CHICAGO BOOTH

Project: Identifying high risk pregnancies

This is an individual project. Hand in by December 11th at Gleacher Center Suite 430 or Harper Center Suite 448

Babies who are born with low birth weight (under 2500g) face increased risks of a range of adverse conditions later in life. Thus, there is considerable interesting in identifying patients at risk for delivering a low birth weight baby, both to facilitate preemptive medical care such as improved nutrition, and also to simply foster medical vigilance.

As one might imagine, the distribution of birth weights varies across different populations of mothers. The goal of this project is to use a large publicly available data set to try and ascertain which populations are most likely to deliver low birth weight babies.

The National Bureau of Economic Research maintains detailed birth records for the United States. The data set under consideration consists of information on single-child, live births to mothers between the ages of 18 and 45 in the year 1997. The goal is to characterize the relationship between the distribution of birth weights and various measured attributes of the mother and child. Specifically, the included features are:

- `BirthWt`: weight of baby at birth in grams
- `Boy`: male baby
- `Married`: mothers marriage status
- `Black`: mother's race
- `Age`: mother age
- `HighSchool`: mother has high school diploma
- `SomeCollege`: mother has done college course work
- `College`: mother has college degree
- `NoPrenatal`: mother had no prenatal care
- `PrenatalSecond`: mother had prenatal care starting in second trimester
- `PrenatalThird`: mother had prenatal care starting in third trimester
- `NonSmoker`: mother is non-smoker
- `Cigarettes`: number of cigarettes smoked per day
- `Weightgain`: weight gained over course of pregnancy in pounds

This analysis is intentionally open ended. While you explore the data, recall the tools you have learned in class, including (but not limited to):

- frequency tables (sometimes called pivot tables)
- conditional probabilities and conditional expectations
- scatter plots
- density plots and histograms
- linear regression and prediction intervals
- hypothesis tests and confidence intervals
- simulation based permutation tests

The maximum length of your report should be fifteen pages, including plots and tables. The general format should be an *introduction* outlining the aspects you plan to explore and describing the data, the *analysis* itself, including plots, diagrams, and tables, and a *discussion/conclusion* section summarizing your findings.

In class of week nine we will work on this project, so be prepared with questions. In the meantime, here are the sorts of questions that you might pose and try to answer.

- For younger mothers, do race and education level appear to be associated with more or less low-weight births? To answer this question, look only at all mothers less than 19 years old and create a table breaking down the various combinations of race and education, examining the frequency of births less than 1500 grams. What patterns of dependence do you see?
- Based on the table you create above, is P(low birth weight | black and college educated) $<$ P(low birth weight | white and college educated)?
- If you subsample the data and redo the table from above, what is the distribution of P(low birth weight | black and college educated)$-$P(low birth weight | white and college educated). Does the histogram strengthen your trust in the pattern in the data or weaken it?
- Perform a linear regression using age, race and education as predictors. Based on the results, what is the predicted probability that a 19 year old white, college educated mother gives birth to a baby weighing less than 1500g? What is this probability for a 19 year old, college educated black mother? How does this result compare to the previous result based on the table?
- Run a linear regression including all of the provided variables as predictors of birth weight. Why can this analysis not be done easily using a table?
- Re-run the linear regression above, but excluding the variable which records the mother's weight gain during pregnancy. Does the coefficient for smoking change? Why might the latter regression be a more appropriate analysis? (Think about how smoking might affect birth weight.)

Good luck!

Marie Tang

Birth Data Analysis

Statistics 41000-83

Fall 2013

Overview

This paper analyzes a dataset of the birth weights (in grams) for all live births in the United States and the District of Columbia in the month of June 1997 based on characteristics of the birth mother as well as the gender of the baby. The mother's characteristics included in this dataset are:

1. **Marital Status**- Married or Unmarried.
2. **Race**- Black or Non-Black.
3. **Age**- Mother's age at the time of baby's birth. This dataset includes mothers age 18-45.
4. **Education Level**- Specifically whether the mother has a high school degree, some college education, or a college degree.
5. **Amount of Prenatal Care**- Defined by the time at which the mother started receiving prenatal care if at all. (No Prenatal if the mother did not receive any prenatal care during the pregnancy, PrenatalSecond if the mother began receiving prenatal care during the second trimester, PrenatalThird if the mother began receiving prenatal care during the third trimester and full prenatal care unless otherwise noted.)
6. **Smoking Status**- Smoker or Nonsmoker.
7. **Cigarettes**- Number of cigarettes the mother smoked each day.
8. **Weight Gain**- Mother's total weight gain during the pregnancy.

