Logistic Regression: Estimating the Probability of Readmittance of Premature Babies into a Neonatal Unit

Nicholas Abad

January 21, 2018

Abstract

Because premature births are at the root of many short-term and long-term health related problems throughout a child's lifespan, the original aim of my project was to use the data that Langley et al. gathered in 2002 in order to come up with a statistical model that estimates the probability that a premature baby is re-admitted into a Neonatal Unit (NNU) within a year.[1] *By using backwards elimination and by analyzing the deviances between nested models using the ANODE test, I found that of the original* 9 *covariates, it is possible to predict our binary response variable, re-admittance, using logistic regression all while using only* 6 *of the original covariates in addition to* 2 *interaction terms. By training our model on a pseudo-random subset of 1200 observations and testing this model on the remaining 288 observations, my model was able to predict the probability of re-admittance with nearly 58% accuracy.*

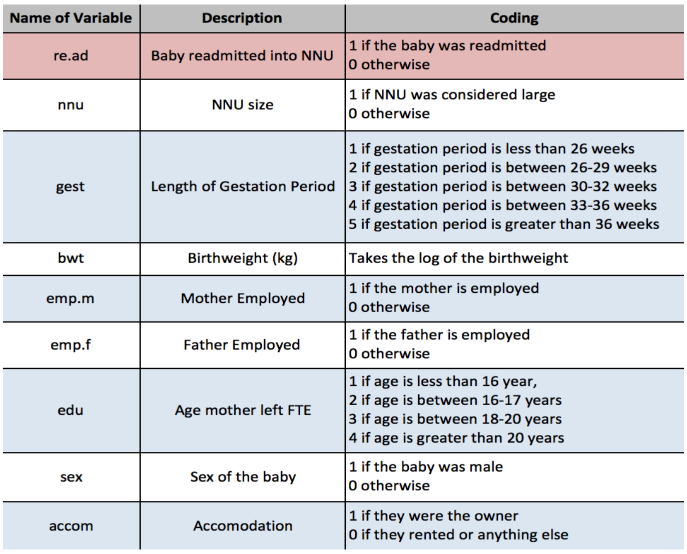
1 Introduction

According to a recent study done by the World Health Organization (WHO), there is an estimated 15 million babies being born prematurely throughout the world every single year, which is more than $\frac{1}{10}$ of all total yearly births. With this number steadily increasing year by year, there has undoubtedly been a large amount of studies conducted in order to learn more about the causes and subsequent effects that this has had, specifically on the baby being born. By definition, a baby is considered to be premature if his/her gestation period lasts less than 37 weeks, counting from the first day of the last menstrual period of the mother. [1]

Of those approximate 15 million premature births, approximately 1 million of those babies pass away due to complications at birth. For the fortunate survivors, however, there is still a significant number of both short-term and long-term effects that these babies must endure throughout their lifetimes. One of the most common problems that arise is that premature birth has been proven to effect the baby's brain, which can lead to a stunt in the baby's intellectual and developmental abilities as well as cause problems that negatively effect physical development, communication abilities, and learning capabilities.[1][2] Additionally, premature births are at the root of some intestinal problems, infections such as meningitis and pneumonia, vision problems, hearing-related issues, and dental problems.

When a baby is born prematurely and requires intensive care within 48 hours of his/her birth, hospitals typically take a precautionary measure and admit the baby into the Neonatal Unit (NNU). Knowing that this is typically the case, Langley and others [4] wanted to investigate the effects that communal neonatal services (CNS) has had on these babies who were admitted to the NNU within the first year of their life. The overall aim of their study was to decide whether or not a CNS can improve the overall satisfaction of both the premature baby and the mother.

In conducting their study, Langley's research group gathered a plethora of data not only related to the CNS but also about the birth weight, the employment of the mother and father, the sex of the baby, and several other categories on 1488 babies. With the dataset that Langley produced, the specific aim of my own project was to formulate a statistical model in order to predict whether a baby was readmitted into the NNU within a year given the details on those aforementioned categories (birth weight, employment, etc.). Ideally, this predictive model could be used to estimate the risk that a baby has of re-admittance into the NNU, given the combination of these categories. The list of those categories, their meaning, and the possible values that they could take can be seen in Figure 1.



