Statistical Methods for Bioinformatics

## Tutorial Group 1:

1. Sam Haberman r0727487

2. Michiel Van der Haegen r0372530

3. Nora Verplaetse s0201725

4. Ajay Sundaresan r0728544

5. Nnamdi Joseph Asouzu r0605462

6. Olawale Omosanya r0653440

# Introduction

The data in this report comes from a study at the Baystate Medical Center in Springfield, Massachusetts, that attempted to understand how behaviour of the mother before and during pregnancy affected the birth weight of the baby. What makes the birth weight of a baby so important is that if a baby has a low birth weight, defined as less than 2500 grams, it has a higher rate of infant mortality and birth defects. The data set contains variables hypothesised to be of clinical concern to physicians. The main variables are the weight of the mother at her last menstrual period (LWT), her race (RACE), her age in years (AGE) and the number of physician visits during the first trimester (FTV). Other variables deemed to be of importance are smoking status of the mother during pregnancy (SMOKE), history of premature labor (PTL), history of hypertension (HT) and presence of uterine irritability (UI). In total, 189 observations were made.

The outcome of a low birth weight baby, represented by the binary variable LOW was first modelled with logistic regression as a function of the four main variables, and a second time taking account of all eight variables. All data analysis was performed using RStudio, and the RACE variable, a categorical variable with three levels, was split into two dummy variables RACE.w and RACE.b.

# Analysis 1

Model Building

The first part of our analysis covers modelling the binary LOW variable as a logistic regression model as a function of the four most clinically relevant variables (LWT, RACE, AGE and FTV), through a model building exercise. To determine which variables and interactions were of importance, we subsequently used a statistical and logical approach. We opted for a Spearman approach as our data is not normally distributed, which calculated a critical r-value (r=0.143) based on the number of observations in the data set. All absolute r-values between two variables below this critical value were not considered for the models.

The correlation matrix listed below only had one significant main effect, LWT, but two significant interactions AGE with FTV and AGE with LWT. These significant interactions are highlighted in the below table. While in some cases one dummy of RACE was significant, both were never significant at the same time, leading to our decision to count the entire interaction as not significant.

LOW AGE LWT FTV RACE.w RACE.b

LOW 1.00000000 -0.08451021 -0.18169054 -0.0860172363 -0.1591516 0.0955912948

AGE -0.08451021 1.00000000 0.18606140 0.2341718477 0.1818312 -0.1416201455

LWT -0.18169054 0.18606140 1.00000000 0.0890236806 0.1228817 0.1724374005

FTV -0.08601724 0.23417185 0.08902368 1.0000000000 0.1423395 -0.0004628739

RACE.w -0.15915160 0.18183124 0.12288168 0.1423394848 1.0000000 -0.4057766142

RACE.b 0.09559129 -0.14162015 0.17243740 -0.0004628739 -0.4057766 1.0000000000

Using the statistically significant variables and interactions as a baseline we were able to come up with logical reasons these covariates should be either a part of or removed from the models we are testing. For the variable LWT, we were only presented with the weight of the mother, and not her height. This makes the weight more difficult to evaluate as we are not given information such as BMI which provides a more accurate assessment of the mother and their respective body composition. RACE.w and RACE.b were also under consideration as it is plausible that RACE does not have an impact on birth weight. The tested interactions of AGE with FTV and AGE with LWT make sense as a pregnancy at later age carries more risk, and thus would result in more check-ups by a physician. Secondly, the weight of a person varies with age.

|  |  |  |
| --- | --- | --- |
| **Model** | **AIC score** | **AIC weight** |
| LOW ~ LWT+RACE.w+RACE.b+AGE+FTV | 235.03 | 0.01 |
| LOW ~ LWT+RACE.w+RACE.b+AGE+FTV+AGE\*FTV | 227.24 | 0.29 |
| LOW ~ LWT+RACE.w+RACE.b+AGE+FTV+AGE\*LWT | 237.19 | 0.00 |
| LOW ~ LWT+AGE+FTV+AGE\*FTV | 225.70 | 0.62 |
| LOW ~ AGE+FTV+AGE\*FTV | 229.72 | 0.08 |

To start our model building process, we took the full model with our four significant variables and checked how the AIC score changed when we added the two significant interactions separately. Only the AGE\*FTV interaction improved the model, so we tested how this model changed when removing both RACE variables. The removal of the RACE dummy variables improved our model leading to the final step where we tested the removal of LWT, which did not improve the model.

