Spring 2015

Exploring autism prediction through logistic regression analysis with corrections for rare events data

Jennifer Hunter

EXPLORING AUTISM PREDICTION THROUGH LOGISTIC REGRESSION ANALYSIS WITH CORRECTIONS FOR RARE EVENTS DATA

A Thesis

Submitted to the McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for the degree of

Master of Science in Computational Mathematics

By

Jennifer Hunter

May 2015

EXPLORING AUTISM PREDICTION THROUGH LOGISTIC

REGRESSION ANALYSIS WITH CORRECTIONS FOR RARE EVENTS DATA

By

Jennifer Hunter

Approved April 1, 2015

_____	_____
John C. Kern II, Ph.D. Chair, Associate Professor of Statistics Department of Mathematics & Computer Science (Committee Chair)	Frank J. D'Amico, Ph. D. Professor of Statistics Department of Mathematics & Computer Science (Committee Member)
_____	_____
James Schreiber, Ph. D. Professor of Education Department of Educational Foundations & Leadership (Committee Member)	Jeffrey Jackson, Ph.D. Professor of Computer Science Director of Graduate Studies Department of Mathematics & Computer Science

_____

James Swindal, Ph.D.
Dean, Professor of Philosophy
McAnulty College and Graduate School of Liberal Arts

ABSTRACT

EXPLORING AUTISM PREDICTION THROUGH LOGISTIC REGRESSION ANALYSIS WITH CORRECTIONS FOR RARE EVENTS DATA

By

Jennifer Hunter

May 2015

Thesis supervised by Dr. John Kern, Associate Professor, Department Chair

The study of rare events data in which observations of non-event outcomes far outnumber event outcomes makes inference under these circumstances quite difficult. Ideally, for a binary dependent variable, one would like sample data to contain enough observations from both outcome categories. With rare events data, however, this is usually impossible and/or costly to achieve with random sampling. This exploratory research aims to find a set of potential predictors that could be used to quantify a person's risk for developing autism spectrum disorder. A more efficient data collection strategy will be employed that allows for a smaller sample size of more meaningful data. Then, a statistical correction to the standard logistic regression model will be applied to yield adjusted predictions that take into account the prevalence of autism cases both in the sample data and in the population of interest.

DEDICATION

This thesis is dedicated to family who has always supported me in my pursuit of happiness and success. My parents, Ellen and Scott Hunter, have always believed in me to accomplish whatever I set my mind to, and I would not be who or where I am today without them. I would also like to thank my partner in crime and twin sister, Nicole, who has been by my side through all of the difficult and stressful adventures since moving to Pittsburgh with me in 2010.

In addition, I would like to say a special thank you to my Statistics professors, Dr. John Kern and Dr. Frank D'Amico, for inspiring my new found love of statistics and analyzing data to solve problems. I have always loved math, but it was not until I entered Duquesne that I realized how applicable and necessary data analysis is in today's society.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1: Introduction

1.1 Autism Spectrum Disorder

Found in only about 1% of the world's population (1 in 68 children in the US), according to statistics from the Center for Disease Control, Autism Spectrum Disorder (ASD) is a rare developmental disability that usually appears in early childhood. The specific signs and symptoms vary widely but include impaired social and communication skills and repetitive or stereotypical behaviors that can be hard to diagnose. Currently there is no medical test to diagnose the disorder, and diagnosis somewhere on the spectrum is usually dependent on the study/observation of a child's behavior and development by a medical professional and/or specialist (CDC, 2015).

This research used quantitative measures, mainly from a blood test, to craft an extensive list of explanatory variables comprised of various elements measured within a person's blood and hair. The goal was to find a "best-fitting" model to describe the relationship between ASD and some subset of the covariates using logistic regression described below. The ability to predict the probability (risk) of a child having ASD through a simple blood test could greatly contribute to the difficult diagnosis process currently in place.

1.2 Logistic regression

When modeling the relationship between a binary outcome variable and one or more independent predictor variables, logistic regression is the standard method of analysis. The dependent variable Y follows a Bernoulli distribution with parameter π that

takes on the value 1 with probability π and 0 with probability *1*- π. The probability for a given Y_i, also called its likelihood, is given by the following function:

$$P(Y_i|\pi_i) = \pi_i{}^{Y_i}(1-\pi_i)^{1-Y_i} \text{ for } Y = 0,1 \quad (1)$$

In any regression, one seeks to model the expected value of Y given the value(s) of the independent variable(s). Since Y is binary, the expected value of a (0, 1) variable is equivalent to the probability of Y taking on one of its two possible values, usually Y=1. In this study, let the rare event of interest be $P(Y = 1)$ and a non-event or control be $P(Y = 0)$. The expected value of Y_i in terms of the independent variables can be modeled by the probability form of the logistic regression formula

$$E(Y) = P(Y_i = 1|\boldsymbol{\beta}) = \pi_i = \frac{e^{x\beta}}{1+e^{x\beta}} \quad (2)$$

where $\boldsymbol{x\beta}$ is a vector of length n with $\boldsymbol{x} = (1, X_1, X_2, \ldots X_n)$ representing the independent variables along with their respective unknown parameters $\boldsymbol{\beta} = (\beta_o, \beta_1, \ldots \beta_n)$. The use of the logistic model ensures that the expected value is between (0,1) for any value of the domain $(-\infty, \infty)$. It also defines the relationship between π_i and the independent variable(s). In linear regression, the best fit line is determined by finding the values of $\boldsymbol{\beta}$ that minimize the sum of squared differences between the observed and predicted values of the model called residuals. However, in logistic regression the best fit "line" is determined by finding values of $\boldsymbol{\beta}$ that maximize the likelihood of obtaining the observed data. This requires changing the form of the likelihood function (1) for Y to one that is in terms of x and our unknown β parameter(s) from the logistic model.

To begin we will work with (1) and (2) to develop a likelihood function, *L*, given $Y_1, Y_2, \ldots Y_n$ independent, identically distributed Bernoulli random variables. Using (1)

this function is obtained by taking the product of the marginal distributions of all Y_i's in the sample as follows:

$$L(\pi|Y_1, Y_2, \ldots . Y_n) = \prod_{i=1}^{n} \ \pi_i{}^{Y_i}(1-\pi_i)^{1-Y_i}.$$

Next by taking logs, substituting (2), and using algebra, we can simplify the original likelihood function as follows (King, 2001, p.140):

$$ln(L(\pi|Y_1, Y_2, \ldots . Y_n)) = ln(\prod_{i=1}^{n} \ \pi_i{}^{Y_i}(1-\pi_i)^{1-Y_i})$$

$$= \sum_{i \in (Y_i=1)} \ln(\pi_i) + \sum_{i \in (Y_i=0)} ln(1-\pi_i)$$

$$= \sum_{i \in (Y_i=1)} \ln\left(\frac{e^{\boldsymbol{x_i\beta}}}{1+e^{\boldsymbol{x_i\beta}}}\right) + \sum_{i \in (Y_i=0)} ln\left(\frac{1}{1+e^{\boldsymbol{x_i\beta}}}\right)$$

$$\ln(\mathrm{L}(\beta|Y_1, Y_2, \ldots Y_n) = \sum_{i=1}^{n} Y_i(\boldsymbol{x_i\beta}) - ln\left(1+e^{\boldsymbol{x_i\beta}}\right) \qquad (3)$$

The above equation is the unconditional log-likelihood function used in logistic regression analysis to find the unconditional probability of obtaining the particular data set being studied. Maximum likelihood estimation (MLE) then fits the logistic model to the data by producing estimates for the unknown population parameters β, denoted $\hat{\beta}$, which maximize the log-likelihood (3). A computer software algorithm usually performs this iterative process since (3) is a nonlinear function with respect to β. The maximum log-likelihood value output from $ln(\mathrm{L}(\beta|Y_1, Y_2, \ldots Y_n)$ will also be used for statistical inference described later.