Name of Variable	Description	Coding
re.ad	Baby readmitted into NNU	1 if the baby was readmitted 0 otherwise
nnu	NNU size	1 if NNU was considered large 0 otherwise
gest	Length of Gestation Period	1 if gestation period is less than 26 weeks 2 if gestation period is between 26-29 weeks 3 if gestation period is between 30-32 weeks 4 if gestation period is between 33-36 weeks 5 if gestation period is greater than 36 weeks
bwt	Birthweight (kg)	Takes the log of the birthweight
emp.m	Mother Employed	1 if the mother is employed 0 otherwise
emp.f	Father Employed	1 if the father is employed 0 otherwise
edu	Age mother left FTE	1 if age is less than 16 year, 2 if age is between 16-17 years 3 if age is between 18-20 years 4 if age is greater than 20 years
sex	Sex of the baby	1 if the baby was male 0 otherwise
accom	Accomodation	1 if they were the owner 0 if they rented or anything else

Figure 1: Description of Variables used. Note that the variable re.ad is highlighted in red due to it being the response variable.

2 Underlying Statistical Theory and Methods

2.1 Deciding on Variable Importance

Although it's possible that every single variable from an original dataset is associated with the response variable, this is a rare occasion and it's often the case that the response is only related to a smaller subset of those original variables. Because of this, we need to decide exactly which of these features are important by statistical methods such as **backwards elimination**.

The first step of this method is to have a model that contains all of the original p features and calculate the corresponding p-values of those features. Once calculated, remove the feature with the largest p-value from the model and create a new model that has only those remaining $(p-1)$ predictors. With the new model, calculate again the corresponding p-values and again remove the single feature with the highest p-value to create an even newer model with now $(p-2)$ features. Continue this process until all variables have a p-value that meets a certain criterion such as 0.20, 0.10, or 0.05. Choosing this criterion is left completely up to the person conducting the analysis.

2.2 Additive Model and Interaction Terms

Consider an additive model $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$ that has 6 predictors that are associated with a certain response variable Y. If two of these covariates, say X_2 and X_5 for example's sake, have a simultaneous influence on the response variable Y, we say that X_2 and X_5 are **interacting** with one another. If this is the case, we could model Y in terms of the following: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_2 X_5$.

In a more general sense, we could extend the model to have n predictors, rather than 6 in the previous example. With $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + ... + \beta_n X_n$, we can categorize an interaction term to be $X_i X_j \forall i \neq j$. When considering interaction terms that have 2 covariates, there are $\binom{n}{2} = \frac{n!}{2!(n-2)!}$ possible combinations that are possible.

In order to test whether or not this new interaction term has any effect on the model, we could use the ANODE test, which analyzes the difference in deviances and is further discussed in detail in Section 2.4. However, it's important to note that in both nested and unnested models, if you were to find an interaction term between $X_i X_j$ to be statistically significant (p-value is less than $0.10, 0.05$, or 0.01) and include that interaction in your model, you consequently need to include individually both X_i and X_j in the model even if they were to have insignificant p-values for those variables.

2.3 Modeling Data using Logistic Regression

Consider a dataset in which the response variable y that you are trying to model is binary and could therefore only take a value of 0 or 1. If you were to model this data using something such as linear regression, it's entirely possible for $p(Y = 0|X) < 0$ and/or for $p(Y = 1|X) > 1$. This model seemingly becomes insensible since probabilities could only range from 0 to 1, which would make this model useless since it is invalid.

