PREDICTING READMISSION FOR HIGH-RISK INFANTS WITHIN 1 YEAR OF INITIAL DISCHARGE: A GENERALISED LINEAR MODEL

32102717

Abstract

Around 8% of babies in the UK are born prematurely [15]. Premature birth can lead to severe health complications in both early life and adulthood [12]. Such complications are the leading cause of death in children under 5 [21]. We aimed to create a generalised linear model to predict the probability that a high-risk infant will be readmitted to the neonatal unit within 1 year after their initial discharge. We utilised data on 1488 high-risk infants collected for a multi-centre study by Langley et al. (2002) [10]. Our model was built using backwards elimination and multiple methods of interaction selection, with the Hosmer-Lemeshow test and cross validation being the main metrics by which we judged its quality. We suggest a model that could be used with some alteration to the sensitivity, to fit the needs of the hospital, but recommend that further research be undertaken that considers more advanced model building techniques and additional explanatory variables that are known to affect the probability of readmission in high risk infants.

---

1. Introduction

Premature birth is defined as occurring before 37 weeks of pregnancy [12;15]. Incidence of pre-term births is already high in the UK, with 8 in every 100 babies being born prematurely [15], with such figures continuing to increase globally [21]. Pre-term babies may suffer from a large number of health problems, the most common being jaundice and feeding problems [17] and the most serious being bleeding in the brain [2;12]. They can also suffer from long-term issues that can affect their whole lives [12]. The shorter the gestation period of the baby the greater the risk of these short and long term health issues [2]. It is also known that pre-term babies are 1.5 to 3 times more likely to be readmitted to the hospital within the first year of life than their full term counterparts [17].

During this study we wished to develop a generalised linear model to predict the probabilities of readmission within 1 year after initial discharge for a set of high risk neonatal survivors , based on a data collected by Langley et al. (2002) [10]. As previously mentioned there are large health risks caused by premature birth, but there are also extreme costs associated with the care of high risk infants for hospitals, and limited resources available [16]. We could not justify prioritising one over the other, so we compromised to aim to create the most accurate model possible, and gave preferential treatment to those variables that we had reason to believe would improve the predictive accuracy of the model.

1.1. The Langley CNS Study (2002) [10].

During this investigation we used data collected by Langley et al. (2002) [10]. They investigated 2181 infants who had a birth weight less than or equal to 1500g or who had received level I intensive care for at least 48 hours after birth [10]. Many infants, however, had missing information, which reduced the data set to 1488 infants for our purposes. The Langley et al. (2002) [10] study excluded multiple births (such as twins), infants who had died, and infants with severe congenital abnormalities [10]. The study occurred over 32 centres across the UK, and was designed to support previous small sample, single centre studies that claimed that with the use of Community Neonatal Services (CNS) the initial length of stay of a high risk infant can be reduced with no subsequent increase in readmission probabilities. In total there were 10 explanatory variables (covariates) for the response variable *readmission*. Details of the variables considered are found in Table 1. Let it be noted that in Langley et al. (2002) [10] *length of stay* is treated as a response variable. In our investigation we decided to use *length of stay* as an additional explanatory variable for *readmission*, thus the model created in this investigation is designed to be consulted at the initial discharge of the infant.

2. Methods: What is a Generalised Linear Model?

In a simple regression model, a 1 unit increase in one of the covariates, x_m, leads to a β_m change in the response variable, where β_m is the coefficient for the explanatory variable x_m. This relationship does not hold for generalised linear models, thus we employ them when the response variable does not change linearly with the covariates. A GLM takes the form:

Variable	Description	Levels (Coding)	Level Description
re.ad	Readmission (response variable)	N/A	No =0, Yes = 1
cns	Was a CNS provided?	cns0	No
		cns1	Yes
size	Size of Neonatal Unit (NNU)	size0	Small
		size1	Large
gest	Gestation Period in Weeks	gest1	$<$ 26 weeks
		gest2	26-29 weeks
		gest3	30-32 weeks
		gest4	33-36 weeks
		gest5	$>$ 36 weeks
bwt	Birth Weight	N/A	N/A
emp.m	Mother employed?	emp.m0	No
		emp.m1	Yes
emp.f	Father/partner employed?	emp.f0	No
		emp.f1	Yes
edu	Age (in years) Mother left full time education	edu1	$<$ 16 years
		edu2	16-17 years
		edu3	18-20 years
		edu4	$>$ 20 years
los	Time until initial discharge in log(days)	N/A	N/A
sex	Sex of Baby	sex0	Female
		sex1	Male
accom	Parents own house?	accom0	No
		accom1	Yes

TABLE 1. All variables considered for the construction of a generalised linear model to predict the probability of readmission for a high risk infant, their descriptions, and their levels.

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \left(\beta_m * x_{m,i}\right) + (\text{interaction terms}), \tag{1}$$

where,

- β_0 is the intercept,
- β_m is the coefficient of the explanatory variable x_m,
- m is the index of the parameters, for $m \in (1, \ldots, M)$,
- η_i is the linear predictor for $\boldsymbol{x}_i$, the observations for the i^{th} infant,
- The mean of the response variable, $p_i = g^{-1}(\eta_i)$,
- and g is known as a link function.

2.1. **Assumptions of a GLM.** A generalised linear model has a set of assumptions [14] that must be satisfied for the model to be a valid GLM, however these assumptions are relaxed enough to encompass a large range of models, while also being strict enough to provide unified methods of estimation and inference. These assumptions are:

(i) **The response variable**: The response variable, Y, is one dimensional, indexed by i for $i \in (1, \ldots, N)$. The observations, $y_1, \ldots, y_N$, of the response variable, Y, are all independent observations of a random variable.

(ii) **The explanatory variables**: The explanatory variables, $\mathbf{X} = (X_1, ..., X_M)$, are also one-dimensional, where $M < N$ (otherwise we will have more variables than observations).

(iii) **Exponential Family**: The distribution of the response variable conditional on the values of the explanatory variables, $Y|\boldsymbol{x}$, is a member of the Exponential Family (see §2.2).

(iv) **Link function**: There exists a single linear predictor, η_i, which influences the distribution of $Y_i|\boldsymbol{x}_i$, where $x_{m,i}$ does not influence the distribution of $Y_i|\boldsymbol{x}_i$ if and only if $\beta_m = 0$. The mean, p_i, is related to η_i through a link function g (see §2.3).

2.2. **Exponential Family.** A random variable, Y, is said to belong to the exponential family if its probability density function can be expressed in the form [14]:

$$f(Y|\theta,\phi) = \exp\left\{\frac{y\theta - k(\theta)}{\phi} + c(y,\phi)\right\}, \tag{2}$$

where θ is the canonical parameter, ϕ is the scale parameter, and k and c are specified functions such that f integrates to 1. It is known that θ and ϕ are not unique.