The final model, selected based on its lowest AIC score, looks like this:

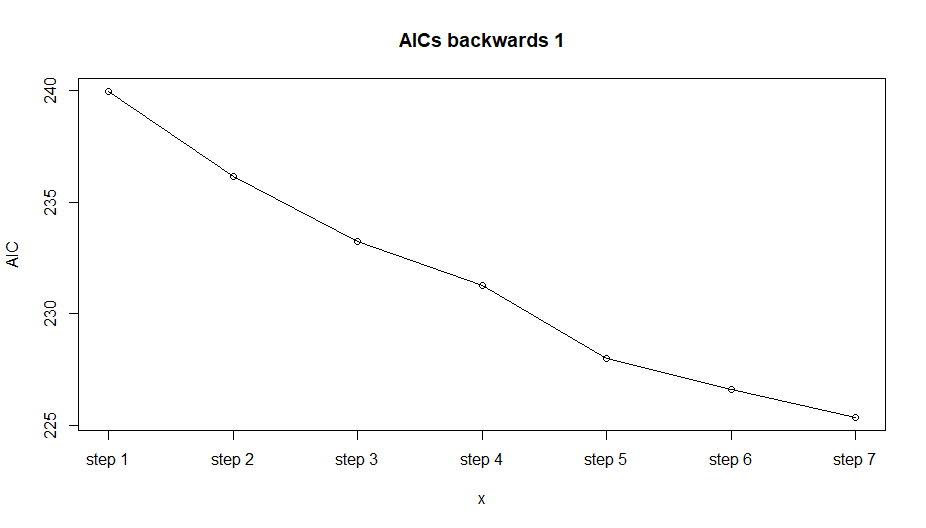
E(LOW) = - 0.165923 - 0.015526\*LWT + 0.065848\*AGE + 2.843866\*FTV - 0.130837\*AGE\*FTV

Backwards Elimination

The backwards elimination procedure was started from the full model with four variables and their respective interactions, as shown below.

E(LOW)=LWT+RACE.w+RACE.b+AGE+FTV+LWT\*RACE.w+LWT\*RACE.b+LWT\*AGE+LWT\*FTV+RACE.b\*AGE+RACE.w\*AGE+RACE.w\*FTV+RACE.b\*FTV+AGE\*FTV

A significance level of a = 0.10 was used to control the type I error at each stage. As a rule of thumb, the term with the highest p-value is removed. In cases where a main effect had the highest p-value but was still involved in an interaction, the interaction with highest p-value was deleted from the model. As RACE was split in two dummy variables, each time one of the dummy variables had the highest p-value, the other dummy variable was removed as well. At each step, the AIC score was checked, and a graph of the AICs is shown below.



After consecutive steps we were left with the following model.

E(LOW) = - 0.165923 - 0.015526\*LWT + 0.065848\*AGE + 2.843866\*FTV - 0.130837\*AGE\*FTV

After completing the backwards elimination process we arrived with the same model as in our model building process using AIC scores.

# Analysis 2

Model Building

The second part of our analysis covers modelling the binary variable LOW as a logistic regression model as a function of all eight variables (LWT, RACE, AGE, FTV, SMOKE, PTL, HT and UI). Once more, correlations between variables were inferred on both statistical and logical relevance.

The significant main effects were LWT, SMOKE, PTL, HT and UI, which are highlighted in the table below along with the interactions above the critical r-value, LWT with AGE, LWT with HT, LWT with UI, AGE with FTV, SMOKE with PTL and PTL with UI. The same reasoning concerning the RACE dummies applies covered in Analysis 1 holds here.

LOW AGE LWT FTV RACE.w RACE.b SMOKE PTL HT UI

LOW 1.00000000 -0.084510206 -0.18169054 -0.0860172363 -0.15915160 0.0955912948 0.161404311 0.260918194 0.152370249 0.16904283

AGE -0.08451021 1.000000000 0.18606140 0.2341718477 0.18183124 -0.1416201455 -0.047680690 0.118501012 -0.005180455 -0.07549513

LWT -0.18169054 0.186061402 1.00000000 0.0890236806 0.12288168 0.1724374005 -0.085300786 -0.110858547 0.169946075 -0.17538950

FTV -0.08601724 0.234171848 0.08902368 1.0000000000 0.14233948 -0.0004628739 -0.083294928 -0.014321796 -0.079767454 -0.05624939

RACE.w -0.15915160 0.181831238 0.12288168 0.1423394848 1.00000000 -0.4057766142 0.312510945 -0.029962259 -0.047535274 -0.03641187

RACE.b 0.09559129 -0.141620145 0.17243740 -0.0004628739 -0.40577661 1.0000000000 -0.005661528 -0.010650165 0.084992861 -0.03683439