One way that you could alleviate this is by choosing a different statistical model to model your data such as **logistic regression**. Logistic regression is based upon the logistic function:

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}} \tag{1}$$

After some manipulation, you could find the following:

$$\frac{p(Y = 1|X)}{1 - p(Y = 1|X)} = e^{\beta_0+\beta_1 X_1+...+\beta_n X_n} \implies log(\frac{p(Y = 1|X)}{1 - p(Y = 1|X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{2}$$

By using the method of maximum likelihood, we ultimately want the best estimates $\hat{\beta_0}$, $\hat{\beta_1}$, ..., and $\hat{\beta_n}$ of our actual β_0, β_1, ..., and β_n. Once we have these estimates, we could then plug these estimates into the original logistic function and calculate $\hat{p}(Y = 1|X)$ using $\frac{e^{\hat{\beta_0}+\hat{\beta_1}X_1+...+\hat{\beta_n}X_n}}{1+e^{\hat{\beta_0}+\hat{\beta_1}X_1+...+\hat{\beta_n}X_n}}$. By using this method, this alleviates the chance that probabilities could go below 0 and also eliminates the possibility that probabilities exceed 1, which was the original problem with linear regression. If the response variable were to have 3 or more possible outcomes, however, one should use other methods to model the data such as Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA). Since this is not the case, however, and since our response variable is binary, the best statistical model to use is logistic regression.

2.4 Analysis of Deviance (ANODE)

Consider two models in which a new model S is nested inside the more complete/complex model C. Let C have n predictors and let S have every predictor that C has except for the last predictor, predictor n. Therefore and in order to be explicit, model C has predictors $\{p_1, p_2, p_3, ..., p_n\}$ while model S has predictors $\{p_1, p_2, p_3, ..., p_{n-1}\}$.

By analyzing the difference of deviance between these two models, one could check whether or not this new model, model S in this case, is statistically more valid than the more complete model, model C. In order to do so, first compute the deviance of model S and subtract the deviance of model C and compare this value to a $\phi\chi^2_{c-s}$ distribution where $(c - s)$ is simply the difference in degrees of freedom between model C and the degrees of freedom in model S. Also, ϕ must be known and is equal to σ^2 in this case.

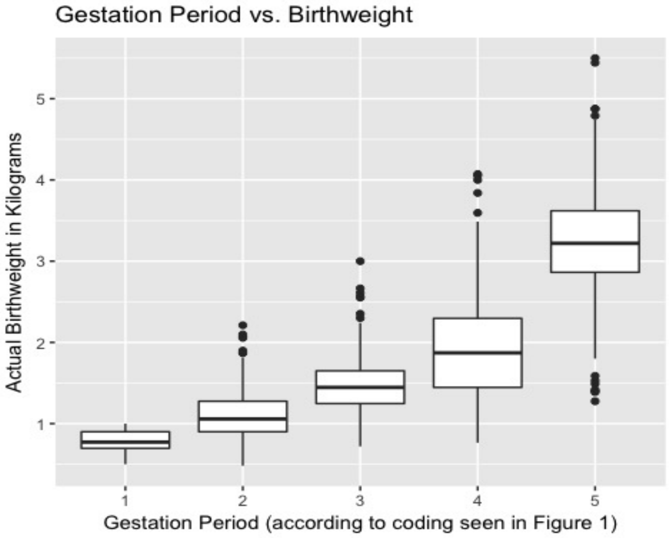
Choose the null hypothesis (H_0) to be that the true model is S and likewise, choose the alternative hypothesis (H_A) to be that the true model is actually C. If the deviance difference between both models is large in comparison to a $\phi\chi^2_{c-s}$ distribution, this would result in a small p-value, which would entail that the null hypothesis is not true meaning that the alternative hypothesis is more valid. On the other hand, however, suppose that you received a large p-value. This would entail that you are failing to reject the null hypothesis meaning that the null hypothesis is actually more valid.

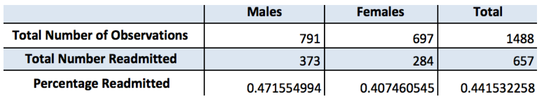
Therefore, if you want to show that the nested model is more valid than the more complete/complex model, a large p-value when comparing it to the $\phi\chi^2_{c-s}$ distribution will statistically show this.

3 Exploratory Analysis and Results

3.1 Understanding the Data

In order to get a good feel for the data and understand it better, it's always a good idea to try to come up with some exploratory plots and tables. However, because a majority of the variables were factors and only took on integer values, exploratory data analysis didn't prove to be the most useful exercise. Although this was the case, I was still able to find that gestation period plays





	Males	Females	Total
Total Number of Observations	791	697	1488
Total Number Readmitted	373	284	657
Percentage Readmitted	0.471554994	0.407460545	0.441532258

a huge roll in determining birth weight and I was also able to see that the overall percentage of males readmitted in this dataset was about 47% while the percentage of females readmitted was only about 41%.