In total, there were 203,849 live births during this month. The goal of this paper is to identify risk factors for low birth weight which is defined as a birth weight of less than 2500 grams. The data used in this analysis was collected by the National Center for Health Statistics (NCHS) and is available in the National Vital Statistics Report which can be found at the following website: http://www.nber.org/natality/1997/docs/Nat1997doc.pdf. Additional characteristics are included in the full report, but for the purpose of this analysis, I will only be considering the factors listed above.

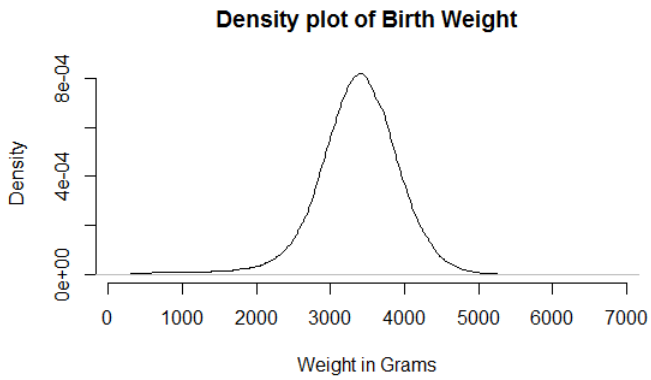
Data Summary

A basic review of the dataset shows that:

1. The average birth weight of babies in this dataset is 3,368.6 grams with a standard deviation of 572.2 grams and approximates a normal distribution.
2. Most mothers (72%) were married at the time they gave birth.
3. A small proportion of the mothers (16%) were Black.
4. The education levels varied, but 84% of mothers had at least a high school education.
5. A vast majority of mothers (99%) had some level of prenatal care.
6. A majority of mothers (87%) were non-smokers.

Summary Statistics

	var	n	mean	sd
BirthWt	1	203849	3368.57	572.24
Boy	2	203849	0.51	0.50
Married	3	203849	0.72	0.45
Black	4	203849	0.16	0.37
Age	5	203849	27.44	5.71
HighSchool	6	203849	0.35	0.48
SomeCollege	7	203849	0.24	0.43
College	8	203849	0.25	0.43
NoPrenatal	9	203849	0.01	0.09
PrenatalSecond	10	203849	0.13	0.33
PrenatalThird	11	203849	0.02	0.15
NonSmoker	12	203849	0.87	0.34
Cigarettes	13	203849	1.50	4.72
Weightgain	14	203849	30.69	12.89



Original Regression Model

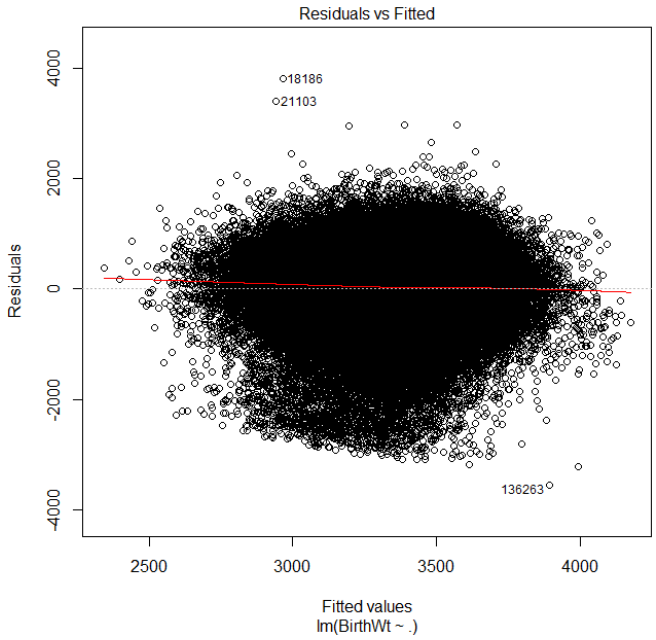
A linear regression of all the variables outlined in the Overview, shows that the three variables with the strongest impact on birth weight are Race, Age and Smoking Status. The effect of smoking is also further exacerbated by the number of cigarettes the mother smokes each day.

```
Call:
lm(formula = BirthWt ~ ., data = birthdata)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2711.11947    9.35805 289.710  < 2e-16 ***
Boy             108.94676    2.39536  45.482  < 2e-16 ***
Married          71.21675    3.17549  22.427  < 2e-16 ***
Black          -199.09087    3.54565 -56.151  < 2e-16 ***
Age               5.01447    0.24031  20.866  < 2e-16 ***
HighSchool       20.77620    3.72275   5.581 2.40e-08 ***
SomeCollege      41.76117    4.11494  10.149  < 2e-16 ***
College          47.53766    4.42203  10.750  < 2e-16 ***
NoPrenatal     -189.80064   13.43583 -14.126  < 2e-16 ***
PrenatalSecond   11.18814    3.74200   2.990 0.002791 **
PrenatalThird    30.10409    8.23880   3.654 0.000258 ***
NonSmoker       168.09716    6.24319  26.925  < 2e-16 ***
Cigarettes       -3.54497    0.44343  -7.994 1.31e-15 ***
Weightgain        8.93287    0.09356  95.476  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 540.2 on 203835 degrees of freedom
Multiple R-squared:  0.1088, Adjusted R-squared:  0.1087
F-statistic:  1913 on 13 and 203835 DF,  p-value: < 2.2e-16
```

