Applied Statistical Modeling & Inference: Project #1

Due on April 30, 2015 at 5:00pm

Professor Ying Lu & Professor Daphna Harel

Sriniketh Vijayaraghavan

Introduction

Low birth weight is a major public health problem in the United States and other countries. It contributes substancially to both infant mortality and childhood handicap. We know that poverty is strongly consistently associated with low birth weight, but the social and environmental conditions that produce preterm delivery have not been elucidated.

We also know that premature birth is one of the strongest reasons for a baby being born. Fetal growth restriction also is another main cause for low birth weight. Low birth weight is the primary cause of birth defects and infections in babies. So, we can establish that determining the cause for low birth weight in babies is very important.

Low birth weight is currently the leading cause of infant mortality. There are many risk factors of low birth weight and/or pre-term birth. These risk factors often include infection, history of pre-term birth, diabetes, hypertension, women who are pregnant with twins, triplets, or more, and women with cervical and/or uterine irregularities. Many of these causes are medical complications that cannot be managed but there are behavioral risk factors that women can control. Such factors include smoking tobacco products, drinking alcohol, using illicit drugs, late or no prenatal care, and obesity. Although obesity is not a direct cause of pre-term birth and low birth weight, obesity does increase rates of medical complications such as diabetes and hypertension which can lead to low birth weight[Tiffany Pelletier].

In this assignment, we try and trace the causes of low birth weight by analysing a sample of 200 mothers.

Data and Descripive Statistics

Variable	Sample Mean	Sample Standard Deviation	Sample Median	Min	Max
MOTHAGE	23.2381	5.2986	23	14	45
MOTHWT	129.8148	30.5793	121	80	250
PREM	0.1957	0.4933	0	0	3
PHYSVIS	0.79365	1.0592	0	0	6

Figure 1: Data descriptions for the continuous variables

In Figure 1, we see that the data is relatively normal. It is not skewed by having an excess of premature births or overweight women. This allows us to determine the cause for low birth weight in a normal sample population.

In Figure 2, we see that categorical variables and their frequency of ocurrence and also their proportion in

Variable	0	1	2	3
RACE	NA	96/0.5079	26/0.1375	67/0.3544
SMOKE	115/0.6084	74/0.3915	NA	NA
HYPER	177/0.936	12/0.0634	NA	NA
URINIRR	161/0.851	28/0.1481	NA	NA

Figure 2: The number of items in each category and their proportions

the sample population. We notice that a large number of women here suffer from hypertension and urinary irritation. This may skew the sample a little bit, but these symptoms are not uncommon in any type of

pregnant women. There is an equal distribution of smokers and non smokers while more than half of the sample contains data from white women.

Now, we plot some scatterplots of the continuous variables with birth weight to determine if we can see visually if there is any strong correlation.

In Figure 3, we see the mother's weight and her age plotted against birth weight of the child. We see that the sample seems like a normal distribution with a few outliers on the higher end of the age and weight group. We also notice that a large sample of the population is clustered towards the lower end. We also see that very low birth weights are ocurring in both lighter and older women. If we plot mother's age versus birth weight, we notice a positive trend, which seems non-intuitive since most of the outliers in birth weight occur in the cases of older women. This is the same for the mother's weight.

Next, we plot a histogram of the child's birth weight to get an idea of the kind of distribution. We see from Figure 3, that it is a relatively normal distribution. We have also fit a normal distributio curve along with the histogram of the birth weight with probabilities shown. This can be seen below.

We also plot a scatter plot matrix to show the correlation between the continuous variables and the birth weight of the baby. This plot (Figure 5), shows us that the variables with the highest correlation are mother's weight with the birth weight, mother's age with mother's weight and mother's age to the number of physician visits. We also see a lesser correlation between birth weight and mother's age and physician visits. The least correlation is the one between birth weight and the number of months that a baby is premature.