3.2 Choosing the Generalized Linear Model

Before performing any type of statistical analysis on the data, we need to first define a specific generalized linear model that best describes the problem at hand given the dataset and what we're trying to predict, which in this case is the binary variable of "re.ad" in which "re.ad $=$ 0" indicates that a baby is not re-admitted while "re.ad $=$ 1" indicates that the baby is in fact re-admitted. Rather than going into this process blindly and randomly choosing between the many possible generalized linear models and different link functions, I think that the most important thing that someone can do is to critically think about what type of response variable that they are trying to measure and see which model most appropriately describes this process. As previously mentioned, our response variable is binary meaning that it could only take on two values so with this in mind, I decided to choose logistic regression as this would make the most sense.

When using logistic regression, it's important to note that this belongs to the binomial family and can be further specified with the "logit" link function. With the given birthweight dataset we are given, we could thus rewrite Equation 1 as the following:

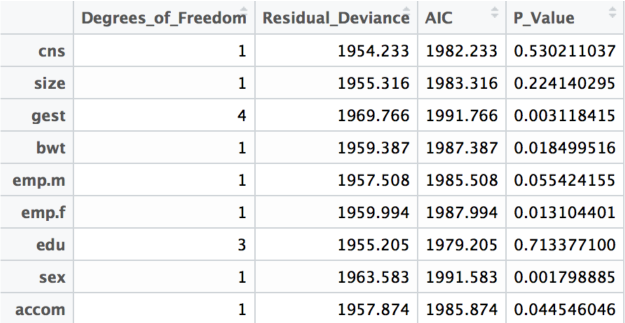
$$p(Y=1|X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}} = \frac{e^{\beta_0+\beta_1(CNS)+\beta_2(SIZE)+...+\beta_n(ACCOM)}}{1+e^{\beta_0+\beta_1(CNS)+\beta_2(SIZE)+...+\beta_n(ACCOM)}} \tag{3}$$

When rewriting this equation, it's necessary to first turn all variables that can be written as words into factors, such as CNS, SIZE, GEST, etc.. This just means that we get separate estimates for each value that that specific variable can take. In actuality, for our dataset, the only variable that was not turned into a factor was BWT (Birth weight).

In order to exemplify what turning a variable into a factor does, consider the variable GEST (Gestation) in which GEST can take any value within the set $\{1,2,3,4,5\}$. Rather than estimating gestation with a single estimate β, turning gestation into a factor would imply that we will find different estimates $\beta_{GEST=1}$, $\beta_{GEST=2}$, ..., and $\beta_{GEST=5}$ where each of these estimates do not have to be equal to each other. The original portion of the model $\beta\times$(Value of GEST) now turns into $\beta_{GEST=1}\times$(1 if GEST $=$ 1, 0 otherwise), $\beta_{GEST=2}\times$(1 if GEST $=$ 2, 0 otherwise), ... $\beta_{GEST=5}$(1 if GEST $=$ 5, 0 otherwise).

3.3 Creating the Additive Model

The next step in modeling our premature birth dataset is feature engineering, which consists of choosing what features or variables to include and exclude from our model. In order to do so, I chose to use **backwards elimination**, which was mentioned previously in Section 2.1.



	Degrees_of_Freedom	Residual_Deviance	AIC	P_Value
cns	1	1954.233	1982.233	0.530211037
size	1	1955.316	1983.316	0.224140295
gest	4	1969.766	1991.766	0.003118415
bwt	1	1959.387	1987.387	0.018499516
emp.m	1	1957.508	1985.508	0.055424155
emp.f	1	1959.994	1987.994	0.013104401
edu	3	1955.205	1979.205	0.713377100
sex	1	1963.583	1991.583	0.001798885
accom	1	1957.874	1985.874	0.044546046