Controlling for the other variables, the regression model shows:

1. The expected birth weight decreases by 199 grams if the mother is Black.
2. The expected birth weight increases by 5 grams for each additional year of the mother's age.
3. The expected birth weight is 168 grams lower if the mother is a smoker. In addition, the expected birth weight decreases by 3.5 grams for each additional cigarette the mother smokes on average each day.

A look at the residual plot of the fitted values shows that there are a few outliers, but nothing that would drastically change the equation.



Areas of Interest

This paper examines three primary areas of interest to identify risk factors for low birth weight babies. The primary areas of interest are:

1. **Race**- How do black mothers differ from non-black mothers? What other characteristics might help explain the lower birth weights from black mothers?
2. **Age-** While the original regression model indicates an expected increase in birth weight corresponding with an increase in age of the mother, is that still true for mothers over the age of 34? I would expect mothers over the age of 34 to have lower weight babies because that is the age at which pregnancies start being considered "high risk".
3. **Smoking-** Since smoking has such a large negative impact on the expected birth weights, what are some other characteristics of mothers that might help offset this impact? If a mother smokes only a few cigarettes a day, does smoking still have a large impact on the expected weight of the baby?

Analysis

Race

First, I subset the dataset into black mothers and non-black mothers, and ran regression models for each subset separately. In the original regression model, all of the variables were significant, but in the sub-setted regression models, age and high school education only appear to matter for the non-black mothers. The impact of age and high school education is no longer significant at the 95% significance level in the regression model for the black mothers' subset.

Black Mothers Subset

```
Call:
lm(formula = Birthwt ~ . - Black, data = blackds)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2620.7480    25.8236 101.487  < 2e-16 ***
Boy             104.3452     6.6512  15.688  < 2e-16 ***
Married          79.7039     7.9418  10.036  < 2e-16 ***
Age               1.2744     0.6519   1.955 0.050594 .
HighSchool       10.8558     9.2244   1.177 0.239265
SomeCollege      31.6161    10.4972   3.012 0.002599 **
College          44.5934    13.4709   3.310 0.000933 ***
NoPrenatal     -209.6615    24.1836  -8.670  < 2e-16 ***
PrenatalSecond   20.0651     8.4680   2.370 0.017817 *
PrenatalThird    51.2765    17.2437   2.974 0.002945 **
NonSmoker       140.0790    18.1869   7.702 1.38e-14 ***
Cigarettes       -4.5215     1.6630  -2.719 0.006554 **
Weightgain        9.5651     0.2350  40.709  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 600 on 32588 degrees of freedom
Multiple R-squared:  0.07709, Adjusted R-squared:  0.07676
```

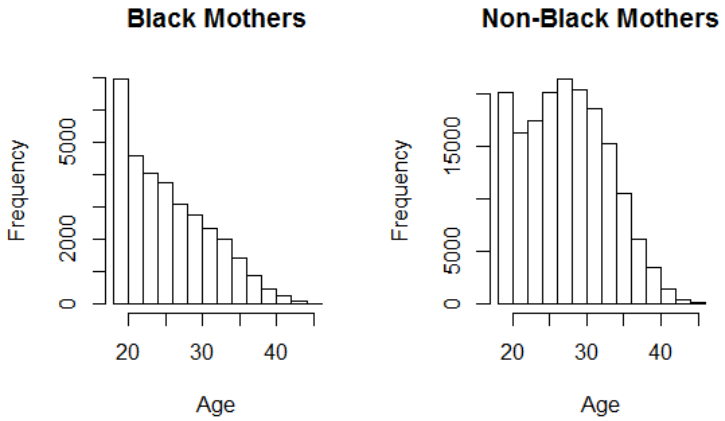
Non-Black Mothers Subset

```
Call:
lm(formula = Birthwt ~ . - Black, data = whiteds)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2692.0361    10.0034 269.112  < 2e-16 ***
Boy             109.8518     2.5544  43.005  < 2e-16 ***
Married          70.4721     3.4774  20.265  < 2e-16 ***
Age               5.7279     0.2583  22.178  < 2e-16 ***
HighSchool       23.1955     4.0784   5.687 1.29e-08 ***
SomeCollege      44.1037     4.4765   9.852  < 2e-16 ***
College          47.1735     4.7104  10.015  < 2e-16 ***
NoPrenatal     -174.7788    16.7659 -10.425  < 2e-16 ***
PrenatalSecond    8.6339     4.2010   2.055   0.0399 *
PrenatalThird    22.3495     9.5077   2.351   0.0187 *
NonSmoker       170.8926     6.6460  25.714  < 2e-16 ***
Cigarettes       -3.3787     0.4590  -7.361 1.83e-13 ***
Weightgain        8.7973     0.1020  86.242  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 528 on 171235 degrees of freedom
Multiple R-squared:  0.08711, Adjusted R-squared:  0.08705
```