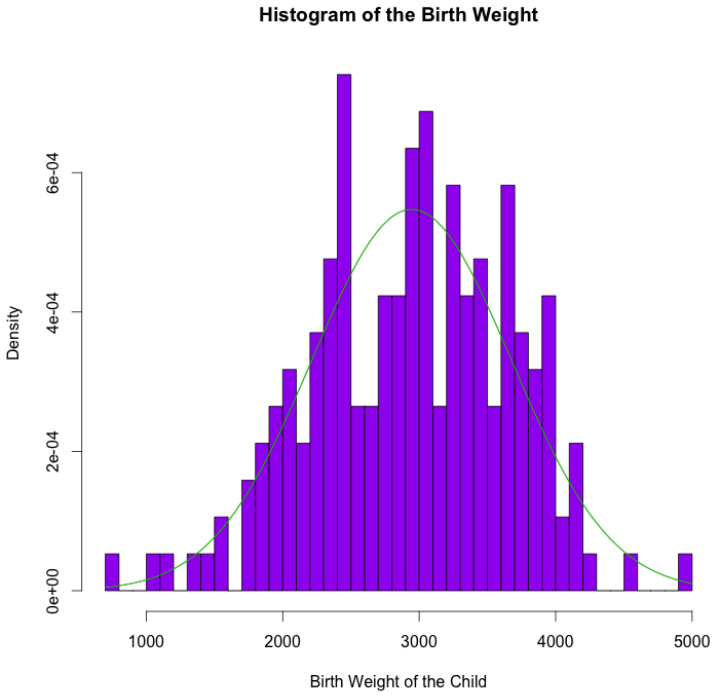

Figure 3: Birth weight with normal curve fitted line

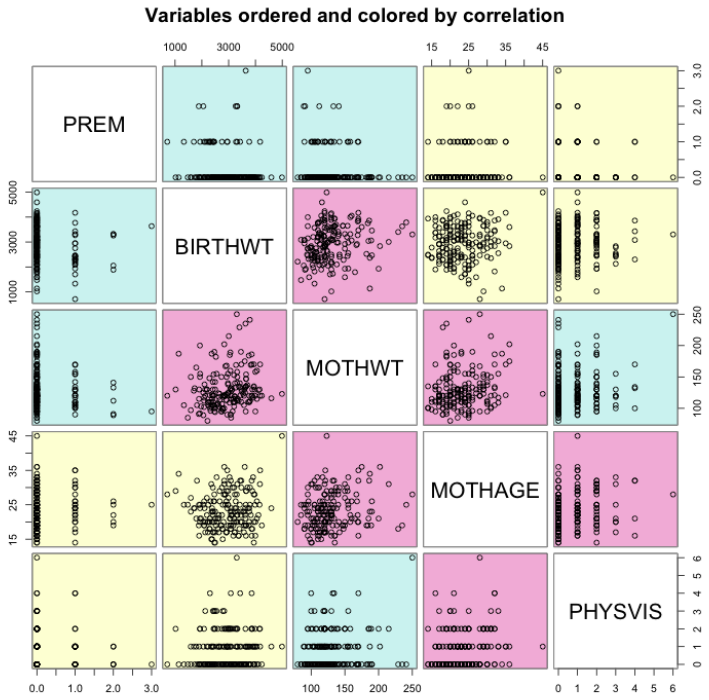

Figure 4: A scatterplot matrix of all the continuous variables

Finally, we plot box plots for all of the categorical variables to see their distributions. Here we also take PREM and PHYSVIS as categorical to see their distributions. This is done as we use PREM and PHYSVIS as both continuous and categorical in the following linear and logistic regressions.

Hypothesis Testing

In this subsection, we perform a hypothesis test to determine whether there is indeed a difference in birth weight of babies based on whether or not the mother was a smoker. First we separate the birth weight data into 2 separate vectors based on whether the smoking status of the mother was 0 or 1. Then, we make an assumption for the null hypothesis. We assume that the mean of the birth weights of babies whose mothers were smokers was the same as the mean of the birth weight of babies whose mothers were not smokers.