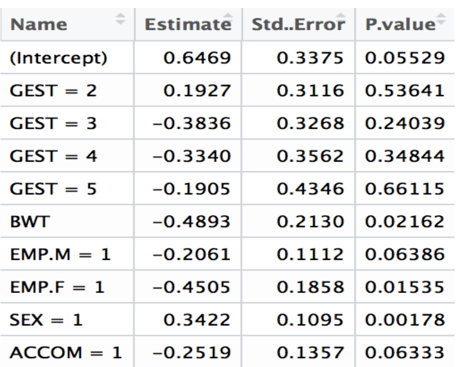
Figure 2: Degrees of freedom, residual deviance, AIC, and p-value corresponding to the list of variables in the complete model

By first starting with the full/complete model, I wanted to find the variable that had the largest p-value and eliminate it from the model. In this specific case and with an approximate p-value of 0.7133 (which can be seen in Figure 2), I temporarily eliminated the variable CNS from the model. In order to determine whether or not this new model without CNS is valid, however, I needed to conduct an ANODE test, which compares the deviances of both models. In this case, the null hypothesis (H_0) is that the true model is the new model, which we considered to be the model without the variable CNS. On the other hand, the alternative hypothesis (H_A) is that the true model is the complete model that includes all variables including CNS. By comparing the difference in deviances of both models and comparing this difference to a χ^2_3 distribution in which 3 denotes a difference of 3 degrees of freedom between the null and complete model, we found an insignificant p-value of 0.7134, which means that there is evidence that the new model is better fit for our data than the complete model. After doing this several times, we finally stopped using backwards elimination once every remaining variable was found to be statistically significant at the 0.10 level. **The variables that remained in the model after using backwards elimination were found to be GEST, BWT, EMP.M, EMP.F, SEX, and ACCOM.**

Because of this, we now have an additive model of the following form

$$log(\frac{P(Re.ad=1|X)}{1+P(Re.ad=1|X)}) = \hat{\beta_0}+\hat{\beta_1}(GEST=2)+\hat{\beta_2}(GEST=3)+\hat{\beta_3}(GEST=4)+\hat{\beta_4}(GEST=5)$$
$$+\hat{\beta_5}(BWT)+\hat{\beta_6}(EMP.M=1)+\hat{\beta_7}(EMP.F=1)+\hat{\beta_8}(SEX=1)+\hat{\beta_8}(ACCOM=1) \quad (4)$$

where all covariates except for BWT are indicator variables, signifying that the value is 1 if that covariate is true and 0 otherwise. For example, if an observation has $GEST = 2$, we would substitute in 1 for $GEST = 2$ in Equation 4 and similarly substitute in 0 for $GEST = 3$, $GEST = 4$, and $GEST = 5$. The estimates, standard errors, and p-values can thus be found in Figure 3.



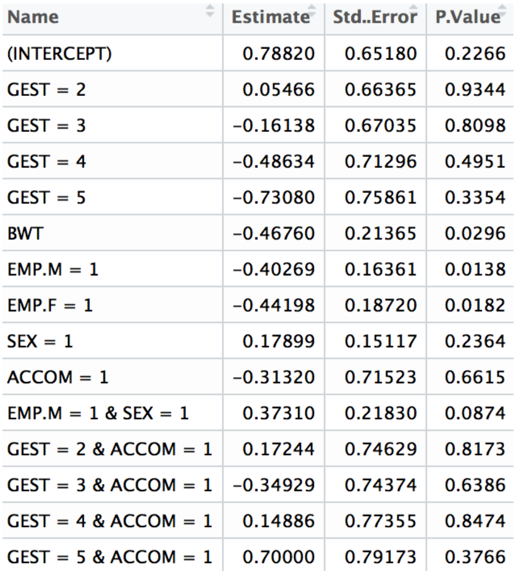
Name	Estimate	Std..Error	P.value
(Intercept)	0.6469	0.3375	0.05529
GEST = 2	0.1927	0.3116	0.53641
GEST = 3	−0.3836	0.3268	0.24039
GEST = 4	−0.3340	0.3562	0.34844
GEST = 5	−0.1905	0.4346	0.66115
BWT	−0.4893	0.2130	0.02162
EMP.M = 1	−0.2061	0.1112	0.06386
EMP.F = 1	−0.4505	0.1858	0.01535
SEX = 1	0.3422	0.1095	0.00178
ACCOM = 1	−0.2519	0.1357	0.06333

Figure 3: The remaining covariates left in the additive model.

