

with an adjusted R^2 score of 0.3608. Lets create a multiple regression model with these:

```
mothers_MRM <- lm(births2$totounc ~ births2$sex + births2$race +
    births2$mothage + births2$mothed + births2$gest + births2$marital +
    births2$cigs + births2$apgar1 + births2$plural)

summary(mothers_MRM)
```

```
##
## Call:
## lm(formula = births2$totounc ~ births2$sex + births2$race + births2$mothage +
##     births2$mothed + births2$gest + births2$marital + births2$cigs +
##     births2$apgar1 + births2$plural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.010 -10.639   0.291  10.503  60.159
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23.7286    11.2605  -2.107 0.035349 *
## births2$sex       3.4511     1.0883   3.171 0.001565 **
## births2$race     -2.1873     0.5659  -3.865 0.000118 ***
## births2$mothage   0.1854     0.1077   1.721 0.085595 .
## births2$mothed    0.3814     0.2494   1.529 0.126484
## births2$gest      4.0715     0.2261  18.011  < 2e-16 ***
## births2$marital   2.1204     1.3433   1.578 0.114773
## births2$cigs     -5.0289     1.4822  -3.393 0.000719 ***
## births2$apgar1   -2.8719     1.0595  -2.711 0.006833 **
## births2$plural  -22.6358     3.6457  -6.209 7.85e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.04 on 985 degrees of freedom
## Multiple R-squared:  0.3667, Adjusted R-squared:  0.3609
## F-statistic: 63.37 on 9 and 985 DF,  p-value: < 2.2e-16
```

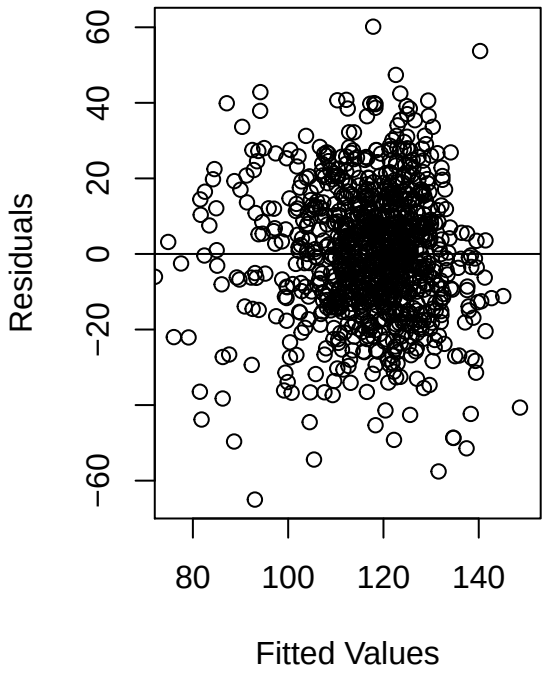
With our multiple regression model above, we get the following equation:

```
Birth Weight = -23.7286 + 3.4511 * Sex - 2.1873 * Race
+ 0.1854 * Mother's Age + 0.3814 * Mother's Education Level
+ 4.0715 * Gestation Period + 2.1204 * Marital Status
- 5.0289 * Cigarette Use - 2.8719 * Kessner Score
- 22.6358 * Number of Babies Born
```

Before we can accept our multiple regression model, we must check our assumptions for the multiple regression model.

