CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Stratum Estimates of the Probability of an Insurance Claim Based on Logistic Regression

AUTHOR: Kailey Pickles

SUPERVISOR: Patrick Farrell

DATE: August 21st, 2018

Table of Contents

Abstract

In this project, I seek to answer the question actuaries in the insurance industry have been trying to answer for years; “What is the probability a person will make a claim?”. This very important question is one of the biggest factors in determining insurance rates. Using the variables of age, gender, and marital status, the probability of a person making a claim will be estimated using logistic regression. Due to lack of available data, the observational data will need to be randomly generated based on already known probabilities. Using this generated data, logistic regression will be performed to find an equation that will predict the probability a person will make a claim.

Chapter 1: Introduction to Logistic Regression

Regression is an important tool for data analysis when trying to determine the relationship between one or more independent variables and a dependent variable. In this project, the case where the dependent (or response) variable is discrete (claim or no claim will be made) will be studied. In this situation, regular linear regression techniques will not properly fit the data so logistic regression will be used. The 'S' shaped curve of a logistic model is a much better fit for the relationship between a certain variable (or variables) and the probability of a success/failure than the straight line of a linear model. Logistic regression seeks to model the probability of an event occurring depending on the values of the independent variables. Also, it seeks to estimate the probability that an event occurs for a randomly selected observation and predicts the effect of a series of variables on a binary response variable. Statistical software will be used for creating the logistic regression equation however in this chapter, how the logistic equation was formed and how it is used to find a regression equation will be explored. This chapter will also look at different ways of analyzing the regression equation once it is found. To illustrate this, the case where there is only one independent variable will first be investigated before expanding into the case of multiple independent variables.

To begin, one first must understand the logistic equation. Under the assumption that each observed event in an experiment is independent, the dependent variable in logistic regression is a Bernoulli random variable with probability π_i. This is due to the fact that each event can be observed as either a success or failure because the dependent variable is discrete. Determining which events are consider a success is up to the user and what is being researched. For example, in this project a success will be if a person made a claim and a fail will be when they did not make a claim. In any case, letting n be the number of events and letting Y represent the dependent variable we have

$$Y_i \sim \text{Bernoulli}(\pi_i) \qquad (i=1,..,n)$$

where

$$f(y_i)= \pi_i^{\,y_i}(1-\pi_i)^{\,y_i}.$$

$Y_i = 1$ would be considered a success while $Y_i = 0$ would be considered a failure. By assumption, the Y_i's are independent, thus their joint density function is

$$f(y_1,\ldots,y_n)=\prod_{i=1}^{n} f(y_i).$$

Which can be written as

$$f(y_1,\ldots,y_n)=\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad (1)$$

where

$$\pi_i = \pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}. \quad (2)$$

The equation for π_i is very important in logistic regression as it is also equal to the probability of a success, $P(Y_i=1)$. This is the typical equation used when writing a logistic equation.

In linear regression, to find the unknown parameters (β_0,β_1) the least squares method would be used. Unfortunately, when this method is used with a discrete outcome the estimates are no longer valid. Thus, in the case of logistic regression, an approach called the Maximum Likelihood Method is used to estimate the unknown parameters in (2). For this method, the natural logs of both sides of (1) are taken and differentiated with respect to both parameters. Thus, yielding

$$\frac{d(\ln (f(y_1,\ldots,y_n))}{d\beta_o}=\sum_{i=1}^{n} y_i \left(\frac{d(\ln(\pi_i))}{d\beta_0}\right) + \sum_{i=1}^{n}(1-y_i)\left(\frac{d(\ln(1-\pi_i))}{d\beta_0}\right) \quad (3)$$

and

$$\frac{d(\ln (f(y_1,\ldots,y_n)}{d\beta_1}=\sum_{i=1}^{n} y_i \left(\frac{d(\ln(\pi_i))}{d\beta_1}\right) + \sum_{i=1}^{n}(1-y_i)\left(\frac{d(\ln(1-\pi_i))}{d\beta_1}\right). \quad (4)$$