As an example, we will show how the probability density function of $Y \sim \text{Bernoulli}(p)$ can be expressed in the exponential family form:

$$f(Y|\theta,\phi) = p^y(1-p)^{1-y}, \tag{3}$$

becomes,

$$f(Y|\theta,\phi) = \exp\left\{y\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\} \tag{4}$$

where $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$, $k(\theta) = -\log(1-p) = \log(1+e^\theta)$, and $c(y,\phi) = 0$.

2.3. **The Link Function.** The link function[14], g, is a monotonic function which maps the range of the mean, p, to the range of linear predictor, η. Thus any monotonic function that maps $[0,1]$ to $\mathbb{R}$ is a valid link function for a binary response variable, such as Bernoulli. We will, however, focus our attention on three common choices;

(1) The logit: $g(p) = logit(p) = \log(\frac{p}{1-p})$. The canonical link of a Bernoulli random variable[14].
(2) The Probit: $g(p) = \Phi^{-1}(p)$, where Φ is the cumulative distribution function of a N(0,1) distribution.
(3) The complementary log-log: $g(p) = \log\{-\log(1-p)\}$.

While all three of these link functions satisfy the relationship $g(p) = \eta$, they imply different relationships between the explanatory variables. We choose the link function which best matches the relationships in the data. When the link function is the logit, this is known as logistic regression, when the link function is Probit, this is known as probit regression. The canonical link function[14], g_c, is derived from the exponential family form of a distribution and satisfies the relationship $\theta = g_c(p) = \eta$. The canonical link is found by calculating $[k'(\theta)]^{-1}$.

2.4. **Deviance and ANODE.** The residual deviance of a GLM is defined as[14]:

$$D(\hat{\boldsymbol{\beta}}) = 2\{\ell(\mathbf{y},\phi) - \ell(\hat{\boldsymbol{\beta}},\phi)\}, \tag{5}$$

where,

- $\hat{\boldsymbol{\beta}}$ is the vector of coefficient estimates,
- $\mathbf{y}$ is the observed values of the response variable,
- $\ell(\mathbf{y},\phi)$ is the log-likelihood for $\boldsymbol{p}$ evaluated at $\mathbf{y}$, with respect to the known scale parameter, ϕ,
- and $\ell(\hat{\boldsymbol{\beta}},\phi)$ is the log-likelihood for $\boldsymbol{p}$ evaluated at the fitted values, $\hat{\boldsymbol{\beta}}$, with respect to the known scale parameter, ϕ.

Thus the deviance of a model is a measure of the difference between the observed values of the response variable and the fitted values of the response variable. This can also be used to compare whether two nested models are significantly different. To do this we use an analysis of deviance (ANODE)[14], also known as a likelihood ratio test.

Consider two models; GLM_C is the complex model with t explanatory variables, $\boldsymbol{X}_C$, and GLM_S is the simple model with $k < t$ explanatory variables, $\boldsymbol{X}_S \in \boldsymbol{X}_C$. For analysis of deviance,

$$\text{H}_0 : \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_S, \quad vs. \quad \text{H}_\text{A} : \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_C,$$

where $\boldsymbol{\beta}$ is the true coefficient values, and $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_S$ are the coefficient estimates of the complex and simple models respectively. Under the null hypothesis,

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}_S,\phi)\} - \ell(\hat{\boldsymbol{\beta}}_C,\phi)\} \sim \chi^2_{(t-k)}. \tag{6}$$

Thus we can say, using a 5% significance level, that if the p-value for W is < 0.05, then there is a significant difference between the models, and the simple model is not a valid reduction of the complex model.

3. Methods: Model Selection

The overall goal of this investigation is to build a generalised linear model based on the observed covariates that accurately predicts the probability of readmission for a high risk infant given their covariate values. This model then also needs to generalise to the population of all pre-term babies well, not just the data that the model is built upon. This raises two questions; how do we build such a model, and how do we know if the model is good? We will begin by discussing the second question. Assume that we have our final model and it takes the form:

$$\eta_{\text{i}} = \beta_0 + \sum_{m=1}^{M} \left(\beta_m * x_{m,i}\right) + (\text{interaction terms}), \tag{7}$$

where β_0 is the intercept, β_m are the coefficients of the covariates x_m, and the probability of readmission, $p_i = g^{-1}(\eta_i)$, where g is known as a link function. Coefficient estimates, β_0 and β_m for $m \in (1, \ldots, M)$ are computed using iteratively re-weighted least squares [14].

How do we know if this is a good model, and how do we know if it is better than another model we could consider? There are many factors that make up a good model; how well it fits the data that estimate its covariates, how well it generalises to the population as a whole, and how complex it is.

The first thing we wish to do is test how well the model fits the data that estimates its coefficients. To do this we use a goodness of fit test. In general there are many GoF tests that exist, but for a Bernoulli response variable there is only one that we can use, the Hosmer-Lemeshow test (See §3.4). If the p-value of the test is significant, tested at the 5% significance level, then we say that the model is a poor fit for the data, it does not predict well. If it is non-significant then we say that the model fits the data well.

One method to test how well the model generalises to new data is by dividing our original data set into k groups; $(k-1)$ of the groups are combined to create a training set used to fit the coefficient estimates of the model, and the final group is used as a test set of unseen data to evaluate the accuracy of the model for new data. This is repeated k times, each time using a different group as the test set, and the results are averaged. This is known as cross validation [4].

The complexity of the model is defined by how many terms it has and what kind of terms they are. For instance, a main effect term is less complex than an interaction term, and a two-way interaction term is less complex than a three-way interaction term. However, a model with 2 main effect terms and their interaction is less complex than a model that has 20 main effect terms but no interactions. When building a model we want to find the most parsimonious model. This is the model with the least number of terms possible to explain the greatest amount of information possible. There is a trade off between the number of terms in the model and the amount of variance that is explained, a good model wants to strike a balance between the two. Having too many terms, and overly complex terms, can usually lead to over fitting, which means that while the model may predict the training data well, it will not generalise to new data. Now that we understand what is considered a good generalised linear model, we can begin to form one for a given set of data.

3.1. **Pre-modelling checks.**

After exploring the data we wish to fit a generalised linear model for the response variable, readmission. Our first step is to make an assumption about the distribution the response variable given the observed values of the explanatory variables, $Y|\boldsymbol{x}$. The conditional distribution of Y is a member of the exponential family, with mean p and fixed (known) scale parameter ϕ. In the context of our study the response variable is binary, taking a value of 1 if the baby is readmitted to the neonatal unit within of year of their initial discharge, and 0 if not. Clearly the best choice is to assume that $Y_i|\boldsymbol{x}_i \sim \text{Bernoulli}(p_i)$.

Under this assumption of a Bernoulli distribution, we now want to check for multicollinearity. Multicollinearity occurs when one of the explanatory variables can be expressed as a linear combination of one or more of the other explanatory variables in the model [14]. In more precise terms, if multicollinearity is not present then the design matrix has full rank, and we can assume that all the explanatory variables are independent [14]. To test for multicollinearity we can calculate the (generalised) variance-inflation factor [5] for each explanatory variable in the additive model. If for all covariates the $(\text{GVIF})^{\frac{1}{d}}$ is less than 5, where d is the number of discrete levels of a variable (degrees of freedom), then we can state that there is no instance of multicollinearity. The GVIF is the extension of the VIF for explanatory variables with more than one coefficient (i.e. factors). For explanatory variables that only require 1 coefficient, the GVIF and $\text{GVIF}^{\frac{1}{d}}$ are equivalent to the VIF, and we take GVIF to the power of $\frac{1}{d}$ to make it comparible for different levels of d [5].