First, we try and determine the type of distribution that the data follows for both the smokers and non smokers. This picture can be seen on the next page in Figure 6.

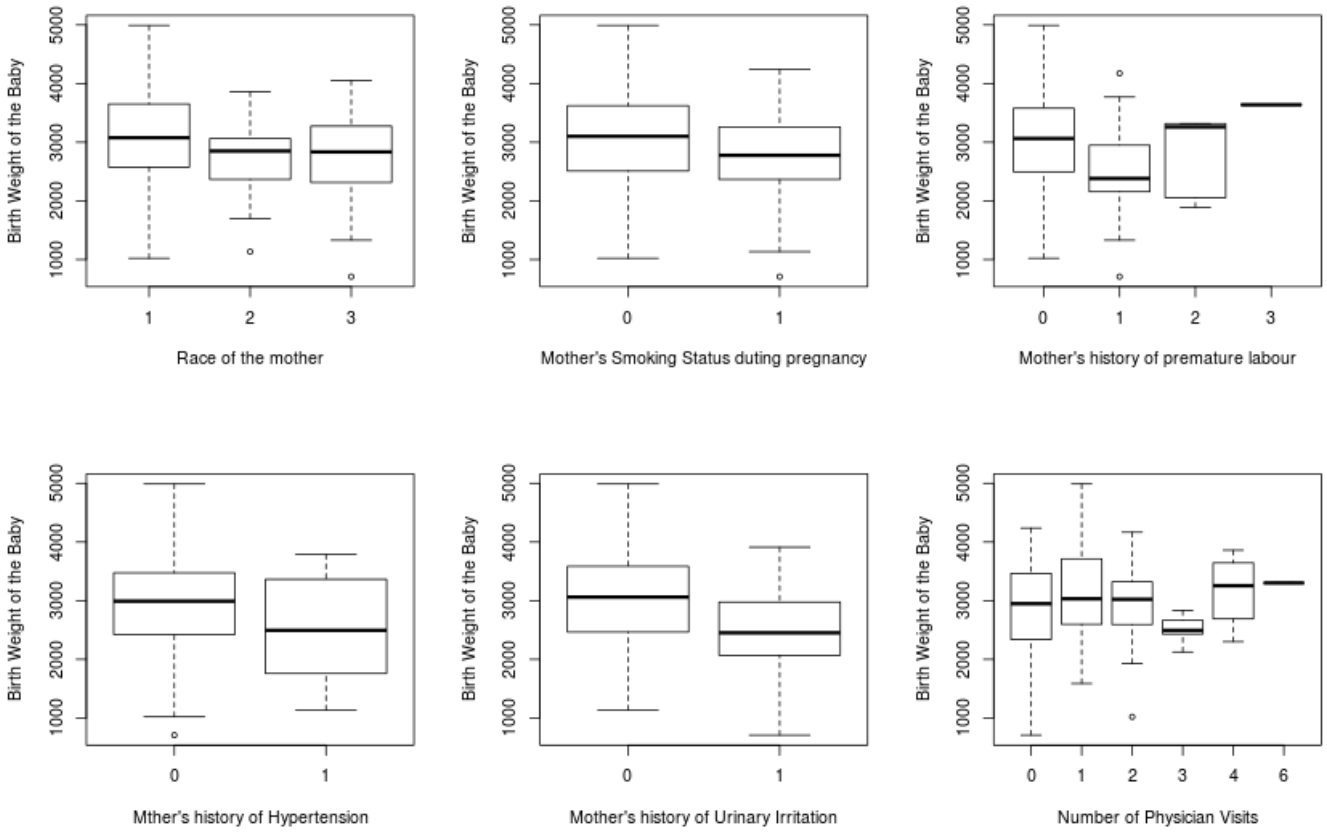

Figure 5: Box Plots of all the Categorical variables

Now, we see that both the data falls under a normal distribution. So this can be used as our sampling distribution for hypothesis testing. We implement the two tailed test to see if the Null Hypothesis holds. We see that the Null hypothesis does not hold as there is some difference in the data. The p-value we get for the Welch Two Sample t-test is 0.00743. This means that we can say that there is a strong evidence that smoking among pregnant mothers either positively or negatively affects birth weight among their children. If we try to fit the lower tail test. We see that it does not hold and gives a p-value close to 1.