So how is logistic regression analysis affected when the data set under consideration is for a rare event? Well, as stated earlier, the goal of this study was to determine which subset of predictor variables, if any, is significantly related to the diagnosis outcome of ASD. Under normal circumstances where observations are collected totally randomly or randomly within strata defined by the independent

variables, $\hat{\beta}$ is consistent and asymptotically efficient (King, 2001, p.141). However, because ASD is a rare event, it can prove very challenging to collect meaningful data (i.e. enough 1's in the sample) by either method, leading to models where the probability of an event are underestimated and/or biased.

The statistical importance of having "enough" 1's in rare event sample data is illustrated by examining the variance matrix, $V(\hat{\beta})$, given by

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \pi_i(1-\pi_i)x_i' x_i\right]^{-1}$$

Specifically, the focus is on the factor $\pi_i(1-\pi_i)$. Typically for rare events, the estimates of $P(Y_i = 1|x_i) = \pi_i$are very small for all observations. If the logit model has some explanatory power, however, then the estimate for π_i for rare event cases ($Y_i = 1$) will likely be larger (closer to .5 because probabilities in rare event studies are usually very small) than estimates for non-event cases ($Y_i = 0$). This results in a larger $\pi_i(1-\pi_i)$ for 1's than for 0's and thus a smaller variance. The conclusion is that in the rare event circumstance, having additional ones is more informative because it leads to smaller variance (King, 2001, p.141).

To account for this challenge, a more efficient data collection strategy was implemented that allows for the collection of "enough" 1's to inform the model along with a correction to the logistic regression model. These two things eliminate the need for oversampling with rare events data. These will be discussed in detail in subsequent sections of this paper. Chapter 2 will deal with the data collection strategies and predictor selection. Then Chapter 3 will cover methods to develop a multivariable

logistic model and Chapter 4 will discuss how to correct such a model for rare events data based on the data collection strategy used.

Chapter 2: Data

2.1 Selection on Y

To prevent bias in an analysis, one wants to ensure that data are collected through random sampling so that the sample drawn is representative of the entire population of interest. A random sample can be obtained by selecting all observations at random or it can be a cross-sectional random sample where observations are selected at random within stratum defined by X depending on the goal of the study. For example, instead of randomly selecting 50 people from some population of interest, a cross sectional study would involve selection based on some independent variable say Gender. Then one would take a random sample of women and a random sample of men from the same population of interest. Both of these random sampling methods provide sufficient samples for statistical analysis.

Unfortunately when dealing with a population where the event of interest is extremely rare to observe, considerable time and money can be wasted trying to collect enough observations so that the sample includes both case (rare event) and control observations. In this study, the rare event would be persons with ASD. By randomly selecting within categories of the dependent variable Y, the data collection process is much more efficient. The sampling strategy is known as a case-control design. First, either randomly collect observations for the "cases" (Y=1), and then randomly select observations for the "controls" (Y = 0). Knowledge of the population fraction of ones, in this study is the average 1 in 68 ASD cases in the US, will also be used (King, 2001, p.141-142).

Case-control sampling is simple and straightforward to implement, but can have serious consequences if not conducted appropriately. The correction method that will be implemented based on this sampling design, called prior correction, requires observations for Y=1 and Y=0 to be independent random (or complete) selections. Second, attention must be paid to the sampling process to ensure that within the selection on Y we are not inadvertently introducing bias by selecting differently on X between the two groups. Careful attention must be paid to who is selected into the sample in order to control for inherent selection on the same set of explanatory variables. Finally, the trade-off between collecting more observations versus better or more explanatory variables must be addressed. In our case, our available explanatory variable list is an extensive one that is fairly inexpensive to collect (i.e. one blood/hair sample). The major cost in this study was from the collection of observations (patient participants), of which it was decided to collect an equal number of 1's and 0's. Since 1's contribute more "information" to the model in a rare event study it would be beneficial to collect as many 1's as possible and at least as many 0's as 1's. The optimal number of 0's to collect is situation dependent on the trade-off between collecting more observations and the value of the explanatory variables used (King, 2001, pp. 142-143)

2.2 Data Collection

Following guidelines for the case-controlled sampling design discussed above, the study consisted of data collected on 60 patients. Study participants were collected by the Children's Institute of Pittsburgh. The same set of tests was performed on each patient to officially diagnose ASD (or not). This diagnosis outcome is the dependent variable in

this study (AUT). Random selection was then made to obtain 30 participants within each of the two Y groups: the cases (AUT=1) and controls (AUT =0). The random selection of cases came from the first 30 newly diagnosed ASD patients from the start time of the study who also met certain criteria for participation established by the clinical professionals involved. Similarly, the 30 controls came from the first 30 volunteers who met the same participation criteria. Some of the participation criteria included conditions like age and gender to be matched between groups to control for bias between samples discussed in the previous section of this paper. The overall sample can be considered representative of the population of the local Western Pennsylvania region. The observations are independent in that no two subjects are related, thus the measures taken from the blood test from one subject have no effect on those from another subject. The independence and randomness (within Y) of the observations satisfy the conditions for inference in logistic regression. The third condition, linearity, concerns the appropriateness of the fit of the logistic model to the data and so will be addressed later.

Blood was drawn and a hair sample was taken from each of the 60 participants. Due to the limits/availability of different resources for analysis, the plethora of measures obtained from each patient's blood/hair came from the combined work of two different labs. Part of each sample was sent to the Quest Diagnostic laboratory and part of each sample was sent to the Duquesne University chemistry laboratory. The majority of the independent variables were created from measuring the levels of various elements (found on the periodic table) in a patient's blood. For each blood sample, these element levels were measured within each of three components of the blood: the red blood cells (RBC), the plasma (Pla), and the serum (Ser). Additionally, each element was measured using

two different methods for measurement established by the Environmental Protection Agency, denoted by a 60 or a 68. For example, the variable *LiRBC60* would be the measure of the element Lithium within the red blood cells measured using method 60. Similarly, the variable *FeHair* is the measure of the amount of iron in a patient's hair. There are other entities that were measured such as the number of natural killer cells (*NK*) in a patient's blood, which will only be elaborated on as needed.

2.3 Predictor Selection

The extensive list of predictors for use in this exploratory analysis begins with 257 possible variables. This number comes from first removing any "rater" variables used for the initial diagnosis of *AUT* =1 or 0, and then also removing any variables that had greater than 50 missing and/or zero-valued observations from the 60 total observations for each variable. In order to maintain as large a sample size as possible, it was necessary to remove any variables that would require excluding observations due to missing information. To address any concern for exclusion of important or clinically relevant variables, the assumption has been made that no single measure MUST be included in the final model due to clinical significance.