From this, the derivatives of (3) and (4) can be calculated. First note that,

$$\frac{d\ln(\pi_i)}{d\beta_0} = \frac{(1+e^{\beta_0+\beta_1 x_i})e^{\beta_0+\beta_1 x_i}-(e^{\beta_0+\beta_1 x_i})^2}{(1+e^{\beta_0+\beta_1 x_i})^2}=\frac{e^{\beta_0+\beta_1 x_i}(1+e^{\beta_0+\beta_1 x_i}-e^{\beta_0+\beta_1 x_i})}{(1+e^{\beta_0+\beta_1 x_i})^2}$$

$$=\frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})} * \frac{1}{(1+e^{\beta_0+\beta_1 x_i})}$$

$$= \pi_i(1-\pi_i)\ .$$

Also note,

$$\frac{d(\ln(1-\pi_i))}{d\beta_0} = \frac{-1}{1-\pi_i}\left(\frac{d\pi_i}{d\beta_0}\right) = -\pi_i.$$

Thus, putting the two equations above into (3) yields

$$\frac{d(\ln(f(y_1,\ldots,y_n))}{d\beta_o} = \sum_{i=1}^{n} y_i(1-\pi_i) - \sum_{i=1}^{n} \pi_i(1-y_i).$$

Expanding this equation and doing some simple cancellations returns

$$\frac{d(\ln(f(y_1,\ldots,y_n))}{d\beta_o} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i \quad . \tag{5}$$

To maximize β_0 set (5) equal to zero. Similar steps can be taken to maximize β_1. These steps follow straight from the steps above to solve for (5) so they will be omitted. Maximizing β_1 yields

$$\frac{d(\ln(f(y_1,\ldots,y_n))}{d\beta_1} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \pi_i = 0. \tag{6}$$

From here, it becomes very difficult to evaluate by hand and software must be used.

Though it is hard to calculate by hand, the theory behind how the software solves for (β_0,β_1) can be explained. To get estimates for β_0 and β_1, statistical softwares use the Newton-Raphson Method. This method takes the Maximum Likelihood Equations ((5) and (6)) to form

$$q' = \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i, \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \pi_i\right)$$

and the variance-covariance matrix,

$$H = \begin{vmatrix} \frac{d^2(\ln(f(y_1,\ldots,y_n))}{d{\beta_o}^2} & \frac{d^2(\ln(f(y_1,\ldots,y_n))}{d\beta_o d\beta_1} \\ \frac{d^2(\ln(f(y_1,\ldots,y_n))}{d\beta_1 d\beta_0} & \frac{d^2(\ln(f(y_1,\ldots,y_n))}{d{\beta_1}^2} \end{vmatrix} = \begin{vmatrix} -\sum_{i=1}^{n} \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_i \pi_i(1-\pi_i) \\ -\sum_{i=1}^{n} x_i \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_i^2 \pi_i(1-\pi_i) \end{vmatrix},$$

to form an iterative procedure for estimating $\boldsymbol{\beta} = (\beta_0, \beta_1)$. An initial estimate for $\boldsymbol{\beta}$ is made to begin the procedure, this initial estimate is called $\boldsymbol{\beta}^{(0)}$. At the k^{th} step, we obtain $\boldsymbol{\beta}^{(k+1)}$ from

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - (H^{(k)})^{-1} q^{(k)}$$

where $H^{(k)}$ and $q^{(k)}$ are H and q evaluated at $\boldsymbol{\beta}^{(k)}$. This process is repeated until $\boldsymbol{\beta}$ converges to what is the maximum likelihood estimate used for β_0 and β_1. With $\boldsymbol{\beta}$, an estimate for π_i can be found

$$\widehat{\pi_i} = \frac{e^{\widehat{\beta_0} + \widehat{\beta_1} x_i}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1} x_i}}$$

resulting in an equation for the probability of a success. The matrix H above can be used to retrieve some other important information typically given by statistical software.

Taking $-H^{-1}$, the square roots of the diagonal elements of this matrix yields the standard errors of the estimates of β_0 and β_1.