Homoscedasticity of residuals

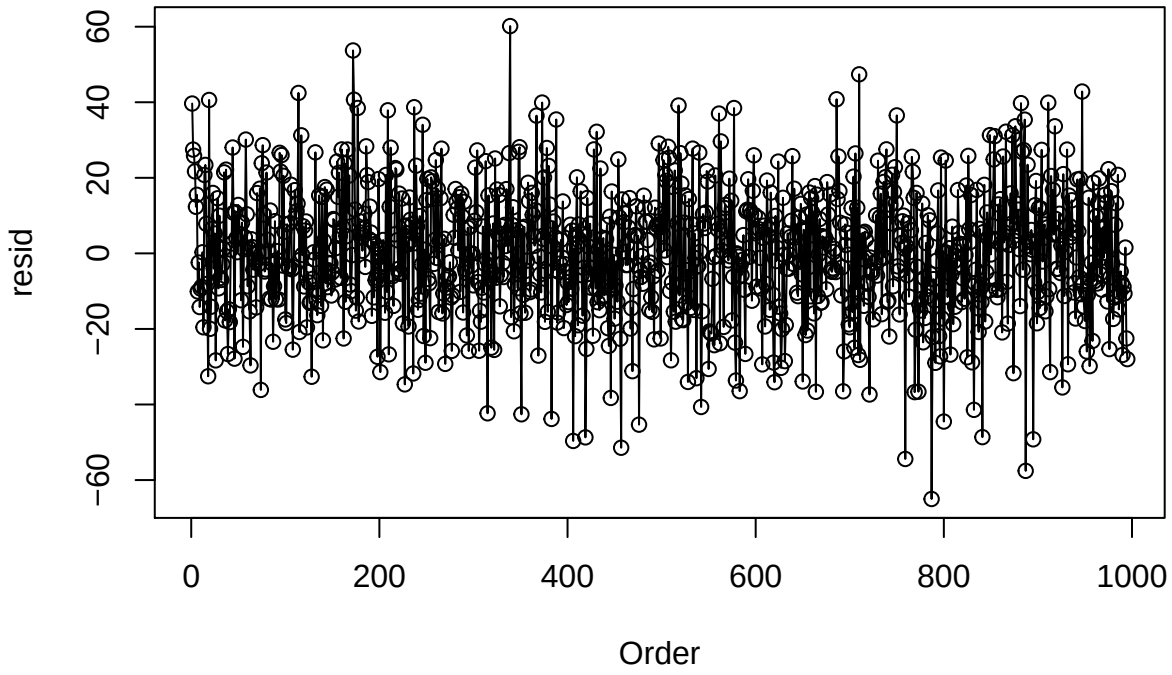
```r
resid <- mothers_MRM$residuals

par(mfrow = c(1, 2))
plot(mothers_MRM$fitted.values, resid, xlim = c(75, 150), xlab = "Fitted Values",
    ylab = "Residuals")
abline(h = 0)
```



From our plot of residuals above, we can see that there is no particular clustering or wedging. It does look like someone dropped a pile of coins onto a table.

Autocorrelation

```
plot(seq(1:length(resid)), resid, xlab = "Order")
lines(seq(1:length(resid)), resid)
```



In our ordered residual plot above, we do not see any patterns that stand out. We can conclude that we do not have autocorrelation.

To sum up the question, the variables that have the best prediction values for birth weight are:

```
Sex, Race, Mother Age, Mother Education, Gestation Period,
Marital Status, Cigarettes, Kessner Score, # of Babies
```

and their individual weights are given by the model:

```
Birth Weight = -23.7286 + 3.4511 * Sex - 2.1873 * Race
+ 0.1854 * Mother's Age + 0.3814 * Mother's Education Level
+ 4.0715 * Gestation Period + 2.1204 * Marital Status
- 5.0289 * Cigarette Use - 2.8719 * Kessner Score
- 22.6358 * Number of Babies Born
```

In our model above, the coefficients of our variables give a relative weight to each parameter. From the model, we can see that the effect on babies weight is given by the following (largest effect to smallest effect):

- Number of Babies Born
- Cigarette Use
- Gestation Period
- Sex
- Kessner Score
- Race
- Marital Status
- Mother's Education Level
- Mother's Age

The variables that increase weight are:

- Gestation Period
- Sex
- Marital Status
- Mother's Education Level
- Mother's Age

and the variables that reduce weight are:

- Number of Babies Born
- Cigarette Use
- Kessner Score
- Race

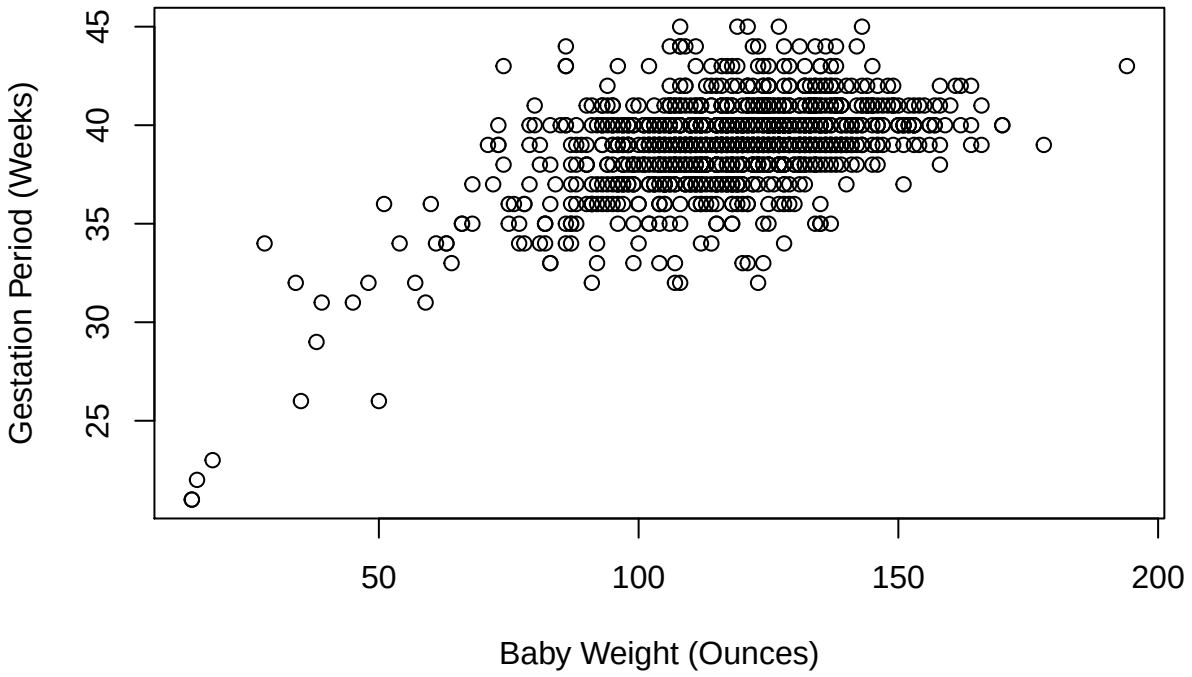
What about Gestation specifically?

We can measure the correlation between Gestation and Birth Weight independently with the following:

```
cor(births2$totounc, births2$gest)
```

```
## [1] 0.5366387
```

```
plot(births2$totounc, births2$gest, xlab = "Baby Weight (Ounces)",
    ylab = "Gestation Period (Weeks)")
```



From our correlation value above, we can see that Gestation and Birth Weight are closely related. Our correlation coefficient is measured from -1 <= r <= 1. We have a mildly positive correlation. From our plot, we can see there is some grouping around the normal baby weight values. All the outliers are relatively isolated from the majority, which is expected.

What about Kessner Score specifically?

Our Kessner score is defined by a series of prenatal tests that result in a score in the range 1-4.

- 1: Adequate
- 2: Intermediate
- 3: Inadequate
- 4: Unknown

From our tables above, we had the following values:

```
format_table(df_apgar1, formatters = list(Kessner = color_tile("lightblue",
    "lightblue")), format("pandoc"), align = "l", digits = 1)
```