Next we consider whether a transformation of any of the continuous explanatory variables would lead to a better fitting model. Since our response variable is binary, a plot may not help us to decide which transformation of the variable is a better fit. Thus we decide to choose the form of the covariate that fits the data best under a univariate model. This is evaluated using the Hosmer-Lemeshow test (see §3.4 for more details on the Hosmer-Lemeshow test).

3.2. **Preliminary Model selection.** Once we have completed the pre-modelling checks, we can begin to build a model. The first step is to consider which explanatory variables will make up our additive model. To do this we fit univariate models for each of our M explanatory variables of the form:

$$\eta_m = \beta_{0,m} + \beta_m * x_m, \tag{8}$$

where β_0 is the intercept, β_m are the coefficients of the covariate x_m, $m \in (1, ..., M)$, and the probability of readmission, $p_i = g^{-1}(\eta_{m,i})$, where g is known as a link function.

If the covariate in a univariate model is not statistically significant under a Z-test then we do not include it in our initial investigation. For the Z-test;

$$\mathrm{H}_0 : \beta_0 = 0, \quad vs. \quad \mathrm{H}_\mathrm{A} : \beta_0 = \beta,$$

where β_0 is the true value of the coefficient and β is the coefficient estimate. Under the null hypothesis;

$$Z = \frac{\beta}{\sigma_\beta} \sim \mathrm{Normal}(0, 1), \tag{9}$$

where σ_β is the standard error of the coefficient estimate. Testing at the $100(1-\alpha)\%$ significance level, a non-significant p-value suggests the term does not affect the response variable. It is important to note that while one level of an explanatory variable may be non-significant, others may be, and one can not remove any levels without removing all levels for that explanatory variable. In this stage we use a significance level of 25%, which is supported by literature [1;11].

We now create an initial additive model containing all the explanatory variables which were significant in the univariate case, as well as any we believe to be of contextual importance. The additive model takes the from:

$$\eta_\mathrm{i} = \beta_0 + \sum_{m=1}^{M} \big(\beta_m * x_{m,i}\big), \tag{10}$$

where β_0 is the intercept, β_m are the coefficients of the covariate x_m, and the probability of readmission, $p_i = g^{-1}(\eta_i)$, where g is known as a link function.

Our goal, as stated above, is to attain the simplest model that contains the most information. Starting from the additive model, we wish to remove terms that do not contribute a significant amount of information to the model. We can do this using backwards elimination. At each step in backwards elimination, we take our current model GLM_C and create k simpler models by removing the k^{th} term in each, for $k \in (1, ..., M_c)$, where M_c is the number of covariates in the current model. We then use analysis of deviance to calculate whether each reduced model is valid by considering its p-value. We have chosen to test at the 15% significance level. If all the reduced models have a p-value below 0.15 then we can state that there are no further valid reductions of the current model. Otherwise, the reduced model with the largest p-value is taken as the new current model, and the process is repeated. We can also use our discretion to included any borderline significant covariates that we feel are contextually important, despite them not being significant. At the end of each step we will also consider the percentage change in each of the coefficients that are common to both the reduced and current models. If any coefficients change by more than 20%, then we can assume that the removed covariate adjusted for one or more of the other covariates, and thus decide to include it in the model, despite it not being significant. At the end of the backwards elimination we will have attained a preliminary main effects model.

It is now worth checking that the continuous explanatory variables in our preliminary main effects model have a linear relationship with g(response variable), where g is the link function. If the relationship between the two is non-linear, we will need to apply other model building techniques such as including higher-power terms, fractional polynomials and spline function [18;19].

3.3. **Interactions.** Interactions between two variables are valid when one variable changes at a different rate depending on the value of another. For instance, the weight at birth for a male baby may be higher on average for that of a female baby, thus there is an interaction between birth weight and gender. We add interaction terms into our model in order to incorporate these relationships. For our modelling process we will only look at two-way interactions, that is interactions between at most two covariates.

There are many ways to add interactions into the model, we will consider two alternatives. The first method we could use is reminiscent of backwards elimination. We can add every possible two-way interaction term into the model, and as we did for the main effects, use backwards elimination to reduce the model one term at time until all the terms were significant, this time testing at the 5% level. This method is in line with what we have already done, however it has some drawbacks. Firstly, the more terms that are in our preliminary main effects model the more interaction terms there are. For a preliminary main effects model with 10 terms there will be 45 possible two-way interaction terms, which will take a long time to run backwards selection on. This method also produces a model with lots more terms than the alternative method that follows, and thus may not give the most parsimonious model.

The alternative is to consider the interaction terms on their own and only choose to include those that make a significant contribution to the model. We create a model for each interaction of the form:

$$\eta_{\mathrm{i}} = \beta_0 + \sum_{m=1}^{M} \left(\beta_m * x_{m,i}\right) + (\text{two-way interaction term})_{m,k}, \tag{11}$$

where β_0 is the intercept, β_m are the coefficients of the covariate x_m in the preliminary model, and the probability of readmission, $p_i = g^{-1}(\eta_i)$, where g is known as a link function. The interaction is between the m^{th} covariate and the k^{th} covariate, where $m < k$ and $m, k \in (1, ..., M)$ where M is the number of explanatory variables in the preliminary main effects model. The interaction term for the m^{th} covariate with the k^{th} covariate is the same as the interaction term for the k^{th} covariate with the m^{th} covariate.

We then use analysis of deviance to test if the removal of the interaction term for each model is valid. For all the models with significant p-values, tested at the 5% significance level, the interaction term is added to the preliminary main effects model to create a preliminary final model. We also have the option of not testing every single interaction if there are a large number, and using intuition and context to test those that we believe may be important. This method has the advantage of potentially being quicker (as we have the option of not considering all terms, and it will have no effect on our analysis of the other terms if we do not), and the models that are created via this method usually contain less interaction terms than the previous method. It has the disadvantage, however, that none of the additional information gained by including each interaction term is accessed in regards to the other interaction terms.

In this investigation we will consider models built using both methods, and will test all possible two-way interaction terms for both methods.

3.4. **Model Diagnostics.** We now have a preliminary final model of the form:

$$\eta_{\mathrm{i}} = \beta_0 + \sum_{m=1}^{M} \left(\beta_m * x_{m,i}\right) + (\text{significant two-way interaction terms}), \tag{12}$$

where β_0 is the intercept, β_m are the coefficients of the covariate x_m in the preliminary model, and the probability of readmission, $p_i = g^{-1}(\eta_i)$, where g is known as a link function.