Race and Age

I was interested in seeing why age was no longer significant for the black mothers subset so I first looked at the average age for the black mothers compared to the non-black mothers. I found that the average age of black mothers was slightly younger, but not by much (the average age for black mothers was 25.9 years old compared to the average age for non-black mothers which was 27.7 years old). By creating a histogram of the ages for each subset, I found that the ages of black mothers were heavily skewed on the lower end with the

highest proportion being under age 21, whereas the ages of non-black mothers were more evenly spread with the highest proportion being in the 26-30 age range. This made it even more surprising that age was no longer a statistically significant predictor variable for the black mothers considering they were disproportionately younger mothers.



Race and Prenatal Care

Next I wanted to look into the impact of prenatal care on each subset because I saw that the regression coefficients were very different between the models.

Black Mothers Subset

```
Call:
lm(formula = Birthwt ~ . - Black, data = blackds)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2620.7480    25.8236 101.487  < 2e-16 ***
Boy          104.3452     6.6512  15.688  < 2e-16 ***
Married       79.7039     7.9418  10.036  < 2e-16 ***
Age            1.2744     0.6519   1.955 0.050594 .
HighSchool    10.8558     9.2244   1.177 0.239265
SomeCollege   31.6161    10.4972   3.012 0.002599 **
College       44.5934    13.4709   3.310 0.000933 ***
NoPrenatal  -209.6615    24.1836  -8.670  < 2e-16 ***
PrenatalSecond 20.0651    8.4680   2.370 0.017817 *
PrenatalThird  51.2765   17.2437   2.974 0.002945 **
NonSmoker     140.0790    18.1869   7.702 1.38e-14 ***
Cigarettes     -4.5215     1.6630  -2.719 0.006554 **
Weightgain      9.5651     0.2350  40.709  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 600 on 32588 degrees of freedom
Multiple R-squared:  0.07709, Adjusted R-squared:  0.07676
```

Non-Black Mothers Subset

```
Call:
lm(formula = Birthwt ~ . - Black, data = whiteds)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2692.0361    10.0034 269.112  < 2e-16 ***
Boy          109.8518     2.5544  43.005  < 2e-16 ***
Married       70.4721     3.4774  20.265  < 2e-16 ***
Age            5.7279     0.2583  22.178  < 2e-16 ***
HighSchool    23.1955     4.0784   5.687 1.29e-08 ***
SomeCollege   44.1037     4.4765   9.852  < 2e-16 ***
College       47.1735     4.7104  10.015  < 2e-16 ***
NoPrenatal  -174.7788    16.7659 -10.425  < 2e-16 ***
PrenatalSecond  8.6339     4.2010   2.055   0.0399 *
PrenatalThird  22.3495     9.5077   2.351   0.0187 *
NonSmoker     170.8926     6.6460  25.714  < 2e-16 ***
Cigarettes     -3.3787     0.4590  -7.361 1.83e-13 ***
Weightgain      8.7973     0.1020  86.242  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 528 on 171235 degrees of freedom
Multiple R-squared:  0.08711, Adjusted R-squared:  0.08705
```

Prenatal care appeared to have a larger impact on the black mothers than the non-black mothers. The most notable impact was no prenatal care, decreasing the expected birth weight by 209.7 grams if the mother was black and by 174.8 grams if the mother was non-black. The table below shows that of the black mothers in our dataset that had no prenatal care, 25.6% of them had low birth weight babies, compared to 14.6% of the non-black mothers.