This means that the mean of the distribution of smokers must be higher than non-smokers. This is proved true when we try the upper-tailed test. Here we see that the p-value is 0.003715, which is the lowest we have gotten. This gives us sufficient evidence that birth weight among smokers is definitely lower than that among non-smokers.

We also plot a box plot to see the distribution of smokers and non-smokers in the sample.

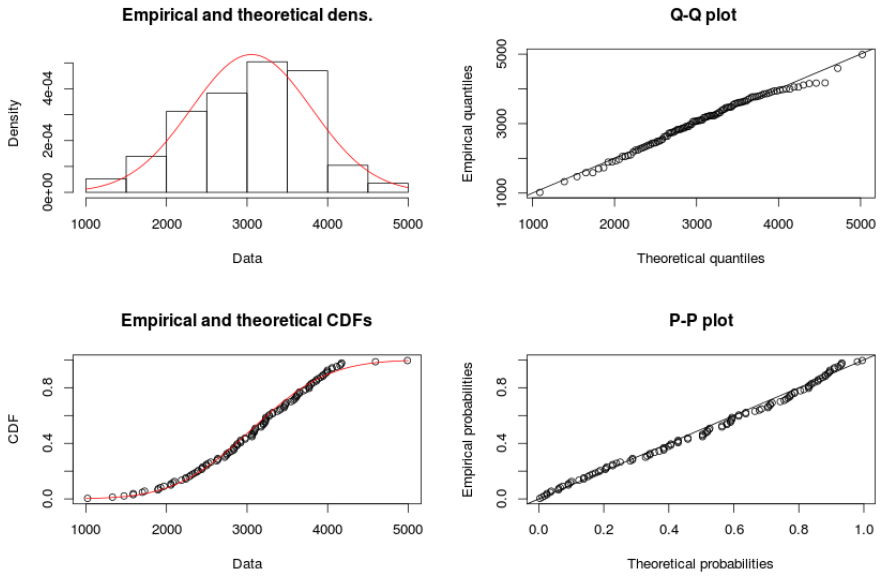

Figure 6: Fitting a normal distribution to the non-smokers birth weight data

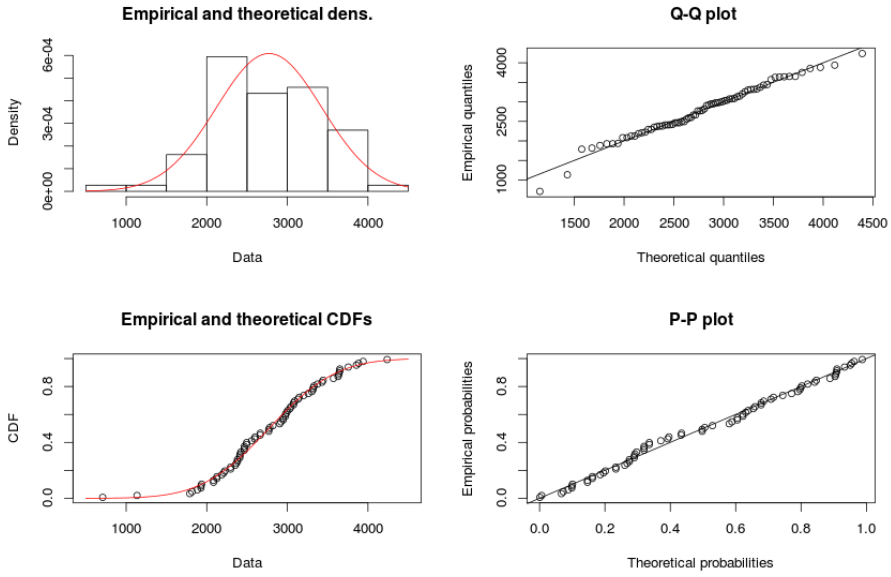

Figure 7: Fitting a normal distribution to the smokers birth weight data

Linear Regression Analysis

In this section, we fit the linear regression between the given variables and the birth weight of the babies. We start off with a simple linear fit between all the additive variables and get an R^2 value of 0.22. First, we treat PREM and PHYSVIS as quantitative. In this case, when we try to determine the best model through a stepwise AIC, we notice that neither PREM nor PHYSVIS make the cut. In the categorical case, PREM makes the cut but PHYSVIS does not.

This makes intuitive sense as we know that as a quantitative variable, the number of months that a baby is

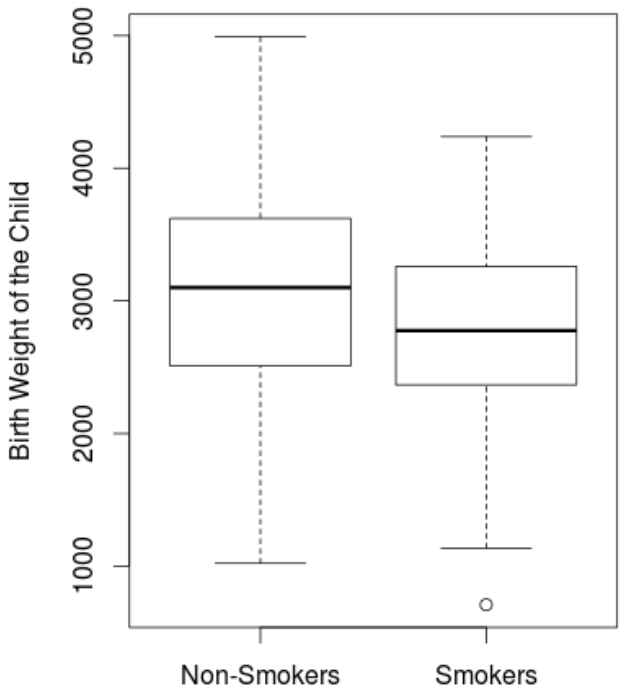

Figure 8: Birth weight of babies of smoker and non-smoker mothers

premature does not affect the model much, but as a categorical variable it is hugely impactful. This makes it an important predictor for low birth weight. PHYSVIS on the other hand has 6-7 categories and makes no sense to include it as a categorical variable. This feature adds more value as a quantitative variable since this allows us to get more information from it. Either way, we see that PREM is much more useful as a categorical variable and PHYSVIS is better as quantitative.

So, for the rest of the analysis, we assume that PREM is categorical and PHYSVIS is quantitative.

The final formula that we get for an additive model with the above assumption after running a stepwise AIC is as follows:

BIRTHWT∼MOTHWT+factor(RACE)+factor(SMOKE)+factor(PREM)+factor(HYPER)+factor(URINIRR)

For this model, we get an AIC value of 2988.345, a BIC of 3024.004, an R^2 value of 0.2746 and an adjusted R^2 value of 0.2381.

Next, we try to fit the model and select the best one while taking into account all additive relationships and two-way interactions between SMOKE and every other variable. In this section, we conduct a stepwise AIC to try and determine the formula which would provide us with the least AIC and BIC without losing much information about the model.