We wish to test how well this model fits our data, and whether there are any changes we can implement to make it fit better. We start by considering how well the model fits the data, to do this we can use a goodness of fit test. One goodness of fit test we can use is known as the Hosmer-Lemeshow test [6;7;8;9]. The Hosmer-Lemeshow test works by the following:

(1) The observations are put into ascending order based on their predicted probabilities for $Y = 1$.
(2) The ordered observations are then split into g approximately equal sized groups.
(3) For each group, the proportion of observations for which $Y = 1$ should be around the average estimated probability for the group.
(4) The same is then done for the probabilities that $Y = 0$.
(5) The Pearson Goodness of Fit statistic is then calculated. Under the null hypothesis,

$$H = \sum_{k=0}^{1} \sum_{l=1}^{g} \frac{(O_{k,l} - E_{k,l})^2}{E_{k,l}} \sim \chi^2_{(g-2)}, \tag{13}$$

where $O_{k,l}$ is the observed number of observations in group l with $Y = k$, and $E_{k,l}$ is the expected number of observations in group l with $Y = k$.

(6) A hypothesis test using H is then conducted, testing:

H_0: The model fits the data, vs. H_A: The model does not fit the data.

Testing at the 5% significance level, if the p-value is small then we can state that the model does not fit the data well.

4. Results

To build our model, we follow the systematic approach outlined in §3, while also aiming to maximise model accuracy with the justifications discussed in §1.

4.1. **Step 1: Pre-modelling checks.** Since the response variable, Y, is binary, taking values 1 if a baby was readmitted to the neonatal unit, and 0 if not, we make the assumption that $Y_i|\boldsymbol{x}_i \sim$Bernoulli$(p_i)$, for the observations $\boldsymbol{x}_i$ of the explanatory variables. We will also start by using the canonical link function for Bernoulli, the logit. We now wish to check for multicollinearity in the explanatory variables. After calculating the generalised variance inflation factor for all the terms in an additive model which contains all the explanatory variables, we see that the value of $(\text{GVIF})^{\frac{1}{d}}$ is less than 5 for all explanatory variables and thus we can state that there is no strong evidence of multicollinearity in the data. As an example, the value of $(\text{GVIF})^{\frac{1}{d}}$ for the covariate *birth weight* was 3.52, which was the largest value for any covariate and is much less than 5.

Now we test if a transformation of any of the continuous explanatory variables would be more appropriate and lead us to a better model for this data. *Length of stay* had already been transformed to be log(*length of stay*) by Bowden & Whittaker (2005)[2], so we chose this transformation for our model. The only other continuous explanatory variable in the model was *birth weight*. We considered two univariate models, one fitted with *birth weight* and the other fitted with log(*birth weight*). Using a Hosmer-Lemeshow test, the p-value for the fit of the model with *birth weight* was 0.000019 while the p-value for log(*birth weight*) was 0.00085. While both clearly fit the data very well, there is stronger evidence for the untransformed *birth weight*.

4.2. **Step 2: Preliminary model selection.** Given our set of explanatory variables and their chosen transformations, we now wish to see which of them we want to include in our initial additive model. This step is of much greater importance when the number of explanatory variables is in the magnitude of the hundreds or thousands. Since we only have 10 explanatory variables it would not be impossible to include them all in our initial additive model, however, we include this step for methodological completeness.

We first create a series of 10 univariate models, one for each explanatory variable. We then use a Z-test to see if there is evidence to suggest that the response variable is unaffected by any of the explanatory variables. We test at the 25% significance level and find that all levels of the explanatory variable *Education* are strongly non-significant, except for the intercept, which is only just significant with a p-value of 0.197. Given that most levels of *Education* are strongly non-significant (p-values > 0.64) and the intercept is only just significant, we have chosen to leave it out of our initial additive model.

We can now create a an initial additive model. This model has the following form, with coefficient estimates and coding key provided in Table 2:

$$\begin{aligned}\eta_{\text{i}} =&\beta_0 + \beta_1 * x_{\text{cns1}} + \beta_2 * x_{\text{size1}} + \beta_3 * x_{\text{gest2}} + \beta_4 * x_{\text{gest3}} + \beta_5 * x_{\text{gest4}} + \beta_6 * x_{\text{gest5}} \\ &+ \beta_7 * x_{\text{bwt}} + \beta_8 * x_{\text{emp.m1}} + \beta_9 * x_{\text{emp.f1}} + \beta_{10} * x_{\text{los}} + \beta_{11} * x_{\text{sex1}} + \beta_{12} * x_{\text{accom1}},\end{aligned} \tag{14}$$

where the predicted probability of readmission $p_i = \frac{e^{\eta_i}}{e^{\eta_i}+1}$.

Coding	Coef. Estimate	Std. Error	Coding	Coef. Estimate	Std. Error
Intercept	-2.92	0.70	–	–	–
cns1	0.04	0.11	bwt	0.04	0.13
size1	0.17	0.12	emp.m1	-0.20	0.11
gest2	0.40	0.31	emp.f1	-0.44	0.19
gest3	0.13	0.33	los	0.75	0.12
gest4	0.44	0.37	sex1	0.27	0.11
gest5	0.54	0.46	accom1	-0.26	0.14

Table 2. The coefficient estimates and their standard errors for each of the terms in the additive model. Descriptions of each term and the terms that make up the intercept can be found in Table 5 in §6. Appendix.

Now that we have an initial additive model, we can use backwards elimination to simplify the model by removing terms that do not add a substantial amount of information. During this process we removed the following terms: *birth weight* then *cns*. None of the terms were found to adjust for any of the others in

the model, and all terms removed had analysis of deviance p-values $\gg 0.15$. We now have a preliminary main effects model of the following form, with coefficient estimates and coding key provided in Table 3:

$$\begin{aligned}\eta_{\text{i}} =&\beta_0 + \beta_1 * x_{\text{size1}} + \beta_2 * x_{\text{gest2}} + \beta_3 * x_{\text{gest3}} + \beta_4 * x_{\text{gest4}} + \beta_5 * x_{\text{gest5}} \\ &+ \beta_6 * x_{\text{emp.m1}} + \beta_7 * x_{\text{emp.f1}} + \beta_8 * x_{\text{los}} + \beta_9 * x_{\text{sex1}} + \beta_{10} * x_{\text{accom1}},\end{aligned} \tag{15}$$

where the predicted probability of readmission $p_i = \frac{e^{\eta_i}}{e^{\eta_i}+1}$.

Coding	Coef. Estimate	Std. Error	Coding	Coef. Estimate	Std. Error
Intercept	-2.80	0.63			
size1	0.17	0.12	emp.m1	-0.20	0.11
gest2	0.40	0.31	emp.f1	-0.44	0.19
gest3	0.13	0.33	los	0.74	0.11
gest4	0.45	0.36	sex1	0.28	0.11
gest5	0.59	0.39	accom1	-0.26	0.14

TABLE 3. The coefficient estimates and their standard errors for each of the terms in the preliminary main effects model. Descriptions of each term and the terms that make up the intercept can be found in Table 5 in §6. Appendix.