SMOKE 0.16140431 -0.047680690 -0.08530079 -0.0832949283 0.31251095 -0.0056615280 1.000000000 0.187267307 0.013407037 0.06215900

PTL 0.26091819 0.118501012 -0.11085855 -0.0143217958 -0.02996226 -0.0106501648 0.187267307 1.000000000 0.001880484 0.19361817

HT 0.15237025 -0.005180455 0.16994608 -0.0797674544 -0.04753527 0.0849928609 0.013407037 0.001880484 1.000000000 -0.10858506

UI 0.16904283 -0.075495128 -0.17538950 -0.0562493871 -0.03641187 -0.0368343870 0.062158997 0.193618175 -0.108585058 1.00000000

Once more, our first model contained all eight variables. We made use of the same logic for interactions as explained in Analysis 1. For the new variables and their interactions their inclusion is explained here such as the interaction between LWT with HT, as heavier individuals tend to have higher blood pressure. PTL has a correlation with both SMOKE and UI, as smoking and uterine irritability increase the chances of premature labour.

|  |  |  |
| --- | --- | --- |
| **Model** | **AIC** | **AIC weight** |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI | 222.52 | 0.00 |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI+LWT\*HT | 224.71 | 0.00 |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI+AGE\*FTV | 213.65 | 0.11 |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI+SMOKE\*PTL | 224.68 | 0.00 |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI+PTL\*UI | 221.96 | 0.00 |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI+LWT\*AGE | 224.77 | 0.00 |
| LOW ~ LWT+AGE+RACE.w+RACE.b+FTV+SMOKE+PTL+HT+UI+AGE\*FTV + PTL\*UI | 212.24 | 0.23 |
| LOW ~ LWT+AGE+FTV+SMOKE+PTL+HT+UI+AGE\*FTV + PTL\*UI | 211.53 | 0.33 |
| LOW ~ LWT+AGE+FTV+SMOKE+PTL+UI+AGE\*FTV + PTL\*UI | 216.49 | 0.03 |
| LOW ~ LWT+AGE+FTV+PTL+HT+UI+AGE\*FTV + PTL\*UI | 212.30 | 0.22 |
| LOW ~ AGE+FTV+SMOKE+PTL+HT+UI+AGE\*FTV + PTL\*UI | 214.78 | 0.06 |

Using the significant interactions that we thought were logical, we did a model building process shown above. To start we compared the full model with different combinations of the significant interactions, before checking to see how the AIC score changed when adding each of the interactions separately. The AGE\*FTV and PTL\*UI interactions both improved the model by AIC score, resulting in our testing of the model with both interactions, which also continued to improve the model. Now that we had a model with the important interactions in it we looked at removing the main effects that were not in an interaction to come up with the best model which removed only the RACE dummy variables.

The final model, selected based on its lowest AIC score, looks like this:

E(LOW) = -1.124271 - 0.015795\*LWT + 0.072131\*AGE + 3.453559\*FTV + 0.638988\*SMOKE + 1.181280\*PTL + 1.863724\*HT + 1.586176\*UI - 0.155139\*AGE\*FTV - 1.605898\*PTL\*UI

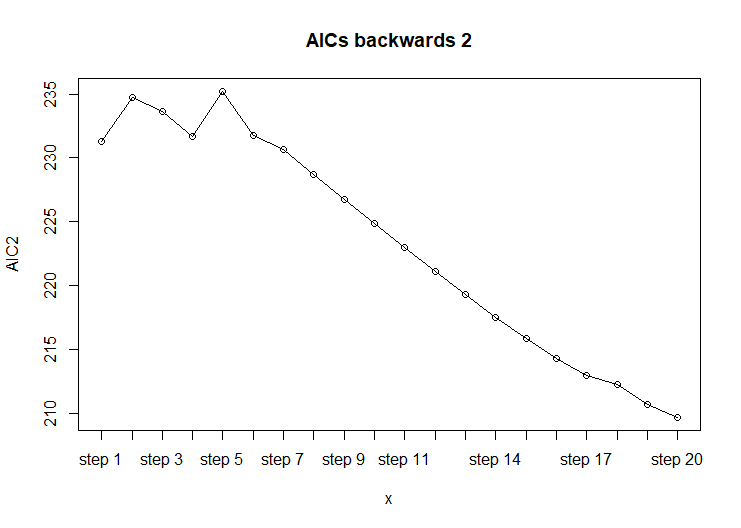
Compared to the first model in Analysis 1, the new model keeps all the same main effects and interactions along with the addition of some new variables and interactions listed above. The additional variables SMOKE, UI, PTL and HT were able to improve our model as they are good predictors for low birth weight and the relative health of the mother.