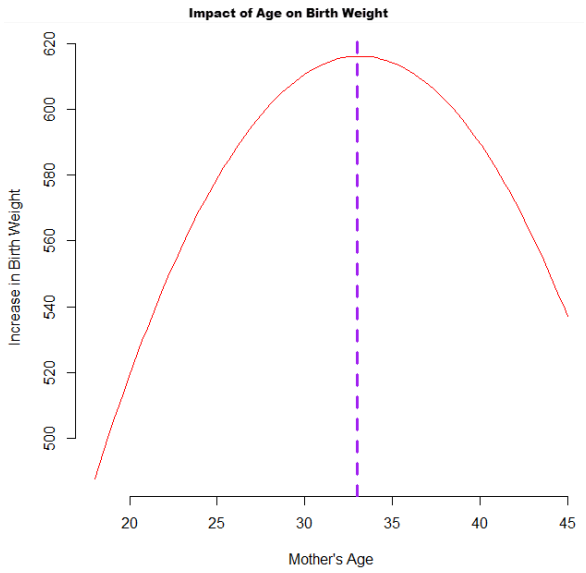
Black	NoPrenatal 0	1
FALSE	0.04665684	0.1462687
TRUE	0.10553000	0.2561728

Age

Next, I wanted to look at the effect of age on birth weight. The original regression model implied that expected birth weight consistently increases with the age of the mother. However, mothers over the age of 34 are considered to have "high risk" pregnancies so I wanted to see if mothers in that age group were expected to have lower birth weight babies. Transforming the age variable and including the age squared in the regression model showed that this is in fact true. The graph below shows that expected birth weight increases with age until 33 years and starts decreasing after that.

Linear Regression with Age Squared

```
Call:
lm(formula = BirthWt ~ . + I(Age^2), data = birthdata)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   2281.43556   27.81187  82.031  < 2e-16 ***
Boy            108.86628    2.39379  45.479  < 2e-16 ***
Married         61.75824    3.22536  19.148  < 2e-16 ***
Black         -199.56182    3.54343 -56.319  < 2e-16 ***
Age             37.19225    1.97612  18.821  < 2e-16 ***
HighSchool      15.26420    3.73544   4.086 4.38e-05 ***
SomeCollege     32.02372    4.15485   7.708 1.29e-14 ***
College         37.40915    4.46204   8.384  < 2e-16 ***
NoPrenatal    -189.00392   13.42709 -14.076  < 2e-16 ***
PrenatalSecond  13.43663    3.74205   3.591 0.000330 ***
PrenatalThird   31.70017    8.23396   3.850 0.000118 ***
NonSmoker      168.06176    6.23909  26.937  < 2e-16 ***
Cigarettes      -3.68468    0.44323  -8.313  < 2e-16 ***
Weightgain       8.97268    0.09353  95.932  < 2e-16 ***
I(Age^2)        -0.56126    0.03421 -16.405  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.9 on 203834 degrees of freedom
Multiple R-squared:  0.1099,    Adjusted R-squared:  0.1099
```



A logistic regression interacting mothers over the age of 34 years with all of the other variables, showed that a given mother increases her odds of having a low birth weight baby by 25% if she is over 34 years old.

Logistic Regression for Age > 34

```
Call:
glm(formula = BirthWt < 2500 ~ . - Age + I(Age > 34), family = binomial,
    data = birthdata)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.0705908  0.0497318 -21.527  < 2e-16 ***
Boy             -0.1437575  0.0195042  -7.371 1.70e-13 ***
Married         -0.3139610  0.0234878 -13.367  < 2e-16 ***
Black            0.7057567  0.0241110  29.271  < 2e-16 ***
HighSchool      -0.0864224  0.0268222  -3.222 0.001273 **
SomeCollege     -0.1560764  0.0308861  -5.053 4.34e-07 ***
College         -0.2772697  0.0349957  -7.923 2.32e-15 ***
NoPrenatal       0.7175752  0.0677506  10.591  < 2e-16 ***
PrenatalSecond  -0.1033782  0.0284650  -3.632 0.000281 ***
PrenatalThird   -0.4184101  0.0664068  -6.301 2.96e-10 ***
NonSmoker       -0.4919791  0.0401889 -12.242  < 2e-16 ***
Cigarettes       0.0122690  0.0026662   4.602 4.19e-06 ***
Weightgain      -0.0405018  0.0008147 -49.714  < 2e-16 ***
I(Age > 34)TRUE  0.2260230  0.0293348   7.705 1.31e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 89184  on 203848  degrees of freedom
Residual deviance: 83116  on 203835  degrees of freedom
AIC: 83144

Number of Fisher Scoring iterations: 6
```

```
> exp(0.2260230)
[1] 1.253604
```

In our dataset, there were 24,973 mothers over the age of 34. Surprisingly, only 6.3% of them had low birth weight babies compared to 5.6% of the mothers aged 34 and under, indicating that these mothers tend to have other characteristics, such as more prenatal care and higher education levels that help offset the risk of low birth weight babies.

```
> highriskweights<-tapply(BirthWt<2500,list(Age>34),mean)
> highriskweights
     FALSE       TRUE
0.05627362 0.06254755
```