A visual inspection of *length of stay* plotted against the predicted probabilities shows a strong linear relationship, so again we do not have to worry about transforming this variable.

4.3. **Step 3: Interactions.** As stated in §3.3 we will be considering two methods for including interactions in our model. The first method is reminiscent of backwards elimination, in which we add all possible 2-way interactions into the model and use analysis of deviance to reduce the model. We have 7 explanatory variables in our model meaning we have 21 possible two-way interactions to consider. Iterating through the process of backwards elimination, we are left with 10 interactions, and a model which we will call final model 1. We did not feel that there were any borderline significant terms that should be included despite being non-significant. Since the form of the model is now quite unwieldy, with 33 terms, we will instead just list the interaction terms that were included: *size:emp.f, size:los, size:sex, size:accom, gest:emp.f, gest:los, gest:sex, gest:accom, emp.m:sex,* and *sex:accom.*

The alternative method is to create a series of 21 models each with one interaction term. Using analysis of deviance to compare each model to the preliminary main effects model, we find that the model with the interaction term *size:emp.f* was the only one that could not be reduced to the preliminary main effects model, testing at the 5% significance level. We add this interaction term into the preliminary main effects model and call this final model 2.

We now wish to investigate whether there is evidence that one final model is better than the other. Fortunately, in this case, the two models are nested, final model 2 is a simplification of final model 1, so we can use analysis of deviance to test if final model 1 can be reduced to final model 2. Analysis of deviance returns a p-value of 0.08463, suggesting that final model 2 is a valid reduction of final model 1. On top this we also know that final model 2 is the far more parsimonious model, and it is worth noting that some of the standard errors in final model 1 are anomalously large, and that almost all of its terms are non-significant at the 5% level under a Z-test. If we also consider some of the interactions that are included in final model 1, we can see that it suggests that the sex of the baby is related to whether the parents own their home, the gestation period of the baby, and the employment status of the mother, all of which seem very unlikely and would certainly lead to over fitting. For these reasons we take final model 2 to be our final model. If there was more ambiguity between models, then we could take all final models through to model diagnostics and choose the model that aligned most with our aims. Our final model takes the following form, with parameter estimates and standard errors found in Table 4:

$$\begin{aligned}\eta_{\text{i}} =&\beta_0 + \beta_1 * x_{\text{size1}} + \beta_2 * x_{\text{gest2}} + \beta_3 * x_{\text{gest3}} + \beta_4 * x_{\text{gest4}} + \beta_5 * x_{\text{gest5}} + \beta_6 * x_{\text{emp.m1}} \\ &+ \beta_7 * x_{\text{emp.f1}} + \beta_8 * x_{\text{los}} + \beta_9 * x_{\text{sex1}} + \beta_{10} * x_{\text{accom1}} + \beta_{11} * (\textit{size1:emp.f1}),\end{aligned} \tag{16}$$

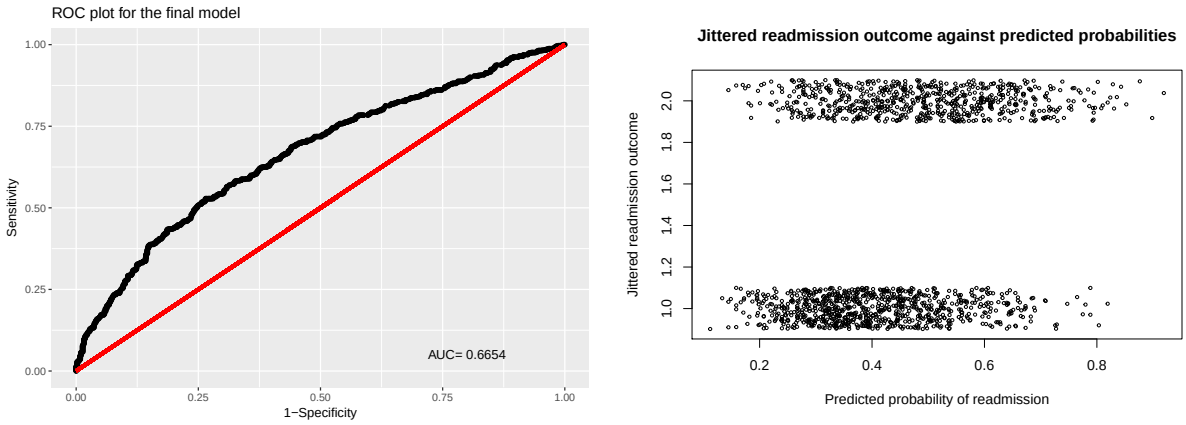
where the predicted probability of readmission $p_i = \frac{e^{\eta_i}}{e^{\eta_i}+1}$.

Coding	Coef. Estimate	Std. Error	Coding	Coef. Estimate	Std. Error
Intercept	-3.18	0.66	emp.m1	-0.20	0.11
size1	0.79	0.34	emp.f1	-0.01	0.29
gest2	0.38	0.31	los	0.74	0.11
gest3	0.13	0.33	sex1	0.27	0.11
gest4	0.44	0.36	accom1	-0.27	0.14
gest5	0.59	0.39	size1:emp.f1	-0.70	0.36

TABLE 4. The coefficient estimates and their standard errors for each of the terms in the final model. Descriptions of each term and the terms that make up the intercept can be found in Table 5 in §6. Appendix.

4.4. **Step 4: Model diagnostics.** Now that we have our final model we need to test it to see how well it fits the current data. We perform a Hosmer-Lemeshow test on final model which returns a p-value of 0.4239 suggesting strong evidence that the model fits the data, testing at the 5% significance level.

We can also use visual inspections of the model, such as those in Figure 1. From Figure 1a we can see that the area under the ROC curve is 0.6654 which suggests quite a poor fit (close to 1 is a perfect fit, close to 0.5 is equivalent to random assignment of response). Figure 1b shows us that the model has quite poor accuracy, given that most observations are centred around a probability of readmission of around 0.4 to 0.5. A model with better accuracy would be able to more definitively decide if a baby was going to readmit, and so the majority of readmitted babies would have a high probability, and the majority of non-readmitted babies would have a low probability.



(A) (B)

FIGURE 1. Visual inspections of the fit and accuracy of the final model: (A) A receiver operating characteristic (ROC) curve showing the true positive rate (sensitivity) against the false positive rate (1−specificity) at various threshold settings for the final model. Area under the curve (AUC) shows an approximate accuracy of the final model. (B) A jittered plot of readmission (not readmitted = 1, readmitted = 2) against predicted probability of readmission for the fitted final model.