3.4 Adding in Interaction Terms

With this final additive model, it is now time to consider interaction terms that may make our model better. Recall from Section 2.2 that in general there is a total of $\binom{n}{2}$ possible interaction terms to consider. In our case however, since there are 6 non-factored covariates, we need to consider $\binom{6}{2} = 15$ total interaction terms.

For each of the 15 interaction terms, I decided to create 15 entirely new models in which the 6 predictors that I found to be significant as well as the single interaction term at hand were included. By using the ANODE test, I compared each of these newer models to the simple additive model and looked for p-values that were statistically significant, keeping in mind the difference in degrees of freedom when comparing the models to a $\chi^2_{(c-s)}$ distribution once more. After conducting 15 of



Name	Estimate	Std..Error	P.Value
(INTERCEPT)	0.78820	0.65180	0.2266
GEST = 2	0.05466	0.66365	0.9344
GEST = 3	-0.16138	0.67035	0.8098
GEST = 4	-0.48634	0.71296	0.4951
GEST = 5	-0.73080	0.75861	0.3354
BWT	-0.46760	0.21365	0.0296
EMP.M = 1	-0.40269	0.16361	0.0138
EMP.F = 1	-0.44198	0.18720	0.0182
SEX = 1	0.17899	0.15117	0.2364
ACCOM = 1	-0.31320	0.71523	0.6615
EMP.M = 1 & SEX = 1	0.37310	0.21830	0.0874
GEST = 2 & ACCOM = 1	0.17244	0.74629	0.8173
GEST = 3 & ACCOM = 1	-0.34929	0.74374	0.6386
GEST = 4 & ACCOM = 1	0.14886	0.77355	0.8474
GEST = 5 & ACCOM = 1	0.70000	0.79173	0.3766

Figure 4: The final model after including interaction terms as well as their corresponding estimates, standard errors, and p-values.

these tests, the following interaction terms that I found to be significant when compared to the additive model were $GEST \times SEX$, $GEST \times ACCOM$, $EMP.M \times SEX$, and $SEX \times ACCOM$.

Now that we have these 4 interaction terms, I decided to create a new base model that included all 6 of the original predictors as well as all of the 4 interaction terms that we just found to be significant. Using backwards elimination once more, I took this model and realized that $SEX \times ACCOM$ had the largest p-value, created a new model without $SEX \times ACCOM$, compared the deviances of the two models and concluded that I should not include $SEX \times ACCOM$ into my model. On the next iteration, I realized that $GEST \times ACCOM$ had the largest p-value and attempted to compare the deviances between the model with that interaction term and the model without that interaction term. When comparing this to a χ^2_1 distribution, however, I found the p-value to be 0.10494, which fell directly on my cut-off point of 0.10. Because it was so close, I ultimately decided to keep this variable in my model and stop the backwards elimination process due to every other variable being significant at this level as well. The final model, with interaction terms this time, has now been chosen to be:

$$\begin{multline}\hat{\beta}_0 + \hat{\beta}_1(GEST = 2) + \hat{\beta}_2(GEST = 3) + \hat{\beta}_3(GEST = 4) + \hat{\beta}_4(GEST = 5) \\ + \hat{\beta}_5(BWT) + \hat{\beta}_6(EMP.M = 1) + \hat{\beta}_7(EMP.F = 1) + \hat{\beta}_8(SEX = 1) + \hat{\beta}_8(ACCOM = 1) \\ + \hat{\beta}_9(EMP.M = 1\&SEX = 1) + \hat{\beta_{10}}(GEST = 2\&ACCOM = 1) + \hat{\beta_{11}}(GEST = 3\&ACCOM = 1) \\ + \hat{\beta_{12}}(GEST = 4\&ACCOM = 1) + \hat{\beta_{13}}(GEST = 5\&ACCOM = 1) \quad (5)\end{multline}$$

As you can see in Figure 4, the p-values for every single term that includes $GEST$ is seemingly insignificant, which would therefore prompt people to delete this term from the model. Although I considered this to be the case, I ultimately decided to leave this term in because I conducted an ANODE test and compared the deviances of the model with all of the $GEST$ terms included to the model without any of the $GEST$ terms included. By doing so and finding a significant p-value, this entails that the alternative hypothesis, which states that the model with variables that contain $GEST$ is the true model, is actually true and that we should reject the null hypothesis.