Smoking

Lastly, I wanted to look at smoking because it had such a large impact in the original regression model. A logistic regression interacting the Non Smoker variable showed that smoking increases the odds of a given mother having a low birth weight baby by 89%.

```
Call:
glm(formula = BirthWt < 2500 ~ . - NonSmoker - Cigarettes + I(NonSmoker ==
    0), family = binomial, data = birthdata)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.6926758  0.0571029 -29.643  < 2e-16 ***
Boy                  -0.1428828  0.0194993  -7.328 2.34e-13 ***
Married              -0.3191732  0.0241716 -13.204  < 2e-16 ***
Black                 0.6974203  0.0240343  29.018  < 2e-16 ***
Age                   0.0063607  0.0018906   3.364 0.000767 ***
HighSchool           -0.0921874  0.0269642  -3.419 0.000629 ***
SomeCollege          -0.1650096  0.0313322  -5.266 1.39e-07 ***
College              -0.2821903  0.0362006  -7.795 6.43e-15 ***
NoPrenatal            0.7292668  0.0676590  10.779  < 2e-16 ***
PrenatalSecond       -0.1010603  0.0284779  -3.549 0.000387 ***
PrenatalThird        -0.4163178  0.0663900  -6.271 3.59e-10 ***
Weightgain           -0.0406509  0.0008152 -49.865  < 2e-16 ***
I(NonSmoker == 0)TRUE  0.6346430  0.0249654  25.421  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 89184  on 203848  degrees of freedom
Residual deviance: 83184  on 203836  degrees of freedom
AIC: 83210

Number of Fisher Scoring iterations: 6

> exp(0.6346430)
[1] 1.886349
```

I divided the dataset into Smokers and Non-Smokers so I could see how these two groups differed with respect to the other variables and ran linear regression models for each subset. The regression summaries below show that the coefficients for age, education level and no prenatal care are very different for Smokers versus Non-Smokers. In addition, the variables indicating prenatal care starting in the second and third trimesters are no longer statistically significant for the Smoker Subset.

Non Smoker Subset

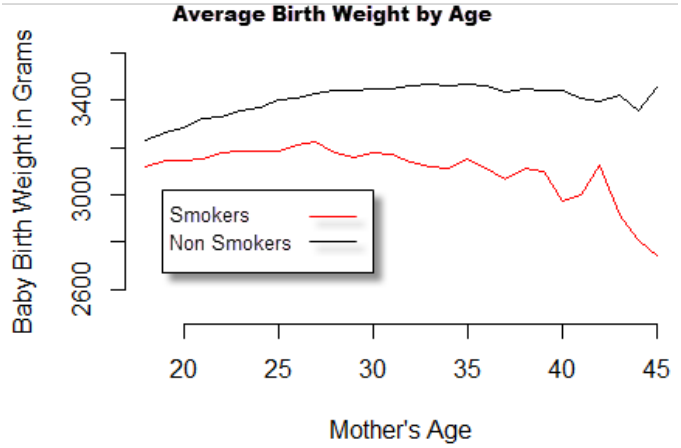
```
Call:
lm(formula = Birthwt ~ . - NonSmoker - Cigarettes, data = birthdata,
    subset = NonSmoker == 1)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2851.4480     8.1940 347.993  < 2e-16 ***
Boy             109.3258     2.5660  42.605  < 2e-16 ***
Married          79.2317     3.5388  22.389  < 2e-16 ***
Black          -196.6071     3.7633 -52.243  < 2e-16 ***
Age               6.0955     0.2603  23.413  < 2e-16 ***
HighSchool       14.5728     4.2341   3.442 0.000578 ***
SomeCollege      32.7669     4.5518   7.199 6.10e-13 ***
College          36.1540     4.7729   7.575 3.61e-14 ***
NoPrenatal     -155.9637    15.8992  -9.810  < 2e-16 ***
PrenatalSecond   16.1040     4.1386   3.891 9.98e-05 ***
PrenatalThird    32.5297     9.1733   3.546 0.000391 ***
Weightgain        8.8674     0.1018  87.100  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.4 on 177071 degrees of freedom
Multiple R-squared:  0.09087, Adjusted R-squared:  0.09082
```

Smoker Subset

```
Call:
lm(formula = Birthwt ~ . - NonSmoker - Cigarettes, data = birthdata,
    subset = NonSmoker == 0)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2842.1070    18.4666 153.905  < 2e-16 ***
Boy             105.1211     6.6544  15.797  < 2e-16 ***
Married          38.1339     7.2060   5.292 1.22e-07 ***
Black          -163.5952    10.9007 -15.008  < 2e-16 ***
Age              -2.4852     0.6332  -3.925 8.69e-05 ***
HighSchool       58.4641     7.9255   7.377 1.67e-13 ***
SomeCollege     103.7445    10.3111  10.061  < 2e-16 ***
College         102.6358    17.6270   5.823 5.86e-09 ***
NoPrenatal     -277.5620    25.3718 -10.940  < 2e-16 ***
PrenatalSecond  -11.8401     8.7817  -1.348    0.178
PrenatalThird    20.5590    18.7836   1.095    0.274
Weightgain        9.4084     0.2364  39.807  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 543.7 on 26754 degrees of freedom
Multiple R-squared:  0.09437, Adjusted R-squared:  0.09399
```