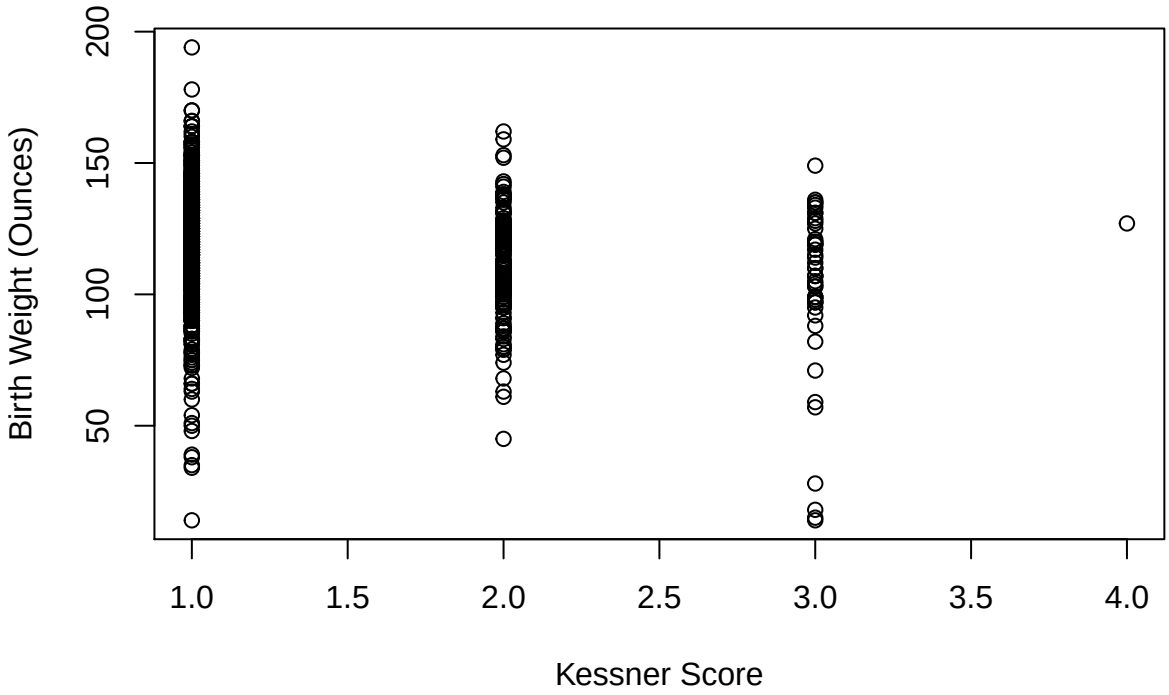
Kessner	Values
Adequate Kessner Score	790.0
Intermediate Kessner Score	159.0
Inadequate Kessner Score	50.0
Unknown Kessner Score	1.0
Mean	1.3
Standard Deviation	0.5

We can see if there is a correlation between our Kessner Score and Birth Weight with the following function:

```
cor(births2$apgar1, births2$totounc)
```

```
## [1] -0.1900242
```

```r
plot(births2$apgar1, births2$totounc, xlab = "Kessner Score",
    ylab = "Birth Weight (Ounces)")
```



From our result and accompanying plot, we can see that there is a weak negative correlation between Kessner Score and Birth Weight. This can be interpreted as mean birth weight goes down slightly as our Kessner Score gets higher (and the Kessner Tests show less adequate scores).

To find out whether babies with normal birth weight tend to have lower Kessner Scores (indicating an adequate health score), we can first seperate our dataset by babies with and without normal birth weights:

```r
# normal birth weight kessner scores
norm_babw_kessner <- c(births$apgar1[births$btotounc == 0], na.rm = TRUE)

print(mean(norm_babw_kessner))
```

```
## [1] 1.243214
```

```r
# below normal birth weight kessner scores
below_babw_kessner <- c(births$apgar1[births$btotounc == 1],
    na.rm = TRUE)

print(mean(below_babw_kessner))
```

```
## [1] 1.469136
```

From our values above, we can see that the means aren't that different for normal and non-normal birth weights. Below normal birth weights tend to have a slightly more likely incidence of intermediate or inadequate scores than normal birth weight babies. This is shown by the higher mean value for below normal weight babies (higher score referencing worse kessner score).

What about Mother's Age?

If we are given the mother's age, what is the probability that the baby will be a healthy weight?

We can look at this with a logistic regression to get our probability for different ages.

```
# Create Logistic Regression Model

births_logreg <- glm(births$btotounc ~ births$mothage, family = binomial(logit))

summary(births_logreg)
```