Next a univariate analysis was conducted for each of the 258 continuous predictors against the dependent variable, AUT, to see if a significant relationship exists. Assessing the fit of a model in the univariate case is equivalent to testing the significance of the estimated coefficient of the predictor in the model. It answers the question as to whether the model including the variable explains more about the outcome than the model without the variable. The Wald z-statistic produced by most statistical software,

tests the null hypothesis that the regression coefficient $\beta_1 = 0$. This test statistic, W_i, is found by dividing the estimated coefficient by its corresponding standard error (SE) as in the following formula:

$$W_i = \frac{\widehat{\beta}_i}{\widehat{SE}(\widehat{\beta}_i)}.$$

An alternative measure to test for model fit that will be discussed in the multivariate case is the likelihood ratio test (LRT). The standard level of significance is $\alpha = .05$, but due to the large number of independent variables alpha was further restricted to the $\alpha = .01$ level. Thus, a p-value of .01 or less for any coefficient suggests rejection of the null hypothesis and provides sufficient evidence that a significant univariate relationship exists. The univariate analysis resulted in 17 significant predictors at the $\alpha = .01$ level. Normally each individual predictor would have been checked for outliers and influential data that could unduly influence the regression coefficients and in turn the significance of the relationship. However, to again preserve sample size, the decision was made to also exclude variables that would only become significant after removal of influential observations. This allows our working sample size to remain as close to 60 observations as possible.

It is important to note that in the predictor selection process, the assumption of "linearity in the logit" was deemed true for each continuous predictor (this assumption is automatic for a nominal predictor). The probability form of the logistic model (2) presented in Chapter 1 commonly has an "S"-shaped curve when viewed graphically. It can be shown through a log transformation of the dependent variable and some algebra that an equivalent linear (logit) form of the logistic model is

$$log\left(\frac{\pi}{1-\pi}\right) = x\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n,$$

which transforms the left-hand side of the equation from $\pi = P(Y = 1)$ to $\log(odds)$. The odds that Y=1 is the ratio of the probability that Y = 1 over the probability that Y= 0. The log referred to here is the natural log, and these two forms of the logistic equation are equivalent and reversible. Since statistical analysis is output in terms of the logit form it just takes some "untransforming" to convert the log(odds) back to the more useful and desired probability. Similar to linear regression, the necessary assumption of linearity in the logit for a continuous predictor allows us to interpret the regression coefficient(s) as a rate of change. The only difference here is that a one unit change in x gives the change in log(odds) for Y. This assumption will be verified later as part of regression diagnostics of the multivariable model.

The 17 predictors significantly related to AUT in the univariate case were derived solely on p-value significance. The univariate model for each predictor was then checked for influential observations that could cause the relationship with AUT to become insignificant if influential data was removed. Similar to linear regression, there are several logistic regression diagnostics that can be used to detect influential data. Common among literature (Hosmer & Lemeshow, 2000; Menard, 1995; Pregibon, 1981) the following diagnostics applied to logistic will be discussed to detect influence: Standardized Pearson residuals, leverage (hat) values, and Cook's distance.

First, residuals can be defined as error estimates used to identify cases for which the model is a poor fit (i.e. a large discrepancy between the observed and predicted values for particular observation). Unlike linear regression, the error variance in logistic regression is dependent on the conditional mean of Y and so must be standardized by

adjusting each residual by its (binomial) standard error. The Pearson residual for the j[th] observation is calculated as follows (Menard, 1995, p.72):

$$r_j = \frac{P(Y_j = 1) - \hat{P}(Y_j = 1)}{\sqrt{\hat{P}(Y_j = 1)[1 - \hat{P}(Y_j = 1)]}}.$$

The standardized Pearson residual is then

$$r_{sj} = \frac{r_j}{\sqrt{1 - h_j}},$$

where h_j is called the leverage or hat value for the j[th] observation. We can assume these residuals have an approximate mean of zero and standard deviation of one so any observation with a residual value outside of (-2, 2) should be given closer inspection.

Second, leverage values can be interpreted as the distance an observation x_j is from the mean of the data in the covariate space. The farther the distance, the greater potential for influence on the slope parameters of the regression line. Leverage values are derived from the diagonal of the "hat" matrix and can range from 0 to 1. The hat matrix maps the vector of observed values Y_i onto the vector of fitted values $\widehat{Y}_j$ in the covariate space. Each matrix value, h_{ij}, quantifies the influence of an observed Y_i in the sample on that of a fitted value $\widehat{Y}_j$. Since the hat matrix is symmetric let $h_{jj} = h_j$ represent any diagonal element of the matrix. Similar to linear regression, it can be shown that h_j corresponds to the influence of Y_j on the fitted values in the sample across all observations. In general, a value larger than two times the average leverage $2(p + 1)/n$, where p is the number of parameters in the model and n is the sample size, should be inspected for influence. In this study, two times the average leverage would be $\frac{2(1+1)}{60} = .0\overline{6}$.

Finally, the last diagnostic tool used for detecting influence in this study combines the first two diagnostics. Cook's distance combines the discrepancy and leverage of an observation j to give a measure of the overall impact of that particular observation on the entire set of regression coefficients. This "distance" is the difference between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}_{-J}}$ where $\widehat{\boldsymbol{\beta}_{-J}}$ is the vector of coefficient estimates with the j[th] observation deleted from the analysis. Hosmer and Lemeshow (2001) give Cook's distance for logistic regression as

$$\Delta\widehat{\boldsymbol{\beta}_J} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}_{-J}} = \frac{r_{sj}^2 h_j}{(1-h_j)}$$

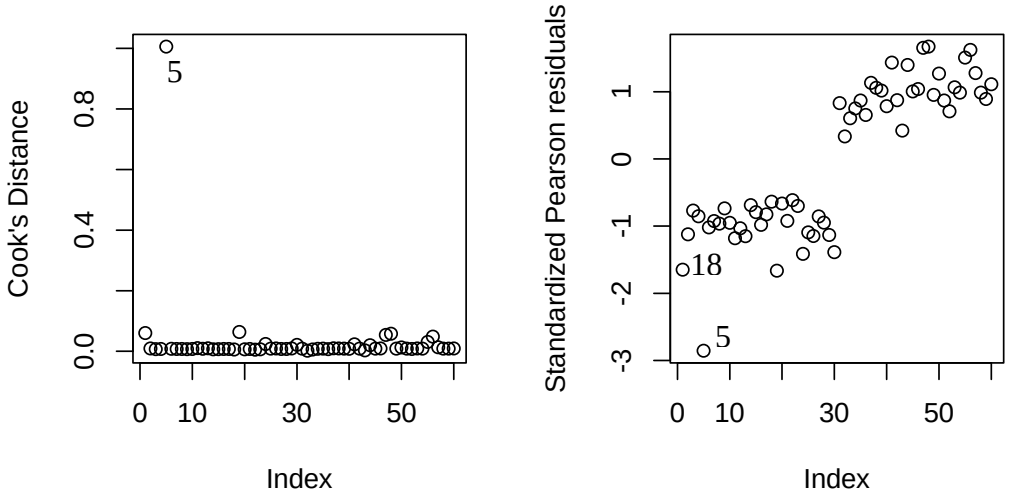
which incorporates the values of the Pearson standardized residual and the leverage value for the j[th] observation. A common alternative to using the Cook's distance is its standardized version called *dfbeta*, which alters the above formula by squaring the denominator. Any Cook's distance larger than 1 will be considered influential and require further inspection of influence on regression coefficient estimates.

The above diagnostics applied to logistic regression were obtained for each of the 17 significant predictors in the study. For any predictor with diagnostic values exceeding limits defined above, especially Cook's, diagnostic plots were obtained to confirm outliers. The logistic model was then refit with influential case(s) removed to assess influence on the regression coefficients. Due to the possibility that additional influential data points could be masked by other influential data, this process was repeated until no influential data was observed.