Using cross validation, we calculate that the mean accuracy of the model for new data to be around 63%. The models mean true positive rate (sensitivity) is around 77.4%, however its mean false positive rate (1−specificity) is around 55.3%. We have assumed that the scale parameter, ϕ, is 1 when building all our models, we can test this by estimating the scale parameter for our final model by assuming a quasi distribution. The estimated scale parameter is 1.01 (very close to 1) so we do not need to change our assumption about the distribution used in the model. We can also test whether a different link function will improve the accuracy of our model. Using a probit link function, we find that there is a negligible increase in average accuracy and specificity, and a negligible decrease in sensitivity (all changes $< 1\%$ in magnitude). Similarly, using a complementary log-log link function, we get a negligible increase in accuracy and sensitivity, and a negligible decrease in specificity (all changes $< 1\%$ in magnitude). These changes are so small that they could easily be due to random fluctuations in the data, thus we choose to continue using the canonical link function, the logit.

Residual diagnostics were also investigated and revealed there may some issues with outliers, leverage and influence. Details of residual diagnostics and the issues they illuminate are detailed elsewhere [4;22].

4.5. **Interpretation of the final model.** Consider two babies, one male, one female. Both babies have a large NNU, a gestation period greater than 36 weeks, both parents employed who own their house, and a log(length of stay) value of 3.55. The predicted probability of readmission for the male is 48.4%, and for the female is 42.19%.

Consider another two babies, both male. Both babies have the same attributes as the male baby in the first example except for the log(length of stay) value. The first baby has a lower log(length of stay) of 3.05 and has a probability of readmission of 39.46%, the other baby has a higher log(length of stay) of 4.13 and has a probability of readmission of 58.85%.

5. Discussion

5.1. **The model.** As discussed in the §4.4 our model had a poor fit. The accuracy of the model overall for new data was around 63%, meaning that a baby was only correctly predicted to be readmitted or not 63% of the time. A model that only correctly predicts 50% of the time is considered equivalent to random assignment of readmission status, thus our model is only slightly better than random. This is very poor and we would ideally be aiming for at least an accuracy of around 80%-90%. On top of the poor overall accuracy, its mean false positive rate (the proportion of babies who will not be readmitted but are predicted to be) was around 55.3%. If all pre-term babies who are predicted to be readmitted are provisioned for, this will put an extremely large and unnecessary strain on hospital resources. The one redeeming factor of our model is that it wasn't awful at predicting when a baby needed to be readmitted to the neonatal unit, it got this correct 77.4% of the time. We tested a large number of models during our exploratory investigation, and all models were poor in their accuracy.

This suggests a few potential flaws in either the model building process or the study design, or both. A more thorough investigation, taking into consideration different techniques of variable selection, higher level interaction terms, other transformations of continuous explanatory variables, and a stronger consideration of the impact of outliers, leverage and influence may lead to a better fitting model. What is more likely, however, is that there are more factors that affect the probability of a baby being born prematurely than those considered in this investigation. For instance, there is evidence that race [3;20], race of mother [20], having a schedualed out patient visit or home visit within 72 hours of discharge [3], having a score ≥ 10 for "Neonatal Acute Physiology, Version II" [3], family income [13], birth facility [3] and geographic location [13], breastfeeding [13;20], being a first born [20], labour and delivery complications [20], being born by Cesarean section [13], and having a young mother [13] can strongly effect the probability that a baby will be readmitted within the first year of life. Also note that jaundice, respitory illness, and feeding problems are overwhelming reasons for readmission [3;13;17], and whether or not a hospital aims to combat these problems before they occur is another variable to consider. For instance it was found that home photo-therapy (a home treatment for jaundice) was the most important factor with respect to rehospitalisation amoung babies with a gestation period greater than 34 weeks [3]. Not including such explanatory variables in our model will mean a lot of the variance between individuals will remain unexplained, and thus our model will have poor prediction power.

If this model is to be used in practice, which we would not recommend, the user could adjust it to fit their circumstances. For instance, currently the accuracy of the model is evaluated on the assumption that a baby with a predicted probability of $\geq 50\%$ will be be readmitted, and those with a probability of readmission $< 50\%$ will not be. If they adjust these levels, they could capture more (or less if they wish) of the babies that will need to be readmitted and increase the true positive rate. This will, however, also increase the false positive rate; there will be more babies that are predicted to be readmitted when they won't actually be readmitted. We can see from Figure 1a that a large increase in the false positive rate ($1-$specificity) will lead to a much smaller increase in the true positive rate (sensitivity).

5.2. **Strengths and limitations of the study.** As previously mentioned, one limitation is the possibility of unexplained variance due to missing explanatory variables in the data. A strength of the study, however, is the large sample size of 1488 infants from a wide range of centres from all over the country. It helps to reassure that the poor fit of the model isn't due to small sample sizes, and that when an accurate model is built in the future, that it may be able to generalise well to any hospital in the country.

5.3. **Avenues for further research.** We suggest that further research is undertaken that includes other explanatory variables that are known to effect the probability of premature birth and readmission. When building a generalised linear model with these new explanatory variables it may be potent to consider additional model building techniques as detailed previously.

References

[1] Robert B. Bendel and A. A. Afifi. Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association*, 72(357):46–53, 1977.

[2] Jack Bowden and Joe Whittaker. A latent variable scorecard for neonatal baby frailty. *Statistical Modelling*, 5(2):159–172, 2005.

[3] G J Escobar, J D Greene, P Hulac, E Kincannon, K Bischoff, M N Gardner, M A Armstrong, and E K France. Rehospitalisation after birth hospitalisation: patterns among infants of all gestations. *Archives of Disease in Childhood*, 90(2):125–131, 2005.

[4] Julian James Faraway. *Extending the linear model with R [electronic resource] : generalized linear, mixed effects and nonparametric regression models*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, 2006.

[5] John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

[6] D. W. HOSMER, T. HOSMER, S. LE CESSIE, and S. LEMESHOW. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, 1997.

[7] David W. Hosmer and Nils Lid Hjort. Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*, 21(18):2723–2738, 2002.

[8] David W. Hosmer and Stanley Lemeshow. *Assessing the Fit of the Model*, pages 143–202. John Wiley & Sons, Inc., 2005.

[9] David W. Hosmer and Stanley Lemeshow. *Model-Building Strategies and Methods for Logistic Regression*, pages 91–142. John Wiley & Sons, Inc., 2005.

[10] D Langley, S Hollis, T Friede, D MacGregor, and A Gatrell. Impact of community neonatal services: a multicentre survey. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 87(3):F204–F208, 2002.

[11] RUTH M. MICKEY and Greenland S.A. The impact of confounder selection criteria on effect estimation. 129:125–37, 02 1989.

[12] March of Dimes. Premature babies. `https://www.marchofdimes.org/complications/premature-babies.aspx`, 2013. [Online; accessed 2017-11-19].

[13] Patricia Martens, Shelley Derksen, and Sumit Gupta. Predictors of hospital readmission of manitoba newborns within six weeks postbirth discharge: A population-based study. 114:708–13, 10 2004.

[14] P. (Peter) McCullagh. *Generalized linear models*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York, 1983.

[15] NHS. Premature labour and birth. `https://www.nhs.uk/conditions/pregnancy-and-baby/pages/premature-early-labour.aspx`, 2015. [Online; accessed 2017-11-19].