After estimating the coefficients, assessing the significance of the variables in the model is the next step for any statistician. A significance test determines whether the independent variables have a significant effect on the response variable. This test is normally part of the output of any statistical software. There are two commonly used methods to determine significance; the P values and the chi-square hypothesis test. P values are the same in both types of regression analysis and therefore it will not be explored in this chapter. The chi-square hypothesis test is similar to the F test in linear regression. The comparison of observed to predicted values using the likelihood function is based on

$$G = -2\ln\left(\frac{(likelihood\ without\ the\ variable)}{(likelihood\ with\ the\ variable)}\right).$$

For the case where there is only one independent variable, the likelihood without the variable is easy to find. If x_i is not significant, $\beta_1=0$, so

$$\pi_i = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \pi. \tag{7}$$

Notice, $\frac{dln(\pi)}{d\beta_0} = \pi(1-\pi)$. Thus, the likelihood function for β_0 without x_i is the same as when x_i is included. Therefore,

$$\frac{d\left(\ln\left(f(y_1,\ldots,y_n)\right)\right)}{d\beta_o} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi = 0.$$

After substituting in (7) and rearranging the above equation, the equation for the sum of all the y_i's can be shown to be

$$\sum_{i=1}^{n} y_i = n\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right).$$

Rearranging and solving for β_0 yields,

$$\widehat{\beta_0} = \ln\left(\frac{\sum_{i=1}^{n} y_i}{n - \sum_{i=1}^{n} y_i}\right).$$

Now that an estimate for $\widehat{\beta_0}$ has been found, G can be solved. In the case with a single independent variable,

$$G = -2\ln\left(\frac{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^{\sum_{i=1}^{n} y_i}\left(\frac{n-\sum_{i=1}^{n} y_i}{n}\right)^{n-\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} \hat{\pi}_i^{\,y_i}(1-\hat{\pi}_i)^{1-y_i}}\right)$$

or,

$$G = 2\left(\sum_{i=1}^{n}[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)]\right) - \left[\left(\sum_{i=1}^{n} y_i\right)\ln\left(\sum_{i=1}^{n} y_i\right) + \left(n-\sum_{i=1}^{n} y_i\right)\ln\left(n - \sum_{i=1}^{n} y_i\right) - n\ln(n)\right]).$$

In the single variable case, the statistic will follow a chi-square distribution with 1 degree of freedom. If $P[\chi^2(1)>G]$ is less than a specified alpha, then the independent variable is significant. In the general case, G follows a chi-square distribution with m degrees of freedom where m is the number of independent variables.

Another important piece of information that can be found from the logistic regression equation is the odds ratio. Before the odds ratio for logistic regression can be derived, odds and the odds ratio must be defined. The odds of something occurring is defined as the probability of success divided by the probability of failure. Thus, if p represents the probability of an event occurring then

$$odds = \frac{p}{1-p}.$$

Therefore, using the estimating equation the odds that $Y_i = 1$, equivalently the odds that Y_i is a success, at x_i, which we will refer to as O, is equal to

$$O(x_i) = \frac{\pi(x_i)}{1-\pi(x_i)} = \frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_i}}} = e^{\beta_0+\beta_1 x_i}.$$

Notice that the odds of x_i+1 is

$$O(x_i+1) = e^{\beta_0+\beta_1(x_i+1)} = e^{\beta_0+\beta_1 x_i} e^{\beta_1}.$$

From here it is easy to find the odds ratio. The odds ratio is the ratio of odds that $Y_i = 1$ given x_i+1 to the odds that $Y_i = 1$ given x_i, as a result,

$$\text{Odds Ratio} = \frac{e^{\beta_0+\beta_1 x_i} e^{\beta_1}}{e^{\beta_0+\beta_1 x_i}} = e^{\beta_1}.$$

Taking the natural log of the odds ratio yields,

$$\ln(\text{Odds Ratio}) = \beta_1.$$

This is an important finding, it means β_1 measures the change in the log-odds that $Y_i = 1$ for a one unit increase in x_i.

With all of the above in mind, the case of multiple variables will now be much easier to derive. In this case, the relationship between the discrete response variable and a set of p independent variables will be explored. A sample of size n will be taken. Let $\mathbf{x}_i$ be a vector of values for the independent variables augmented by the constant one, so that

$$\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ip}).$$

Similarly, there will be p+1 parameters so let $\boldsymbol{\beta}$ be the parameter vector where

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p).$$

Furthermore, $\pi_i = \pi(\mathbf{x}_i)$ using a multiple logistic regression model is very similar to the single variable case as shown below,

$$\pi_i = \hat{P}(Y_i = 1) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}} = \frac{e^{x_i'\boldsymbol{\beta}}}{1+e^{x_i'\boldsymbol{\beta}}}.$$