To illustrate the diagnostic process, the variable *SrSer60* will be used as an example. Diagnostics revealed case 5 of *SrSer60* to have a high Cook's distance (1.006)

and Pearson residual (-2.856), and Figure 2.1 shows diagnostic plots that confirm this fact.

Figure 2-1: Diagnostic plots for Univariate Model with SrSer60



Indeed refitting the univariate model with and without case 5 revealed a significant change in β_1 from 0.07125 to 0.11427 with p-values of 0.00368 and 0.000305 respectively. Repeating the set of diagnostics revealed case 18 to have a high residual that was masked by case 5, however, there was no significant change to the model fit with case 18 removed. Therefore, only case 5 will be removed *for SrSer60*. A similar procedure was followed for the remaining variables though no other cases meriting removal for any other variable.

The final step in the predictor selection process was to remove any predictors that are highly correlated with other predictors to prevent possible collinearity in the multivariate model. Examination of the correlation matrix of the 17 predictors revealed several highly correlated variables. It is easy to see why elements such as *LiRBC60*, *LiPla60,* and *LiSer60* are highly correlated with all pairwise correlations greater than .9. As described in the previous section, these are all measures of the same element just in a different part of the blood so one would expect the measures to be comparable. A similar

argument can be made for *MgPla60* and *MgRBC60* as well as *SrPla60* and *SrSer60*, which also have pairwise correlations greater than .9. Of the three sets, the variable with smallest p-value from the univariate analysis will be chosen for inclusion in initial model. The removal of 4 more variables leaves a total of 12 continuous predictors for inclusion into the multivariable model to predict AUT. Any other high correlations between variables will be addressed as needed in the model selection process. Results from the univariate analysis for each potential predictor variable are illustrated in Table 2.1.

Table 2-1: Results of Univariate Logistic Regression Model

	Variable	$\widehat{\beta}$	Std. Error	$\widehat{OR}$	95% CI for OR	Wald-z	p-value
1	NK	0.00439	0.00148	3.32	(1.65, 8.23)	2.963	0.0030
2	MMAcid	0.01857	0.00642	4.88	(1.85, 16.17)	2.895	0.0038
3	LaRBC60	8.8142	2.9665	3.4	(1.65, 8.35)	2.972	0.0030
4	PrRBC60	24.47695	8.4137	2.69	(1.45, 5.59)	2.909	0.0036
5	NdRBC60	7.5029	2.1204	7.56	(2.69, 26.11)	3.538	0.0004
6	DyRBC60	24.0254	8.5914	2.66	(1.41, 5.59)	2.796	0.0052
7	LiPla60	0.0394	0.0125	3.45	(1.74, 8.51)	3.138	0.0017
8	MgPla60	0.0002	0.0001	3.44	(1.71, 8.33)	3.114	0.0018
9	GaPla60	0.2282	0.0731	3.14	(1.63, 6.99)	3.122	0.0018
10	GdPla60	26.9038	8.6136	3.58	(1.74, 8.72)	3.124	0.0018
11	SrSer60	0.11427	0.03165	4.75	(2.22, 12.29)	2.903	0.0003
12	TiSer68	0.0815	0.0251	5.2	(2.11, 15.71)	3.243	0.0012

*Note: SrSer60 and MMAcid statistics based on n=59 observations.
*OR and 95% CI are for standardized regression coefficients (not shown in table)

2.4 Interpreting Logistic Regression Coefficients

In simple linear regression, the slope coefficient $\widehat{\beta}_\iota$ for an independent variable indicates the rate of change of Y for every one-unit change in x. The sign and magnitude of the slope coefficient illustrates the effect an independent variable has on the response variable. Thus, comparing the slope coefficients of different independent variables for a

univariate regression on the same response variable allows for the comparison of the effects of each predictor on the response. As previously explained, in logistic regression the logit model expresses the change in log(odds) for success (Y=1) to failure (Y=0) for the response as a linear function of a one-unit change in the independent variable. Of course, interpretation here relies on the assumption of linearity in the logit for a continuous covariate. Since interpretation of odds is easier than log(odds), exponentiation of both sides of the logit form of the model gives an expression of the odds of the response, here having ASD, in terms of some x as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \log(odds_x) = \beta_0 + \beta_1 x$$

$$e^{\log(odds_x)} = e^{\beta_0+\beta_1 x}$$

$$odds_x = e^{\beta_0+\beta_1 x}.$$

In order to compare how the odds of having ASD change for a given predictor, a statistic called the odds ratio (OR) is used in logistic regression. The OR for a continuous predictor is constant and gives the ratio of the odds of an event occurring for a one-unit change in x by the following formula:

$$OR = \frac{odds_{x+1}}{odds_x} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}.$$

For a general example, an OR = 2 means that the odds of Y occurring are increased by a factor of 2 for every unit increase in the predictor. Likewise, an OR = -.5 means that the odds of Y occurring decrease by half for every unit increase in the predictor. For continuous covariates, the idea of a meaningful one-unit increase must be considered when there are multiple covariates measured on different scales and/or with different levels of precision. The descriptive statistics for the 12 predictor variables chosen for the

multivariable model selection process are given in Table 2.2 to illustrate the wide range of scales for the different covariates.

Table 2-2: Descriptive Statistics for Continuous Predictors

	Variable	Description	n	Range(Min, Max)	Mean(St. dev)
1	NK	Natural killer cells	60	1277(7,1284)	353.03(273.3)
2	MMAcid	Methylmalonic acid	59	571(77,648)	160.81(85.6)
3	LaRBC60	Lanthanum	60	.63(0, .63)	.15(.14)
4	PrRBC60	Praseodymium	60	.17(0, .17)	.04(.04)
5	NdRBC60	Neodymium	60	1.85(0,1.85)	.23(.27)
6	DyRBC60	Dysprosium	60	.2(0, .2)	.03(.04)
7	LiPla60	Lithium	60	97.5(0,97.5)	20.43(31.5)
8	MgPla60	Magnesium	60	27426.1(18617.5,46043.6)	25550(5917)
9	GaPla60	Gallium	60	23(3.02, 26.02)	9.45(5.01)
10	GdPla60	Gadolinium	60	.21(0,.21)	.033(.047)
11	SrSer60	Strontium	59	63.87(12.2,76.1)	34.95(13.66)
12	TiSer68	Titianium	60	131.99(54.88,186.87)	89.27(20.28)

*All values rounded to the nearest hundredth.

Table 2-2 makes it easy to see, for example, that a one-unit increase in *NK* is very different from a one-unit increase in *PrRBC60*. Therefore, in order to directly compare the effect of each of these predictors with respect to *AUT*, a standardization of each predictor variable is used to compute their respective ORs shown in Table 2-1. Note the corresponding standardized coefficient estimates used to calculate the OR are not those shown above in Table 2-1. Predictors were standardized prior to regression by subtracting their respective means and then dividing by their respective standard deviations. The OR for the standardized coefficients then associates the change in the odds of Y occurring for every one standard deviation change in the predictor variable. Menard (1995) gives evidence for why a one standard deviation change is considered a large enough unit to show an effect, if any, exists on the dependent variable through

Chebycheff's Inequality Theorem. This theorem supports that even for a very nonnormal distribution, at least 93.75% of all cases should lie within 8 standard deviations of the mean and at least 96% within 10 standard deviations. The standardization of predictors to common units allows for a ranking of the effects of the predictors on the dependent variable (Menard, 1995, pp. 44-48). Therefore, even though all 12 predictors are strongly associated with *AUT* as evident by their p-values, the standardized ORs given in Table 2-1 show their relative impact on the risk of *AUT*. For example, the largest impact came from a one standard deviation in *NdRBC60*, periodic element Nd in the red blood cells, which increases the risk of *AUT* by a factor of about 7.5. The interpretation of regression coefficients in the multivariate model is usually of more importance than the univariate case because the effects of the estimated coefficients are adjusted for all other variables also included in the model.