Smoking and Age

The regression summaries above show that in the non-smoker subset, the baby's expected weight increases by 6.1 grams for each additional year of the mother's age. However, if the mother is a smoker, the expected weight of the baby actually decreases by 2.5 grams for each additional year of the mother's age. I suspect that smoking may be causing other health issues with mothers as they get older and that those other health issues are adversely affecting the birth weights of their babies. However, further research would need to be conducted to confirm this.



Smoking and Education

The regressions also show that the amount of education a mother has matters a lot more for the Smokers than for the Non-Smokers. In comparing the linear models for each subset, the regression coefficients are 3-4 times higher at each education level for the Smokers than for the Non-Smokers. A high school education is expected to impact the Non-Smokers by increasing the birth weight of the baby by only 14.6 grams, whereas a high school education is expected to impact the Smokers by increasing the birth weight of the baby by 58.5 grams. Having some college education increases the expected birth weight by 32.8 grams for Non-Smokers, but by a whopping 103.7 grams for Smokers. A full College education also has a similar effect, increasing birth weight by 36.2 grams and 102.6 grams respectively for the Non-Smokers vs. the Smokers.

Based on this information, I assumed that smoking was related to education level and that the more education a mother had, the less likely she was to be a smoker. To test that assumption, I looked at the probabilities of being a smoker at every education level. The results showed that 26.0% of mothers with less than a high school education were smokers, 17.4% of mothers with only a high school degree were smokers, 9.9% of mothers with only some college education were smokers, and only 2.3% of mothers with a college degree were smokers. It was interesting to see that the percentage of mothers who were smokers decreased as education level increased, however we cannot tell if education causes mothers to not smoke or if mothers who get more education tend to not be smokers.

Smoking and Prenatal Care

The regression models also show that having no prenatal care during the pregnancy has a much larger impact on the Smokers than the Non-Smokers. For the Smokers, no prenatal care is associated with a 277.6 gram decrease in the baby's expected birth weight compared to a 156.0 gram decrease for the Non-Smokers. In

addition, prenatal care starting in the second or third trimesters no longer has a statistically significant impact for the Smokers, further indicating the extreme importance for mothers who are smokers to start prenatal care in the first trimester. The WebMD website has the following advice for mothers regarding prenatal care,

> **"Get early and regular prenatal care.** The first eight weeks of your pregnancy are very important to your baby's development. Early and regular prenatal care can increase your chances of having a safe pregnancy and a healthy baby. Prenatal care includes screenings, regular exams, pregnancy and childbirth education, and counseling and support." (http://www.webmd.com/baby/guide/pregnancy-after-35)

Based on this information, prenatal care starting in the first trimester probably has such a large positive impact on smoker mothers because it helps detect other associated health issues early on and it helps educate mothers on ways to increase their odds of having healthy weight babies.

The logistic regression below, looking only at mothers that are smokers, shows that no prenatal care further increases the odds of having a low birth weight baby by 16%.

Logistic Regression for Smokers Subset

```
Call:
glm(formula = BirthWt < 2500 ~ . - NonSmoker - Cigarettes, data = birthdata
    subset = NonSmoker == 0)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.1497627  0.0100138  14.956  < 2e-16 ***
Boy            -0.0200476  0.0036085  -5.556 2.79e-08 ***
Married        -0.0161925  0.0039076  -4.144 3.43e-05 ***
Black           0.0571184  0.0059111   9.663  < 2e-16 ***
Age             0.0026077  0.0003433   7.595 3.18e-14 ***
HighSchool     -0.0234040  0.0042977  -5.446 5.21e-08 ***
SomeCollege    -0.0392872  0.0055914  -7.026 2.17e-12 ***
College        -0.0352443  0.0095586  -3.687 0.000227 ***
NoPrenatal      0.1475470  0.0137583  10.724  < 2e-16 ***
PrenatalSecond -0.0001627  0.0047620  -0.034 0.972738
PrenatalThird  -0.0287213  0.0101857  -2.820 0.004810 **
Weightgain     -0.0029069  0.0001282 -22.681  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.08692702)

    Null deviance: 2419.7  on 26765  degrees of freedom
Residual deviance: 2325.6  on 26754  degrees of freedom
AIC: 10592

Number of Fisher Scoring iterations: 2

> exp(.1475470)
[1] 1.158988
```

Smoking and Number of Cigarettes

Finally, I wanted to further investigate how the number of cigarettes smoked each day impacts the expected birth weight of the baby. The original regression model indicated that the expected birth weight is 168 grams lower for smokers and that the expected birth weight decreases by 3.5 grams for each additional cigarette the mother smokes a day. However, I wanted to investigate the impact of smoking without the classification of "smoker" or "nonsmoker" to see if there was a significant impact on expected birth weight if the mother only smoked 1-3 cigarettes a day.