[16] S Petrou, T Sach, and L Davidson. The long-term costs of preterm birth and low birth weight: results of a systematic review. *Child: Care, Health and Development*, 27(2):97–115, 2001.

[17] Marco Pezzati. Hospital readmissions in late preterm infants. *Italian Journal of Pediatrics*, 40(2):A29, Oct 2014.

[18] P Royston, G Ambler, and W Sauerbrei. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5):964–974, 1999.

[19] Patrick Royston and Douglas G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3):429–467, 1994.

[20] Carrie Shapiro-Mendoza, Kay M Tomashek, Milton Kotelchuck, Wanda Barfield, Judith Weiss, and Stephen Evans. Risk factors for neonatal morbidity and mortality among "healthy," late preterm newborns. 30:54–60, 05 2006.

[21] WHO. Preterm birth. `http://www.who.int/mediacentre/factsheets/fs363/en/`, 2017. [Online; accessed 2017-11-19].

[22] Zhongheng Zhang. Residuals and regression diagnostics: focusing on logistic regression. *Annals of Translational Medicine*, 4(10), 2016.

6. Appendix

6.1. **R code.** The R code that was used to explore the data and run statistical analysis is detailed below.

```
1 #################
2 #Preamble
3 #################
```

Variable	Description	Levels (Coding)	Level Description	Intercept?
re.ad	Readmission (response variable)	N/A	No =0, Yes = 1	-
CNS	Was a CNS provided?	cns0	No	YES
		cns1	Yes	-
size	Size of NNU	size0	Small	YES
		size1	Large	-
gest	Gestation Period in Weeks	gest1	$<$ 26 weeks	YES
		gest2	26-29 weeks	-
		gest3	30-32 weeks	-
		gest4	33-36 weeks	-
		gest5	$>$ 36 weeks	-
bwt	Birth Weight	N/A	N/A	-
emp.m	Mother employed?	emp.m0	No	YES
		emp.m1	Yes	-
emp.f	Father/partner employed?	emp.f0	No	YES
		emp.f1	Yes	-
los	Time until initial discharge in days	N/A	N/A	-
sex	Sex of Baby	sex0	Female	YES
		sex1	Male	-
accom	Parents own house?	accom0	No	YES
		accom1	Yes	-

TABLE 5. The explanatory variables contained in the initial additive model (of which a subset were present in all subsequent models), their descriptions, and the levels of each that were incorporated into the intercept.

```
4
5  library("dplyr")
6  library("ggplot2")
7  library("car")
8  library("caret")
9  library("ResourceSelection")
10 library("Deducer")
11
12 ##########################
13 # Model Fitting Checks
14 ##########################
15
16 ##### Testing for Multicollinearity #####
17
18 vif(additive.model)
19 # Since all the GVIF values are less then 5 we say there collinearity is not an issue.
20
21 ##### Testing bwt vs. log(bwt) #####
22
23 univariable.bwt <- glm(re.ad~bwt, family = binomial(link = "logit"))
24
25 hoslem.test(univariable.bwt$y, fitted(univariable.bwt))
26
27 univariable.log.bwt <- glm(re.ad~log(bwt), family = binomial(link = "logit"))
28
29 hoslem.test(univariable.log.bwt$y, fitted(univariable.log.bwt))
30
31
32 ###################################
33 # Step 1: Univariate Modelling
34 ###################################
35
36
```

```
37 univariable.cns <- glm(re.ad~cns, family = binomial(link = "logit"))
38 summary(univariable.cns)
39
40 univariable.size <- glm(re.ad~size, family = binomial(link = "logit"))
41 summary(univariable.size)
42
43 univariable.gest <- glm(re.ad~gest, family = binomial(link = "logit"))
44 summary(univariable.gest)
45
46 univariable.bwt <- glm(re.ad~bwt, family = binomial(link = "logit"))
47 summary(univariable.bwt)
48
49 univariable.emp.m <- glm(re.ad~emp.m, family = binomial(link = "logit"))
50 summary(univariable.emp.m)
51
52 univariable.emp.f <- glm(re.ad~emp.f, family = binomial(link = "logit"))
53 summary(univariable.emp.f)
54
55 univariable.edu <- glm(re.ad~edu, family = binomial(link = "logit"))
56 summary(univariable.edu)
57
58 univariable.los <- glm(re.ad~los, family = binomial(link = "logit"))
59 summary(univariable.los)
60
61 univariable.sex <- glm(re.ad~sex, family = binomial(link = "logit"))
62 summary(univariable.sex)
63
64 univariable.accom <- glm(re.ad~accom, family = binomial(link = "logit"))
65 summary(univariable.accom)
66
67 # Thus we will leave out edu as all its levels have p-values >0.25
68
69 ###################################
70 # Step 2: Multivariate Modelling
71 ###################################
72
73 ##### Round 1 ######
74
75 drop1(additive.model, test="Chisq")
76
77 # Largest p-value is 0.83360 which means edu does not significantly add to the model
78
79 model.minus.edu <- glm(re.ad~cns+size+gest+bwt+emp.m+emp.f+los+sex+accom, family =
       binomial(link = "logit"))
80
81 anova(additive.model, model.minus.edu, test="Chisq")
82 # There is no significant difference in the models, so it is a valid reduction.
83
84 # This agrees with our univariate investigation
85
86 round(abs((coef(model.minus.edu)-coef(additive.model)[-(11:13)])
87           / coef(additive.model)[-(11:13)]), 3)*100
88 # None of the coefficient estimates change by more than 20% so education is not an
       important adjustment for the effect of other variables
89
90 # and so on...
91
92 ##### Round 4 ######
93
94 drop1(model.minus.edu.bwt.cns, test="Chisq")
95
```

```
# Largest p-value is 0.14911 which means size does significantly add to the model at the 15% sig. level

# We have now obtained a preliminary main effects model.

prelim.main.effects.model <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom, family = binomial(link = "logit"))

anova(additive.model, prelim.main.effects.model, test="Chisq")

# Not significantly different from the additive model so is a valid reduction.


###################################
# Step 3: Linearity Assumptions
###################################

# In the step, continuous variables are checked for their linearity in relation to the logit of the outcome.
# Our continuous variables are "los" only.

summary(neomod$re.ad)
pr <- fitted(prelim.main.effects.model)

scatter.smooth(los,log(pr/(1-pr)),cex=0.5)
# The smoothed scatter plot shows that the variable los are all linearly associated with readmission outcome in logit scale.


###########################################
# Step 4: Interactions amoung covariates
###########################################

# Add in all interactions

glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom, family = binomial(link = "logit"))

interactions.model <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+
                            size:gest+size:emp.m+size:emp.f+size:los+size:sex+size:accom+
                            gest:emp.m+gest:emp.f+gest:los+gest:sex+gest:accom+
                            emp.m:emp.f+emp.m:los+emp.m:sex+emp.m:accom+
                            emp.f:los+emp.f:sex+emp.f:accom+
                            los:sex+los:accom+
                            sex:accom,
                          family = binomial(link = "logit"))

summary(interactions.model)

anova(prelim.main.effects.model, interactions.model, test="Chisq")
# There is no significant difference in the models

##### Round 1 ######

drop1(interactions.model, test="Chisq")

# Largest p-value is 0.89074 which means size:emp.m does not significantly add to the model

int.model.1 <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+
                     size:gest+size:emp.f+size:los+size:sex+size:accom+
                     gest:emp.m+gest:emp.f+gest:los+gest:sex+gest:accom+
                     emp.m:emp.f+emp.m:los+emp.m:sex+emp.m:accom+
```

```
                        emp.f:los+emp.f:sex+emp.f:accom+
                        los:sex+los:accom+
                        sex:accom,
                      family = binomial(link = "logit"))

anova(interactions.model, int.model.1, test="Chisq")
# There is no significant difference in the models, so it is a valid reduction.

anova(prelim.main.effects.model, int.model.1, test="Chisq")
# There is no significant difference in the models, so this model is not significantly
    different to the preliminary one.

# and so on...

##### Round 10 ######

drop1(int.model.10, test="Chisq")

# Largest p-value is 0.47889 which means emp.f:los does not significantly add to the
    model (tested at the 5% level)

int.model.11 <-  glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+
                       size:emp.f+size:los+size:sex+size:accom+
                       gest:emp.f+gest:los+gest:sex+gest:accom+
                       emp.m:sex+
                       sex:accom,
                     family = binomial(link = "logit"))


anova(int.model.10, int.model.11, test="Chisq")
# There is no significant difference in the models, so it is a valid reduction.

anova(prelim.main.effects.model, int.model.11, test="Chisq")
# There is significant difference between the models at the 5% sig level, so the removal
    of the additional interaction terms is not valid.


final.model.1 <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+
                         size:emp.f+size:los+size:sex+size:accom+
                         gest:emp.f+gest:los+gest:sex+gest:accom+
                         emp.m:sex+
                         sex:accom,
                       family = binomial(link = "logit"))


############################################
# Step 4/2: Interactions amoung covariates
############################################

# We wish to test each interaction serparetely, and then include all those that have
    significant effects.

#### size:gest ######

uni.int.1 <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+
                           size:gest, family = binomial(link = "logit"))

summary(uni.int.1)

# None of the levels are significant so we do not include it in the model