Chapter 3: Methods

Once the clinically and/or statistically relevant variables have been chosen for inclusion in the multivariate analysis, it is time to begin building. Several model selection methods exist for building a model or models. Two computer based methods will be used here as guides in this exploratory analysis. Although the initial model building will be done through automated techniques, the results will not be considered definitive. The goal is to find the set of variables that result in the most parsimonious model within the constraints of the data being studied. This chapter will focus on stepwise and best subsets methods for model selection carried out in R and Jmp statistical software.

Before discussing the model selection techniques it is important to understand the test of significance used for the multivariable model. Given a set of n independent variables, let the logit form of the multiple logistic regression model be given by

$$g(x) = \boldsymbol{x\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

where $g(x)$ denotes the log(odds) of the probability π which can be written in probability form as

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}.$$

The MLE procedure produces a log-likelihood estimate of the fitted model that is used to assess its significance. The likelihood ratio test (LRT) relies on the comparison of the deviance between two models. The deviance, similar to the residual sum of squares used in linear regression, compares the difference between the observed and fitted values of a given model and is calculated by

$$D = -2\ln(likelihood\ of\ fitted\ model).$$

The log-likelihood is multiplied by -2 so that the quantity follows an approximate chi-square distribution from which hypotheses can be tested. In the univariate case, the LRT could have been alternatively used to test variable significance by comparing the deviance ratio of the model with and without the variable using the test statistic

$$G = -2\ln\left(\frac{L_0(without\ variable)}{L(with\ variable)}\right) = -2[\ln(L_0) - \ln(L)]$$

where L_0 is the likelihood for the constant model with no variable and L is the likelihood for the fitted model. The test statistic here follows a chi-square distribution with 1 degree of freedom and tests the same hypothesis as the Wald statistic in Chapter 2

$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_A: \beta_1 \neq 0\ .$$

Two similar hypothesis tests can be carried out in the multivariate case using the LRT. The first being the test of overall significance of a multivariate model with n predictors as follows:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_n = 0 \quad \text{versus} \quad H_A: At\ least\ one\ \beta_i \neq 0$$

with test statistic

$$G = -2\ln(L_0) - (-2\ln(L))$$

where L_0 and L represent the likelihood for the constant and full models respectively and G now has n degrees of freedom. The second is the nested LRT that can be used to compare a nested pair of models. The general hypotheses here are

$$H_0: Reduced\ model \quad \text{versus} \quad H_A: Full\ model$$

with test statistic

$$G = -2\ln(L_{Reduced}) - (-2\ln(L_{Full}))$$

where $L_{Reduced}$ represents the likelihood of the smaller model with n_1 predictors and L_{Full} represents the likelihood of the larger model with $n_1 + n_2$ predictors with G having

$(n_1 + n_2) - n_1 = n_2$ degrees of freedom. The *p*-value from all three LRTs come from the upper tail of the χ^2-distribution with the degrees of freedom essentially equal to the difference in number of variables between the two models being compared. If G is statistically significant (*p*-value < .05) we can reject the null hypothesis and conclude that the information about the variable(s) being tested contributes more to explaining the outcome than a model without them (Cannon, Cobb, Hartlaub, Legler, Lock, Moore, Rossman, & Witmer, 2013, pp.480-485, 529). It is important to note that even if the overall model proves significant, it does not mean that ALL predictors contribute to the model. Individual Wald tests or LRTs should be conducted to test for the significance of a variable in the multivariable model adjusted for all other variables already being in the model.

3.1 Stepwise Logistic Regression

Stepwise logistic regression is a computer-controlled sequential model building technique that's usually implemented in one of two ways: forward selection or backward elimination. In backward elimination, one starts with the full model including all potential variables, and then looks for the variable(s) to remove based on some information criteria. This study compared information loss (AICc) at each variable removal/addition step with the goal of minimizing information loss as stopping criteria. Variables that provide the largest reduction in information loss are sequentially removed and models compared until all remaining variables are statistically significant and/or the stopping criterion threshold has been reached. The process for forward selection is

similar except that one would start with the constant model and sequentially add variables to the model that minimizes the AICc.

The statistical criterion used here as a stopping point to determine the "best" model was the Akaike Information Criteria (AICc or AIC depending on software) which estimates information loss for a given model. The two differ only slightly and are given by

$$AICc = 2k - 2\log(L) + \frac{2k(k+1)}{n-k-1} or\ AIC = 2k - 2\log(L)$$

where k is the number of parameters in the model, n is the number of observations, and L is the MLE value of the model. The *AICc* converges to the *AIC* value as n gets very large but for the sample is this study where n is relatively small compared to the available k the *AICc* will be used when possible. The "best" model would be the one with the minimum *AICc* value among all possible models. Based on the above equation, the *AICc* value is a balance between the deviance or fit of the model (-2log(L)) and the number of parameters in the model. The *AICc* just has more of a penalty for extra parameters, which prevents over-fitting. Stepwise regression will be performed in both JMP and R software to account for any possible differences in programmed calculations within the software.

Though this deterministic method is criticized for its reliance purely on statistical criteria, there are benefits to it for the type of predictive exploratory analysis being done in this study. In this case there are many covariates (relative to the small number of observations) of which clinical importance and association with the outcome variable is unknown. Stepwise regression allows for screening of numerous covariates and the comparison of many different models simultaneously with ease (Menard, 1995, pp.54-55). Both stepwise simulations would produce the same result in a perfect world.

Realistically though, they are dependent upon their starting points and stopping criterion, and neither method may converge to the actual optimal model. One method could uncover a relationship that another may have missed so both will be run and results compared since they are easy to implement.

3.2 Best Subsets Logistic Regression

An alternative to stepwise regression is the use of best subsets regression. Any software implementing this method will consider all possible models containing from one to n parameters (the n parameter model being the full model) and return a specified number of "best" models. Each model is given a weight determined by some information criteria such as the *AICc* in this study, and then all models are compared and ranked based on their *AICc* value. Unlike stepwise regression, all possible models are considered for simultaneous comparison instead of just a subset of the possible models. Also, since all possible models are included in this method, one can be sure that the actual "best" model based on the information criteria will be found. Two R packages that implement best subsets regression that were used are the *bestglm* and *multiglm* packages (McLeod & Xu, 2014; Calcagno, 2013). To delve deeper into the specifics behind one way to implement best subsets or stepwise logistic regression, the reader could consult Homser & Lemeshow (2000, pp.116-135).

Chapter 4: Results

4.1 Multivariate Model Selection

The following tables illustrate the results of implementing both stepwise and best subsets regression methods to develop initial multivariate models for prediction of AUT. Both JMP and R software were used to compare stepwise regression model selection output as shown in Table 4-1 below. Then R was used to implement two different best subsets regression methods to compare their model selection output as in Table 4-2.