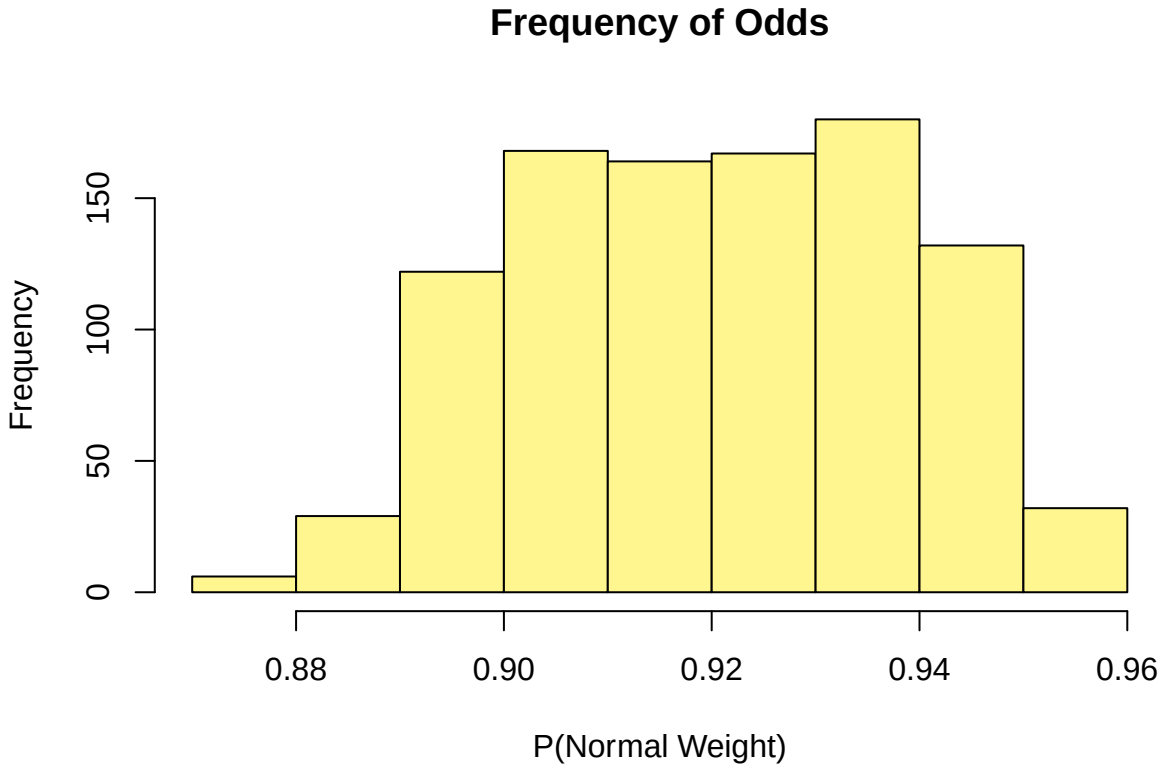
```
##
## Call:
## glm(formula = births$btotounc ~ births$mothage, family = binomial(logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5193  -0.4365  -0.3958  -0.3586   2.4525
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.40458    0.51256  -2.740  0.00614 **
## births$mothage -0.04085    0.02012  -2.031  0.04229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 557.54  on 999  degrees of freedom
## Residual deviance: 553.29  on 998  degrees of freedom
## AIC: 557.29
##
## Number of Fisher Scoring iterations: 5
```

```
# transform data

births_mothage_order <- (births[order(births$mothage), ])
births_mothage <- as.numeric(unlist(births_mothage_order$mothage))



xform <- c(1:1000)
Age_names <- c(1:1000)

for (i in 1:1000) {
    xform[i] <- (1 - (exp(-1.40458 - 0.04085 * births_mothage[i]))/(1 +
        exp(-1.40458 - 0.04085 * births_mothage[i])))
    Age_names[i] <- sprintf("Age = %d", births_mothage[i])
}

xform <- as.numeric(xform)

# Combine variables by Age and assign value to mean of age

xform_agg <- aggregate(x = xform, by = list(Age_names), FUN = mean)

print(xform_agg)
```

```
##     Group.1         x
```

```
## 1  Age = 13 0.8738713
## 2  Age = 14 0.8783054
## 3  Age = 15 0.8826046
## 4  Age = 16 0.8867716
## 5  Age = 17 0.8908089
## 6  Age = 18 0.8947193
## 7  Age = 19 0.8985057
## 8  Age = 20 0.9021707
## 9  Age = 21 0.9057173
## 10 Age = 22 0.9091483
## 11 Age = 23 0.9124665
## 12 Age = 24 0.9156748
## 13 Age = 25 0.9187759
## 14 Age = 26 0.9217727
## 15 Age = 27 0.9246680
## 16 Age = 28 0.9274646
## 17 Age = 29 0.9301652
## 18 Age = 30 0.9327726
## 19 Age = 31 0.9352893
## 20 Age = 32 0.9377182
## 21 Age = 33 0.9400617
## 22 Age = 34 0.9423225
## 23 Age = 35 0.9445030
## 24 Age = 36 0.9466058
## 25 Age = 37 0.9486332
## 26 Age = 38 0.9505877
## 27 Age = 39 0.9524715
## 28 Age = 40 0.9542869
## 29 Age = 41 0.9560362
```

```
hist(xform, col = "khaki1", xlab = "P(Normal Weight)", main = "Frequency of Odds")
```