Our results here are as follows:

Step: AIC=2448.06
BIC=3028.563
formula: BIRTH ~ MOTHAGE + MOTHWT + RACE + SMOKE + PREM + HYPER + URINIRR + SMOKE * MOTHAGE
Multiple R^2: 0.297 Adjusted R^2: 0.2533

This makes intuitive sense. We see from the coefficients of mother's age and smoking and we get this relationship.

$$MOTHWT = MOTHWT[14.2312 - 44.6560 * SMOKE]$$

if the mother is a smoker, then we see that the older the mother is, the lesser the birth weight of the baby will be. The summary and confidence intervals of the model can be seen in Figure 9 below.

```
Call:
lm(formula = BIRTHWT ~ MOTHAGE + MOTHWT + factor(RACE) + factor(SMOKE) +
    factor(PREM) + factor(HYPER) + factor(URINIRR) + factor(SMOKE) *
    MOTHAGE, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1835.8  -414.4    60.5   451.9  1338.9

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)            2481.44374  347.66560   7.137 2.36e-11 ***
MOTHAGE                  14.23124   11.59077   1.228   0.2211
MOTHWT                    4.30251    1.67403   2.570   0.0110 *
factor(RACE)2          -381.38362  149.41908  -2.552   0.0115 *
factor(RACE)3          -277.39195  113.12893  -2.452   0.0152 *
factor(SMOKE)1          712.46069  454.14214   1.569   0.1185
factor(PREM)1          -272.62058  144.39308  -1.888   0.0607 .
factor(PREM)2             0.09624  290.73754   0.000   0.9997
factor(PREM)3          1391.70787  650.83384   2.138   0.0339 *
factor(HYPER)1         -587.85065  195.31604  -3.010   0.0030 **
factor(URINIRR)1       -596.73031  136.91167  -4.359 2.22e-05 ***
MOTHAGE:factor(SMOKE)1  -44.65604   18.91191  -2.361   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 630 on 177 degrees of freedom
Multiple R-squared:  0.297,     Adjusted R-squared:  0.2533
F-statistic: 6.798 on 11 and 177 DF,  p-value: 1.733e-09
```

Figure 9: Summary of the final linear model

From the coefficients, we can see that higher the mother's age, then higher the birth weight of the baby will be, but the mother's age can be a negative factor if she smokes. We see that Black people and people of other races are much more prone to having babies with low birth weights, the factors leading to this cannot be identified from our given data.

We can see from the p-values that mother's age and mother's that have had 2 premature babies are not statistically significant. We also notice that urinary infection and hypertension are really significant in reducing the weight of the baby.

	2.5 %	97.5 %
(Intercept)	1795.3405594	3167.546930
MOTHAGE	-8.6426435	37.105114
MOTHWT	0.9988727	7.606141
factor(RACE)2	-676.2557910	-86.511458
factor(RACE)3	-500.6470540	-54.136846
factor(SMOKE)1	-183.7693961	1608.690766
factor(PREM)1	-557.5741385	12.332986
factor(PREM)2	-573.6618487	573.854319
factor(PREM)3	107.3151426	2676.100594
factor(HYPER)1	-973.2985063	-202.402803
factor(URINIRR)1	-866.9196374	-326.540976
MOTHAGE:factor(SMOKE)1	-81.9778775	-7.334206