Table 4-1: Results of Applying Stepwise Variable Selection Using AICc

JMP Models	*Variables*	*n*	*G*	*p*	*AICc*
Full	All 12 variables	58	-	-	53.64
Forward	MMAcid+LaRBC60+SrSer60+TiSer68*	58	9.34	0.31	39.87
Backward	NK+LaRBC60+LiPla60+GaPla60+SrSer60+TiSer68	59	6.39	0.38	41.96
R Models	**Variables**	**n**	**G**	**p**	**AICc**
Full	All 12 variables	58	-	-	53.64
Forward	MMAcid+LaRBC60+NdRBC60+SrSer60+TiSer68	58	7.88	0.34	40.89
Backward	NK+MMAcid+LaRBC60+LiPla60+GaPla60+SrSer60+TiSer68	58	4.13	0.53	42.44

*Indicates "best" model

In both programs, the forward selection stepwise method resulted in a model with a lower AICc than backward elimination and neither program resulted in convergence to the same model between methods. The $G = -2ln(L_{reduced} - L_{full})$ test statistic for each model is the nest LRT for the variables excluded from each model respectively. The insignificance of the p-values for G for each model supports the exclusion of the remaining variables from the full model. Also, all models included the predictors *LaRBC60, SrSer60, TiSer68* suggesting their importance in the multivariate model.

Based on summary results from Table 4-1, the forward selection model output from JMP would be considered the "best" model based on minimum AICc criterion generated through stepwise regression.

Table 4-2 presents the result of the five "best" models by AIC(c) selected through best subsets regression using *bestglm* and *multiglm* packages in R. It turned out that both packages returned the same five models, the only difference being their rank due to *bestglm* using AIC as the selection criterion and *multiglm* using AICc as the selection criterion. Similar to stepwise regression models, all models here also have insignificant tests for variables excluded from each model. They all also contain the same three predictors mentioned above in addition to MMAcid, which also appears in all models. According to summary results from Table 4-2, Model 2 would be selected as the "best" model based on minimum AICc by best subsets regression.

Table 4-2: Results of Applying Best Subset Regression in R using AIC(c)

Models	Variables	n	G	p	AIC (bestglm)	AICc (glmulti)
Full	All 12 variables	58	-	-	47.088	55.361
1	NK+MMAcid+LaRBC60+SrSer60+TiSer68	58	5.54	0.59	38.625	40.272
2	MMAcid+LaRBC60+SrSer60+TiSer68*	58	7.89	0.44	38.982	40.136
3	NK+MMAcid+LaRBC60+DyRBC60+SrSer60+TiSer68	58	3.99	0.68	39.073	41.313
4	MMAcid+LaRBC60+NdRBC60+SrSer60+TiSer68	58	6.46	0.49	39.525	41.172
5	NK+MMAcid+LaRBC60+NdRBC60+SrSer60+TiSer68	58	4.59	0.60	39.637	40.791

*Indicates "best" model

Overall, though best subsets regression seems to outperform stepwise regression based on the average AICc of the models selected, both methods coincidentally produced the same "best" model for this study. Therefore, the initial multivariable logistic model is given by

$$g(x,\beta) = \beta_0 + \beta_1(MMAcid) + \beta_2(LaRBC60) + \beta_3(SrSer60) + \beta_4(TiSer68).$$

4.2 Goodness of Fit

As in the univariate case, the significance of each predictor needs to be verified in the multivariate model. As illustrated by Table 2-1, each of the four predictors for the initial multivariate model was shown to have a significant univariate relationship with AUT. Table 4-3 presents the results of fitting the multivariate model including a test of significance for each predictor while controlling for all other variables already in the model.

Table 4-3: Results of Fitting "Best" Multivariate Model by AICc

Variable	$\widehat{\beta}$	Std Error	$\widehat{OR}$	95% CI for $\widehat{OR}$	Wald-ChiSquare	P>ChiSq
Intercept	-28.405	9.31006	-	-	9.31	0.0023*
MMAcid	0.01272	0.00910	2.92	(1.15,19.73)	1.96	0.1620
LaRBC60	13.7877	5.24266	6.66	(2.02,39.56)	6.92	0.0085*
SrSer60	0.27834	0.09364	43.26	(5.66,906.84)	8.84	0.0030*
TiSer68	0.17482	0.06486	33.31	(3.98,766.06)	7.26	0.0070*

*OR and 95% CI are for standardized regression coefficients (not shown in table)

When compared to Table 2-1, the multivariate model shows significant but weaker associations for all predictors except for *SrSer60* when adjusted for other variables in the model. Significant relationships exist at the $\alpha = 0.05$ level according the

Wald Chi-Square test statistic (from Chapter 1 $(Wald\ z)^2 = Wald\ \chi^2$) for all variables except for *MMAcid* with p = 0.1620. On the contrary, the significance of the LRT for all variables including *MMAcid* supports the inclusion of all four variables in the multivariate model as seen in Table 4-4. Also, the relative effect of each predictor on *AUT* is again measured by the standardized ORs. The effect size of each predictor is different in order and magnitude in the multivariate model compared to that in the univariate model.

Table 4-4: Effect Likelihood Ratio Tests

Variable	***Nparm***	***DF***	***LR ChiSquare***	***P>ChiSq***
MMAcid	1	1	5.5088376	0.0189*
LaRBC60	1	1	11.388784	0.0007*
SrSer60	1	1	21.0586004	<.0001*
TiSer68	1	1	13.9865893	0.0002*

Both the LRT and the Wald Chi-Square test the null hypothesis that a given slope coefficient is zero. However, they now represent the comparison of the multivariate model with and without the variable in question so are conditional on the other variables in the model. The significance of the LRT for *MMAcid* provides evidence that the variable contributes unique information to the model not accounted for by the other variables. Therefore MMAcid will remain in the model. Additionally, examination of the coefficients for each predictor between the univariate and multivariate models reveals no drastic changes in coefficients. This is further support for the notion that the excluded variables are statistically insignificant in providing predictive information about the response variable, *AUT*.

In addition to the significance of predictors, the overall model goodness of fit indicates whether these predictors actually provide an effective model for predicting *AUT*. Note the main effects model with four predictors will be the final multivariate model after analysis of adding all pairwise combinations of interaction effects to the multivariate model yielded insignificant conclusions. The G statistic from Table 4-2 (G =7.89, p = 0.44) provided support for exclusion of variables from the full 12-variable model to the simpler model with four predictors. Also, the following output in Table 4-5 gives the overall LRT for the final model compared to a model with no predictors. The overall model hypothesis test of

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_n = 0 \quad versus \quad H_A: At\ least\ one\ \beta_i \neq 0$$

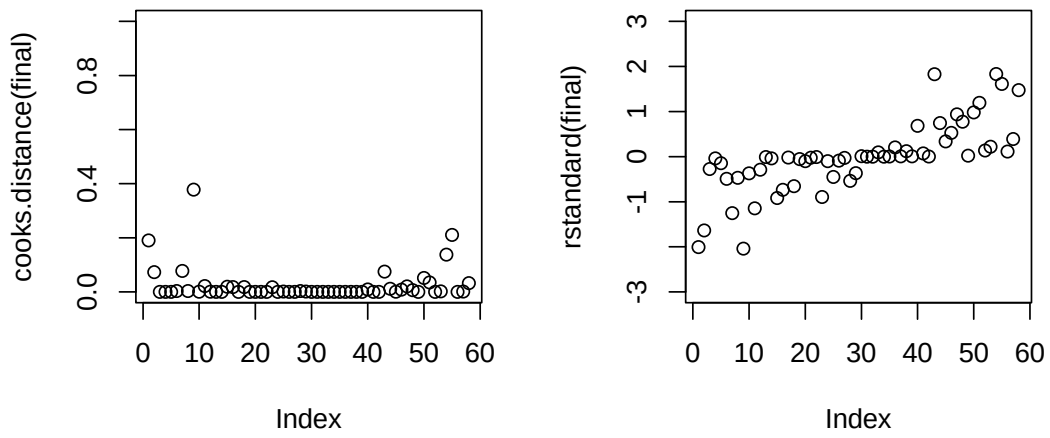
is significant at the $\alpha = 0.05$ level and it was shown in the previous section that each β_i can be concluded to be non-zero by LRT. Table 4-5 also gives the AICc reported earlier along with a pseudo R-Squared, which in logistic regression is the ratio of the Difference to Reduced -LogLikelihood values in the table. Similar to linear regression, this gives a measure of how well the model explains the variability in the response. The model can explain about 64% of the variability in *AUT*.