From our table and histogram above, we can see that the rates for each age group are generally pretty high. They have a slight trend towards more normal weights as age increases. This can be seen in our summary of the logistic regression model, as our B0 and B1 variables are both negative (moving towards 0, which is normal baby weight).

We can say that age of mother gives a slight benefit towards normal baby weights.

Discussion and Results

To reiterate the questions asked in this analysis:

Does Smoking or Drinking have an effect on the weight of the newborn?

- Statistically significant results were not found. Drinking came very close to being statistically significant. These results may be different with a larger sample size, as there was only 17 mothers who had reported drinking during pregnancy

Is there a correlation between the mother drinking and Fetal Alcohol Syndrome?

- We only had 1 case of Fetal Alcohol Syndrome in the data, and the mother reported that she had not drank during the pregnancy. No conclusion can be drawn from this sample of 1000.

Is there a difference between gestation periods for mothers who smoked or drank?

- For mothers who smoked there was a difference of less than 1 week average gestation period over mothers who did not smoke. For mothers who drank there was less than 2 weeks difference between drinking and non-drinking mothers. These numbers are close enough that it is inconclusive.

What factors lead to a change in newborn birth weight?

From greatest to least:

- Number of Babies Born

- Cigarette Use
- Gestation Period
- Sex
- Kessner Score
- Race
- Marital Status
- Mother's Education Level
- Mother's Age

Does length of Gestation correlate to birth weight?

We found a .54 correlation coefficient between length of gestation and birth weight. Our correlation coefficient is measured between -1 <= r <= 1. Our result is not quite strong (0.8), but just beyond the weak boundary (0.5). We have a mildly positive correlation between Gestation length and Birth Weight.

Does Kessner Score correlate with Birth Weight?

From our mean values for below normal weight babies and normal weight babies, we don't see much of a difference. Normal Weight Babies had a mean Kessner Index score of 1.24 and Below Weight Babies had a mean Kessner Index score of 1.46. Below normal birth weights tend to have a slightly more likely incidence of intermediate or inadequate scores than normal birth weight babies, but not significantly so.

Given the mother's age, what is the probability that the baby will be a healthy weight?

```
print(xform_agg)
```

```
##     Group.1         x
## 1  Age = 13 0.8738713
## 2  Age = 14 0.8783054
## 3  Age = 15 0.8826046
## 4  Age = 16 0.8867716
## 5  Age = 17 0.8908089
## 6  Age = 18 0.8947193
## 7  Age = 19 0.8985057
## 8  Age = 20 0.9021707
## 9  Age = 21 0.9057173
## 10 Age = 22 0.9091483
## 11 Age = 23 0.9124665
## 12 Age = 24 0.9156748
## 13 Age = 25 0.9187759
## 14 Age = 26 0.9217727
## 15 Age = 27 0.9246680
## 16 Age = 28 0.9274646
## 17 Age = 29 0.9301652
## 18 Age = 30 0.9327726
## 19 Age = 31 0.9352893
## 20 Age = 32 0.9377182
## 21 Age = 33 0.9400617
## 22 Age = 34 0.9423225
## 23 Age = 35 0.9445030
## 24 Age = 36 0.9466058
## 25 Age = 37 0.9486332
## 26 Age = 38 0.9505877
## 27 Age = 39 0.9524715
## 28 Age = 40 0.9542869
## 29 Age = 41 0.9560362
```

In general, most of the time that baby will be a normal, healthy weight. There seems to be a positive trend

as age increases of having a healthy weight baby.