Figure 10: Confidence intervals of the final linear model

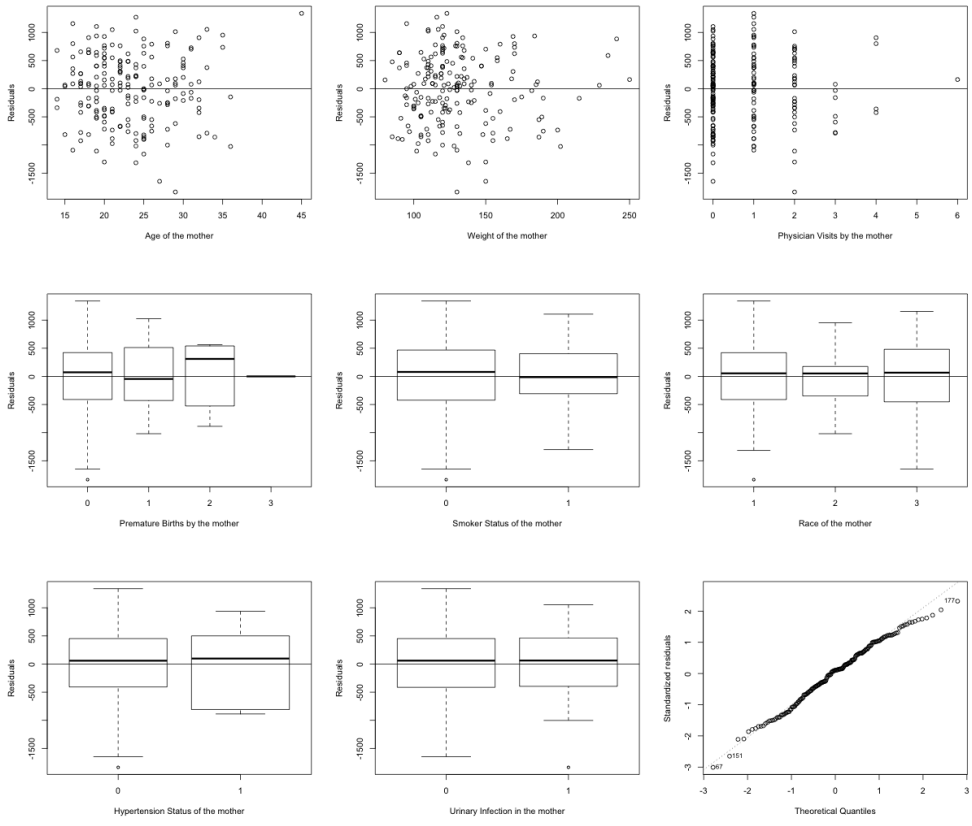

Figure 11: Residual Plots for the final model

We can see that the residuals are normally distributed about 0. We plot scatterplots between continuous variables and residuals and box plots for categorical variables.

Logistic Regression Analysis

Here, we convert all the birth weights below 2500g as 0 and everything above as 1. After conducting a step wise AIC, we see that the final formula is:

BIRTHWT∼MOTHWT+factor(RACE)+factor(SMOKE)+factor(PREM)+factor(HYPER)+factor(URINIRR)

We see that the AIC for the model is 213.5905 and the BIC is 246 as opposed to the initial values that are 228.20 and 296.28 respectively. We notice that none of the two way interactions are in the final formula. This means that none of the two way features were important enough to affect the birth weight of the child significantly. From the summary, we can see that factor(PREM)1, which means the women who have had one premature babies before have the most impact on the birth weight of the child. Women who have had 3 premature babies should not be included in the sample since it has an extremely high standard error and is not statistically significant.

Since 0 stands for low birth weight and 1 for high birth weight, we see that all factors such as RACE,

	OR	lower95ci	upper95ci	Pr(>|Z|)
MOTHWT	1.017321e+00	1.00322227	1.0316188	0.015875454
factor(RACE)2	2.868282e-01	0.10047582	0.8188081	0.019622876
factor(RACE)3	4.508111e-01	0.18758657	1.0833965	0.074926985
factor(SMOKE)1	4.125603e-01	0.18493317	0.9203649	0.030566688
factor(PREM)1	2.327320e-01	0.08608968	0.6291598	0.004063663
factor(PREM)2	7.604463e-01	0.11123516	5.1987032	0.780074972
factor(PREM)3	2.532111e+06	0.00000000	Inf	0.986673465
factor(HYPER)1	1.498371e-01	0.03671575	0.6114860	0.008158136
factor(URINIRR)1	4.089325e-01	0.16288674	1.0266385	0.056911980

Figure 12: Odds ratio and the 95% confidence intervals for logistic regression final model

SMOKE, PREM,HYPER and URINIRR lead to low birth weight. MOTHWT is slightly higher and has a value close to 1, which means that its effect is almost negligible on birth weight. Since we do not have any interaction effects in our final model, we cannot say what the effects are. But in our initial model we found that SMOKE with an interaction with HYPER and PREM leads to low birth weight among babies. Both MOTHAGE and MOTHWT have minimum effect with smoking on birth weight, according to the Odds Ratio. And Other races actually benifit from smoking and it seems to give rise to heavier babies, as does having urinary infections and higher PHYSVIS.