Table 4-5: Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	25.845351	4	51.6907	<.0001*
Full	14.357185			
Reduced	40.202536			
RSquare (U)		0.6429		
AICc		39.8682		
Observations (or Sum Wgts)		58		

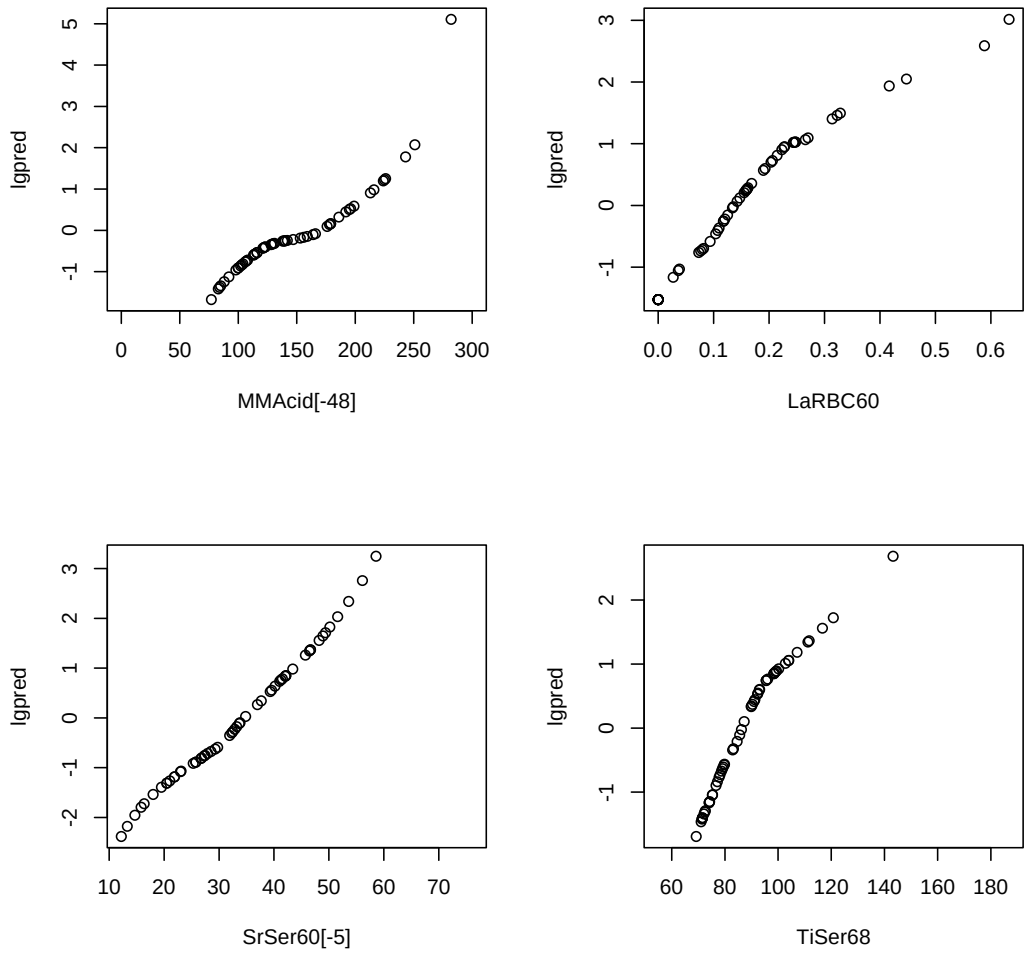
Also, similar to the univariate case, the fit of the multivariate model can be affected by an outlier or outliers. Hence, the same regression diagnostics were performed on the multivariate model revealing no serious influential cases as shown in Figure 4-1 below. The same diagnostics used to assess the univariate model fit also indicate the overall multivariate model is a good fit.

Figure 4-1: Diagnostic Plots for Multivariate Logistic Model



Lastly, for the defined final model, the logistic regression assumption of linearity in the logit for continuous covariates must be checked to determine an appropriate scale for the covariates. Remember this was assumed true for the purpose of univariate analysis. Hosmer & Lemeshow (2000) suggest multiple means of assessing this assumption including a univariable-smoothed scatterplot on the logit scale (p.99). Lowess smoothed plots for each of the four predictors are shown in Figure 4-2. It is reasonable to assume a continuous and linear relationship exists on the logit scale for each covariate.

Figure 4-2: Univariable Lowess Smoothed Logit Versus X



The final model can be expressed in both logit and probability forms respectively by

Logit Form

$$log\left(\frac{\pi}{1-\pi}\right) = -28.4 + 0.013(MMAcid) + 13.79(LaRBC60) + 0.29(SrSer60) + 0.17(TiSer68)$$

Probability Form

$$\pi = \frac{1}{1 + e^{-\left(-28.4+0.013(MMAcid)+13.79(LaRBC60)+0.29(SrSer60)+0.17(TiSer68)\right)}}$$

which can be used to find estimated probabilities or ORs for any given combination of covariate values.

Chapter 5: Discussion

5.1 Prior Correction

The logistic regression model for the AUT data given by

$$\pi = \frac{1}{1 + e^{-\left(-28.4+0.013(MMAcid)+13.79(LaRBC60)+0.29(SrSer60)+0.17(TiSer68)\right)}}$$

is based on a case-control data collection strategy which selected on the two groups of the dependent variable Y. For rare events data, this sample of 30 ASD cases (1's) and 30 non-ASD controls (0's) is far from a typical random sample one would expect to see when 0's dominate the population. Therefore, the model developed in this study is biased towards the proportion of 1's in the sample. King & Zeng (2001) illustrate that this experimental design can be consistent and efficient if the appropriate statistical corrections are made to the MLE estimates. They introduce an estimation method called prior correction, which involves computing the usual MLEs and applying corrections based on the proportion of 1's in both the sample and in the population of interest. The derivation of the method of prior correction for the logit model is presented in Appendix I. Prior correction for the logit model is easy to implement. As long as the functional form and explanatory variables are correct, and case-control sampling is done appropriately, the MLE for $\beta_{1,2\ldots.n}$ are statistically consistent estimates of the "true" slopes and only the estimate for the intercept, β_0, need be corrected (pp.143-144).