3.5 Goodness of Fit

In order to check the goodness of fit of my model (or lack thereof), I decided to use the Hosmer-Lemeshow test, which is based upon dividing samples into their predicted probabilities and is typically chosen as the typical goodness of fit test for logistic regression. According to Hosmer and Lemeshow in a paper that was published in 1980 [5], the test statistic that they use within their test approximately follows a χ^2_{g-2} distribution when the model is correctly specified and where g represents the amount of groups that you have split the probabilities into. When running the Hosmer-Lemeshow test, a small p-value signifies a poor fit of the model. On the other hand, however, it should be noted that a large p-value does not indicate a good fit but just signifies that the model does not fit poorly.

When running the aforementioned Hosmer-Lemeshow test on the final interaction model, I found it to have a p-value of 0.2116, which is not statistically significant. Because this is the case, I concluded that my model did not fit the data poorly but cannot conclude that my model fits the data well.

3.6 Accuracy

With a final model in hand, I decided to try to test the accuracy of this model by splitting my data into a training and testing set. By choosing my training set to be a (pseudo) random subset of 1200 of the 1488 observations and choosing my testing data to be the remaining 288 observations, I was able to discover that my accuracy was exactly 0.5798611.

4 Conclusion and Discussion

Once I was finished with all of my feature engineering and statistical analysis, I was able to create a final model that included a total of 8 total covariates, 2 of which happened to be interaction terms. Those covariates, their estimates, their standard errors, and their final p-values could be found in Figure 4. By using logistic regression in order to model our data, which implies that the model belongs to the binomial family with a logit link function, I was able to come up with an out-of-bag error rate of 0.4201389, which ultimately came as a surprise to me. I was hoping that the final model that I came up with would've done a bit better given the decent amount of observations and the amount of possible predictors but I am however finding a bit of relief knowing that for every step I took during this process, the step at the moment was statistically and intuitively justified. For example, when deciding which variables to take out of my model using backwards elimination and the ANODE test, I chose a relatively strict cut off point of 0.10, since this is a common cut off point when testing statistical significance. Additionally, when deciding which generalized linear model to choose as well, I thought that since the response variable was binary, the most logical choice would be to use logistic regression, which would imply choosing the binomial family as well as the logit link function. My choice was then supported by testing different generalized linear models and different link functions. Regardless, I think that by increasing the total amount of observations, I could have assuredly lowered my out-of-bag error rate as this would have given every one of my models more data observations to train on. In the end and although I was expecting a better result, there is still no doubt in my mind that my model could still help people estimate the probability that a baby is re-admitted into the NNU, given data regarding gestation, birth weight, the employment of the mother, the employment of the father, the sex of the baby, and the accommodation.

References

[1] World Health Organization (WHO) Media Centre: *Preterm Fact Sheet* November 2017 <http://www.who.int/mediacentre/factsheets/fs363/en/>

[2] March of Dimes Organization *Long-term Health Effects of Premature Babies* October 2013 <https://www.marchofdimes.org/baby/long-term-health-effects-of-premature-birth.aspx>

[3] Bowden, J. and Whittaker, J. *A latent variable scorecard for neonatal baby frailty, 159-172* 2005: <http://smj.sagepub.com/content/5/2/159.full.pdf+html>

[4] Langley, D., Hollis, S., Friede, T., MacGregor,D., and Gatrell, A. *Impact of community neonatal services: a multicentre survey. Archives of Disease in Childhood: Fetal and Neonatal Ed. 87:F204-F208* 2002: <http://fn.bmj.com/cgi/reprint/87/3/F204.pdf>

[5] Barlett, Jonathan *The Hosmer-Lemeshow goodness of fit test for logistic regression* February 2014: <http://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/>