In order to look at how smoking 1-3 cigarettes a day affects the birth weight of babies, I created a new variable called CigBuckets that returns a "0" if the mother smokes 0 cigarettes a day, a "1" if the mother smokes 1-3 cigarettes a day, a "2" if the mother smokes 4-20 cigarettes a day, and a "3" if the mother smokes over 20 cigarettes a day. I ran a linear regression model taking out the Non Smoker variable and included the new CigBuckets variable with primary interest in the regression coefficient for the group of mothers who smoked 1-3 cigarettes a day. The summary below shows that even "minimal" smoking (1-3 cigarettes a day) does have a big impact on the expected birth weight of the baby, decreasing the expected birth weight by 148 grams.

Linear Regression including CigBuckets

```
Call:
lm(formula = Birthwt ~ . - NonSmoker - Cigarettes - CigBuckets +
    as.factor(CigBuckets), data = birthdata)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             2879.46562    7.50304 383.773  < 2e-16 ***
Boy                      108.90780    2.39541  45.465  < 2e-16 ***
Married                   71.29776    3.17557  22.452  < 2e-16 ***
Black                   -198.81828    3.54481 -56.087  < 2e-16 ***
Age                        4.97353    0.24024  20.702  < 2e-16 ***
HighSchool                21.37555    3.72223   5.743 9.33e-09 ***
SomeCollege               42.43968    4.11358  10.317  < 2e-16 ***
College                   48.28702    4.42023  10.924  < 2e-16 ***
NoPrenatal              -190.70020   13.43627 -14.193  < 2e-16 ***
PrenatalSecond            11.16233    3.74212   2.983  0.00286 **
PrenatalThird             29.93173    8.23915   3.633  0.00028 ***
Weightgain                 8.94028    0.09355  95.572  < 2e-16 ***
as.factor(CigBuckets)1  -148.09376    9.32139 -15.888  < 2e-16 ***
as.factor(CigBuckets)2  -215.95498    3.99391 -54.071  < 2e-16 ***
as.factor(CigBuckets)3  -248.10235   16.45849 -15.074  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 540.3 on 203834 degrees of freedom
Multiple R-squared:  0.1087, Adjusted R-squared:  0.1087
```

Conclusion

Through this analysis, I found that the population for black mothers is very different from non-black mothers. Black mothers tend to be younger and have a disproportionately higher percentage of low birth weight babies. The probability of having a low birth weight baby is further impacted by the level of prenatal care, and of the black mothers who had no prenatal care, a high percentage of them (25.6%) had low birth weight babies. While this dataset did not include variables for socioeconomic status, further research on the socioeconomic status for black mothers versus non-black mothers would be interesting to help investigate some of the disparate findings.

Looking at age as a predictor variable, I found that the expected birth weight increases as mothers get older, but only up to age 33. Beyond that, the expected birth weight decreases with age. A mother increases her odds of having a low birth weight baby by 25% if she is over 34 years old, yet only 6.3% of mothers over the age of 34 had low birth weight babies. As such, it appears that mothers in this age group are doing other things to help reduce their risk of having low birth weight babies. Further analysis could be done to look at the other characteristics of mothers in this age group.

Lastly, I looked at smoking as a predictor variable and found that smoking is a huge risk factor for low birth weight babies. A given mother increases her chances of having a low birth weight baby by 89% if she is a smoker and even smoking just 1-3 cigarettes a day has a large impact on the expected birth weight.

Linear regression models looking at Smokers and Non Smokers separately showed significant differences in the predictor variables for each subset. Most notable were age, education level and prenatal care. For Smokers, age had a negative impact on the expected birth weight, even when considering the very young mothers. An area for further research would be to see if smoking is causing other health issues as mothers get older and if those other health issues are negatively impacting birth weight. In addition, education and no prenatal care

had much larger impacts on the expected birth weight for smokers, but we can't tell for certain why. It's possible that mothers who are smokers tend to have other unhealthy habits such as drug and alcohol use, but that education and prenatal care reduces the incidence of those other unhealthy habits that can affect birth weights. Further research on how education and prenatal care impacts Smokers, especially how it changes their behaviors, would be interesting to see.