The sample proportion of 1's in this study is, $\bar{y} = .5$, and the population proportion of ones, denoted by τ, is taken to be the prevalence of ASD in the US. As mentioned in the introduction, $\tau = \frac{1}{68} \approx .015\ or\ 1.5\%$. Using this prior knowledge, the following corrected estimate for β_0 is given by

$$\hat{\beta}_0 - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

$$\hat{\beta}_0 - ln\left[\left(\frac{1-.015}{.015}\right)\left(\frac{.5}{1-.5}\right)\right]$$

which is equivalent to $\hat{\beta}_0$ only for the selection on Y sampling design discussed in this paper. This correction factor results in more realistic estimates of the probability of ASD for a given person in the population.

To illustrate the differences, Table 5-1 presents three examples from the sample data for persons with low, middle, and high predicted probabilities π for ASD (Y=1) based on the logistic regression model. Note Person 1 does not actually have ASD whereas Persons 2 and 3 do have ASD. The following output shows the difference in the original probability form of the logistic model with the form including prior correction for rare events data.

$$\boldsymbol{P(Y=1)} = \pi = \frac{1}{1+e^{-(x\beta)}} \quad \text{versus} \quad \boldsymbol{P(Y=1)} = \pi = \frac{1}{1+e^{-\left(x\beta - ln\left[\left(\frac{1-.015}{.015}\right)\left(\frac{.5}{1-.5}\right)\right]\right)}}.$$

Table 5-1: P(Y=1) for Person *i* Before and After Prior Correction

Person	x	Original $\boldsymbol{\pi_i}$	Adjusted $\boldsymbol{\pi_i}$	AUT
1	(100, 0.03644372, 29.29353, 79.70047)	0.0105358	0.0001621258	0
2	(121, 0.313854, 36.984340, 92.263)	0.5465779	0.01802621	1
3	(291, 0.1471794, 46.45581, 89.99044)	0.9798999	0.4260797	1

Chapter 6: Conclusion

6.1 Conclusion

Overall, this exploratory study has shown that there is potential for an adequate model based on quantitative predictors used to diagnose ASD in the future. This study was especially limited due to sample size, and even a large sample would still require a screening of the extensive list of predictor variables to identify potentially "important" predictors. Study of the various elements in the blood with a clinical professional is important in terms of finding an adequate starting point with which to begin model selection. The variable selection process used in this study gave an idea of the statistical association of each of the predictors with the dependent variable. The ability to obtain a larger sample would allow for the consideration of more predictors into the multivariate model.

6.2 Future work

The information gained in this study provides a basis from which to move forward with further research. Though the model obtained worked well for the current sample, there is a possibility that the variation observed within this sample is not a good representation of the population as a whole. A first step toward future work would be to validate this model repeatedly for other sample data to see how well it performs. Second, and possibly most important, would be to increase the sample size if possible so the model is not as sensitive to influential data and over-fitting. Outliers and influential data had much more of an impact in this study in regards to decision making to preserve sample size. The general rule of thumb in logistic regression is to have ten event cases per predictor variable included in the model. Thus the logistic model for this study may

be pushing the possibility for over-fitting with four predictors. Lastly, an in depth discussion with a clinical professional about the clinical importance of the variables both in the multivariate model and even those excluded from the model could have a significant impact in the model building process. The ability to provide a model for the prediction of ASD would have a huge impact on the current diagnosis procedures and this study provided a first look into how our blood may provide the clues necessary to solve the problem.

References

Calcagno, V. (2013). *glmulti: Model selection and multimodel inference made easy.* [Computer software]. Retrieved from http://CRAN.R-project.org/package=glmulti

Cannon, A. R., Cobb, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., … Witmer, J. A. (2013). *Stat2: Building models for a world of data.* New York, NY: W.H. Freeman and Company.

Centers for Disease Control and Prevention. (2015, January 2). *Autism Spectrum Disorder (ASD).* Retrieved from http://www.cdc.gov/ncbddd/autism/index.html

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression*. New York, NY: John Wiley & Sons.

King, G. & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2): 137-163. Retrieved from http://gking.harvard.edu

McLeod, A.I. & Xu, C. (2014). *bestglm: Best subset GLM.* [Computer software]. Retrieved from http://CRAN.R-project.org/package=bestglm

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Appendix I

In this Appendix, the method of prior correction discussed in this paper for the logit model is derived which results in a consistent, easy to apply correction for rare events logistic models. King & Zeng derive this correction factor for several different models, but not all are easily applicable. It is shown that prior correction gives estimates equivalent to maximizing the full information likelihood equation $P(Y, X|\beta)$ (2001, p.159-160).

Suppose X, Y are random variables, then x,y are random variables representing the case-controlled selection of ones and zeros from X,Y. The claim of prior correction is that $P(Y|X)$ can be estimated with an iid sample drawn from its own density $P(X, Y)$ or from $P(x, y)$ multiplied by some correction factor as shown in the general form below.

$$P(Y|X) = P(y|x)\left[\frac{P(Y)P(x)}{P(y)P(X)}\right]$$

Let D and d be the random samples of size n from $P(X, Y)$ and $P(x, y)$ respectively. In binary models using prior correction it is assumed that $P(Y = 1) = \tau$ and $P(y = 1) = \bar{y}$ are known. Thus the correction factors are

$$\boldsymbol{A_1} = \frac{\tau}{\bar{y}}, \boldsymbol{A_0} = \frac{1-\tau}{1-\bar{y}}, \text{ and } \boldsymbol{B^{-1}} = P(y = 1|x, d)\frac{\tau}{\bar{y}} + [1 - P(y = 1|x, d)]\frac{(1-\tau)}{1-\bar{y}}$$

We want to find $P(y = 1|x, d)A_1B$ where $P(y = 1|x, d) = \frac{1}{1+e^{-x\beta}}$ for the logit model which will be denoted P for simplification. Hence

$$P(y = 1|x, d)A_1B = \frac{P\left(\frac{\tau}{\bar{y}}\right)}{P\left(\frac{\tau}{\bar{y}}\right) + (1 - P)\left(\frac{1-\tau}{1-\bar{y}}\right)}$$

**Multiply top and bottom by $\bar{y}(1-\bar{y})$ gives

$$= \frac{P \cdot \tau \cdot (1-\bar{y})}{P \cdot \tau \cdot (1-\bar{y}) + (1-P) \cdot (1-\tau) \cdot \bar{y}}$$

$$= \left(\frac{P \cdot \tau \cdot (1-\bar{y}) + (1-P) \cdot (1-\tau) \cdot \bar{y}}{P \cdot \tau \cdot (1-\bar{y})}\right)^{-1}$$

$$= \left(1 + \left(\frac{1-P}{1}\right)\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right)^{-1}$$

$$= \left(1 + \left(\frac{1}{P} - 1\right)\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right)^{-1}$$

**Substitute back in for P gives

$$= \left(1 + \left(1 + e^{-x\beta} - 1\right)\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right)^{-1}$$

$$= \left(1 + e^{-x\beta} \cdot \left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right)^{-1}$$

$$= \left(1 + e^{-x\beta} \cdot e^{ln\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)}\right)^{-1}$$

$$= \left(1 + e^{-x\beta + ln\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)}\right)^{-1}$$

Recall that the probability form of the logistic model can be written as

$$\pi = \left(1 + e^{-x\beta}\right)^{-1}$$

which demonstrates that since the bias factor $ln\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)$ is a constant term, the MLEs of $\beta_{1,2,\ldots n}$ are not affected by the selection on Y sampling strategy. To correct for the bias that is added to the model by this sampling strategy, simply subtract the bias from the intercept term.