We also plot an ROC curve to determine the tradeoff between Sensitivity and Specificity. The area under the curve that we get is 0.7681. This means that our test is reasonably accurate.

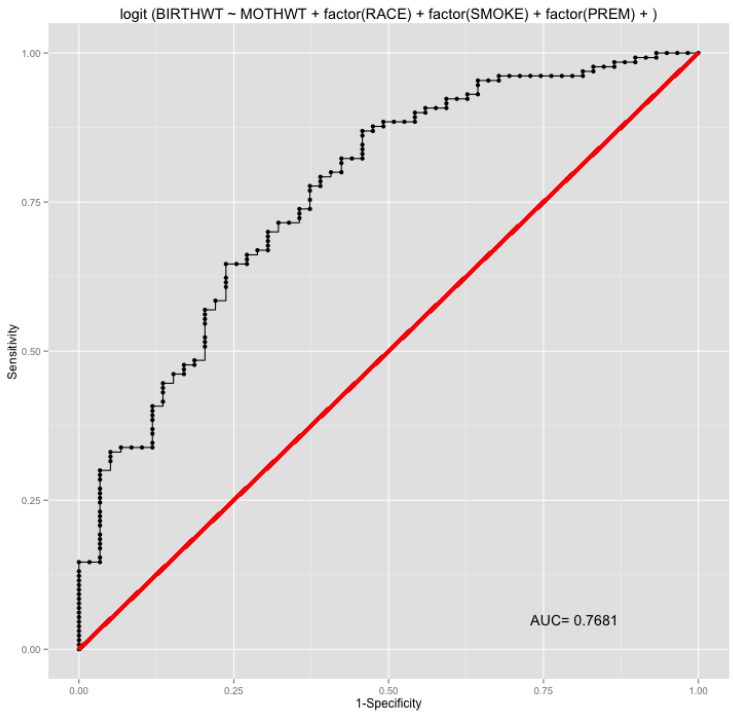

Figure 13: ROC Curve

Conclusion

We have used the birthwt data to fit, model and estimate which features would be the most informative. We know that birth weight is a serious problem that is faced, primarily in the US and we have tried to estimate the cause of it through this exercise. We have been able to come up with a succint analysis and correlated findings.

In the hypothesis testing section, we realized that birth weights in babies whose mothers smoke are significantly lower than those who don't. This hypothesis was further proven when we conducted our linear regression analysis. We determined that both smoking and smoking in older women are major causes to lower birth weight. We also found that women who have already had premature children are more prone to get babies with low birth weight.

In the logistic regression, we saw that hypertension and urinary infection in mothers were primary causes for babies to be predicted to be having low birth weight. We also saw that black people also have higher chances of having babies with lower birth weight. We see that mother's age has been removed as one of the predictors in Logistic Regression from Linear Regression.

We found that the Linear Regression and Logistic Regression model differ in the fact that Linear regression places a higher importance on a mother's weight and its correlation with smoking. This was not seen in the logistic model. The logistic model also shows that having urinary infections with smoking can lead to a more healthy baby. Some of these inferences don't neccesserily make sense, but this is understandeable since the sample size is pretty small. These problems may be eliminated by having a higher sample size.

We realize that this information is useful to try and circumvent certain problems which could be prevented or treated. We found that many of our intuitions were right and we determined to what extent were certain factors important to determining our end goal.