Backwards Elimination

The backwards elimination procedure was started from the full model with eight variables and all interactions. When analysing this model, all terms were highly significant. To solve this issue, all interactions previously shown to have no significance were deleted. Furthermore, the interaction term HT\*UI was not included, as this interaction between two binary variables resulted in unusable data. We used the following model to start the backwards elimination procedure.

E(LOW)=LWT+RACE.w+RACE.b+AGE+FTV+SMOKE+PTL+HT+UI+LWT\*SMOKE+LWT\*PTL+LWT\*HT+LWT\*UI+RACE.w\*SMOKE+RACE.b\*SMOKE+RACE.w\*PTL+RACE.b\*PTL+RACE.w\*HT+RACE.b\*HT+RACE.w\*UI+RACE.b\*UI+AGE\*FTV+AGE\*SMOKE+AGE\*PTL+AGE\*HT+AGE\*UI+FTV\*SMOKE+FTV\*PTL+FTV\*HT+FTV\*UI+SMOKE\*PTL+SMOKE\*HT+SMOKE\*UI+PTL\*HT+PTL\*UI

Each step, type I errors were controlled at a=0.1. The same method of operation was maintained as in the first backwards elimination. A graph of the stepwise AIC score is depicted below.



After consecutive elimination steps we were left with the following model.

E(LOW) = - 0.405554 - 0.015635\*LWT + 0.038048\*AGE + 3.528069\*FTV - 2.776036\*SMOKE + 5.489755\*PTL + 1.950286\*HT + 1.730875\*UI - 0.159603\*AGE\*FTV + 0.153579\*AGE\*SMOKE - 0.178684\*AGE\*PTL - 2.009435\*PTL\*UI

This model has a better AIC score then the model found in the model building exercise (209.7 compared to 211.53). Using the critical-r we removed or did not consider some interactions such as AGE\*SMOKE and AGE\*PTL that ended up being significant in the backwards elimination model. The critical-r method to determine significance is not the optimal way to do the model building process and leads to the main differences in our models, as well as the increased difficulty in model building when using a larger number of variables.

# Conclusion

The final model from our model building exercise based on AIC is

E(LOW) = -1.124271 - 0.015795\*LWT + 0.072131\*AGE + 3.453559\*FTV + 0.638988\*SMOKE + 1.181280\*PTL + 1.863724\*HT + 1.586176\*UI - 0.155139\*AGE\*FTV - 1.605898\*PTL\*UI

This model leads us to believe that the important factors for determining whether a mother will have a low birth weight child are weight at the last menstrual period, age, number of physician visits, smoking status during pregnancy, history of premature labor, history of hypertension, and presence of uterine irritability. Additionally, important are the interactions between age and number of physician visits and history of premature labor and uterine irritability.

The odds ratios of our main variables are listed here, along with their interpretations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| LWT | AGE | FTV | SMOKE | PTL | HT | UI |
| 0.9843294 | 1.0747956 | 31.6127040 | 1.8945621 | 3.2585434 | 6.4477030 | 4.8850352 |

|  |  |  |
| --- | --- | --- |
| **Interaction** | **Change in Behavior/Status** | **Chance of LWB baby** |
| LWT | One pound | 0.98x decrease |
| AGE | One year | 1.07x increase |
| FTV | One physician visit | 31.6x increase |
| SMOKE | YES | 1.89x increase |
| PTL | YES | 3.26x increase |
| HT | YES | 6.45x increase |
| UI | YES | 4.88x increase |

As the table shows, an increase in weight of one pound of the mother results in a 0.98 times decrease in odds of having an LBW baby. An increase in age of one year of the mother results in a 1.07 times increase in odds of having an LBW baby. An increase of one physician visit by the pregnant mother results in 31.6 times increased odds of having an LBW baby. A mother who smokes has 1.89 times increased odds of having an LBW baby. A history of premature labor has 3.26 times increased odds of having an LBW baby. A history of hypertension in the mother has 6.45 times increased odds of having an LBW baby. The presence of uterine irritability in the mother has 4.88 times increased odds of having an LBW baby.

We are confident in the model that we ended with as it is relatively simple and does not have too many interactions. All the interactions that are included make both logical and biological sense and are backed by the statistical methods used in our analysis. We are not quite sure whether using the critical-r method was the best way to help us select our interactions, as it only looks at whether there is a significant linear correlation. This led to the difference in models in Analysis 2 between our model building process and backwards elimination, as well as the increased difficulty of model building when evaluating a larger number of variables and interactions.