```

```
##### size:emp.m ######

uni.int.2 <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+
                    size:emp.m, family = binomial(link = "logit"))

summary(uni.int.2)

# None of the levels are significant so we do not include it in the model

# and so on...

#Thus,

final.model.2 <- glm(re.ad~size+gest+emp.m+emp.f+los+sex+accom+size:emp.f, family =
    binomial(link = "logit"))

anova(prelim.main.effects.model, final.model.2, test="Chisq")
# There is significant difference between the models just above the 5% sig level,
# so the removal of the additional interaction terms is not valid.

anova(final.model.1, final.model.2, test="Chisq")
# There no is significant difference between the models at the 5% sig level,
# so the removal of the additional interaction terms is valid.

final.model <- final.model.2

##############################################
# Step 5: Assessing fit of the model
##############################################


hoslem.test(final.model$y, fitted(final.model))

# The P value is 0.4659, indicating that there is no significant difference between
    observed and predicted values. (tested at the 5% level)

Predprob <- predict(final.model,type="response")
plot(Predprob,jitter(as.numeric(re.ad),0.5),cex=0.5,ylab="Jittered readmission outcome")

rocplot(final.model)


##### Cross validation #####

#Randomly shuffle the data
Data<-neomod.factor[sample(nrow(neomod.factor)),]

#Create 10 equally size folds
folds <- cut(seq(1,nrow(Data)),breaks=10,labels=FALSE)

#Create data frame for results
accuracy <- data.frame(acc= 1:10, LCI=NA, UCI=NA, sens=NA, spec=NA)

#Perform 10 fold cross validation
for(i in 1:10){
  #Segement your data by fold using the which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)
  testData <- Data[testIndexes, ]
  trainData <- Data[-testIndexes, ]

  attach(trainData)
  final.model <- glm(formula = re.ad ~ size + gest + emp.m + emp.f + los + sex +
```

```
271         accom + size:emp.f, family = binomial(link = "logit"))
272   detach(trainData)
273
274   attach(testData)
275   testData$pred <- predict(final.model, testData, type="response")
276   detach(testData)
277
278   confusionMatrix <- confusionMatrix(round(testData$pred),testData$re.ad)
279
280   accuracy$acc[i] <- confusionMatrix$overall[[1]]
281   accuracy$LCI[i] <- confusionMatrix$overall[[3]]
282   accuracy$UCI[i] <- confusionMatrix$overall[[4]]
283   accuracy$sens[i] <- confusionMatrix$byClass[[1]]
284   accuracy$spec[i] <- confusionMatrix$byClass[[2]]
285 }
286
287
288 (sum(accuracy$acc)/10)*100 #avg. accuracy
289 # 62.97161%
290
291 (sum(accuracy$sens)/10)*100 #avg. sensitivity
292 # 77.43853%
293
294 (sum(accuracy$spec)/10)*100 #avg. specificity
295 # 44.69827%
296
297
298 ##### Inference #####
299
300 new.people <- neomod.factor
301 new.people[1,] <- c(1, 1, 5, 1.6, 1, 1, 2, 1, 3.55, 0, 1)
302 new.people[2,] <- c(1, 1, 5, 1.6, 1, 1, 2, 1, 3.55, 1, 1)
303 new.people[3,] <- c(1, 0, 1, 1.6, 0, 0, 2, 1, 3.05, 1, 0)
304 new.people[4,] <- c(1, 0, 2, 1.6, 0, 0, 2, 1, 3.05, 1, 0)
305 new.people[5,] <- c(1, 0, 3, 1.6, 0, 0, 2, 1, 3.05, 1, 0)
306 new.people[6,] <- c(1, 0, 4, 1.6, 0, 0, 2, 1, 3.05, 1, 0)
307 new.people[7,] <- c(1, 0, 5, 1.6, 0, 0, 2, 1, 3.05, 1, 0)
308 new.people[8,] <- c(1, 1, 1, 1.6, 1, 1, 2, 1, 4.13, 1, 1)
309 new.people[9,] <- c(1, 1, 2, 1.6, 1, 1, 2, 1, 4.13, 1, 1)
310 new.people[10,] <-c(1, 1, 3, 1.6, 1, 1, 2, 1, 4.13, 1, 1)
311 new.people[11,] <-c(1, 1, 4, 1.6, 1, 1, 2, 1, 4.13, 1, 1)
312 new.people[12,] <-c(1, 1, 5, 1.6, 1, 1, 2, 1, 4.13, 1, 1)
313 new.people[13,] <-c(1, 1, 5, 1.6, 1, 1, 2, 1, 3.05, 1, 1)
314 new.people[14,] <-c(1, 1, 5, 1.6, 1, 1, 2, 1, 4.13, 1, 1)
315
316 predict(final.model, new.people[1:14,], type="response")
```