Thus, the equation for y_i is still

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

Similarly, the log-likelihood function is the same for both cases

$$L(f(y_1, \ldots, y_n)) = \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n}(1-y_i)\ln(1-\pi_i),$$

so is the maximum likelihood equation for β_0

$$\frac{d(\ln(f(y_1,\ldots,y_n))}{d\beta_o} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0.$$

For the rest of the β's, their maximum likelihood equations follow from the equation got β_1 in the single variable case with a slight adjustment. Let j = 1, …, p, then

$$\frac{d(\ln(f(y_1,\ldots,y_n))}{d\beta_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij}\pi_i = 0.$$

Like the single variable case, it becomes very hard to difficult to evaluate the likelihood equations at this point and statistical software that uses the Newton-Raphson Theorem must be used to solve for an estimate for $\boldsymbol{\beta}$. Using the maximum likelihood functions for each $\boldsymbol{\beta}_l$, q and H are obtained;

$$q' = \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i, \sum_{i=1}^{n} x_{i1} y_i - \sum_{i=1}^{n} x_{i1}\pi_i, \ldots, \sum_{i=1}^{n} x_{ip} y_i - \sum_{i=1}^{n} x_{ip}\pi_i\right)$$

and

$$H = \begin{vmatrix} \frac{d^2L(\boldsymbol{\beta})}{d\beta_0^2} & \frac{d^2L(\boldsymbol{\beta})}{d\beta_0 d\beta_1} & \cdots & \frac{d^2L(\boldsymbol{\beta})}{d\beta_0 d\beta_p} \\ \frac{d^2L(\boldsymbol{\beta})}{d\beta_0 d\beta_1} & \frac{d^2L(\boldsymbol{\beta})}{d\beta_1^2} & \cdots & \frac{d^2L(\boldsymbol{\beta})}{d\beta_1 d\beta_p} \\ \vdots & \vdots & & \vdots \\ \frac{d^2L(\boldsymbol{\beta})}{d\beta_0 d\beta_p} & \frac{d^2L(\boldsymbol{\beta})}{d\beta_1 d\beta_p} & \cdots & \frac{d^2L(\boldsymbol{\beta})}{d\beta_p^2} \end{vmatrix}$$

$$= \begin{vmatrix} -\sum_{i=1}^n \pi_i(1-\pi_i) & -\sum_{i=1}^n x_{i1}\pi_i(1-\pi_i) & \cdots & -\sum_{i=1}^n x_{ip}\pi_i(1-\pi_i) \\ -\sum_{i=1}^n x_{i1}\pi_i(1-\pi_i) & -\sum_{i=1}^n x_{i1}^2\pi_i(1-\pi_i) & \cdots & -\sum_{i=1}^n x_{i1}x_{ip}\pi_i(1-\pi_i) \\ \vdots & \vdots & & \vdots \\ -\sum_{i=1}^n x_{ip}\pi_i(1-\pi_i) & -\sum_{i=1}^n x_{i1}x_{ip}\pi_i(1-\pi_i) & \cdots & -\sum_{i=1}^n x_{ip}^2\pi_i(1-\pi_i) \end{vmatrix}.$$

From here, with the expanded q and H, the steps are the same as in the single variable case. Starting with an initial guess, $\boldsymbol{\beta}^{(0)}$, an iterative procedure is performed until the estimate for $\boldsymbol{\beta}$ converges. At the m^{th} step of this iterative process, $\boldsymbol{\beta}^{(m+1)}$ is obtained using

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - (H^{(m)})^{-1} - q^{(m)}$$

where $q^{(m)}$ and $H^{(m)}$ represent q and H evaluated at $\boldsymbol{\beta}^{(m)}$. If the algorithm converges at iteration m, then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m)}$ is the maximum likelihood estimate for $\boldsymbol{\beta}$. From here, estimates for $\hat{\pi}_i = \hat{\pi}(x_i)$ can be found by putting $\hat{\boldsymbol{\beta}}$ into the equation obtaining

$$\hat{\pi}_i = \hat{\pi}(x_i) = \frac{e^{x_i'\hat{\boldsymbol{\beta}}}}{1+e^{x_i'\hat{\boldsymbol{\beta}}}}.$$

Now the odds ratio can be expanded to the multiple variable case. Recall that in the single variable case the odds ratio is the ratio of the odds that $Y_i=1$ given x_i+1 to the odds that $Y_i=1$ given x_i. In the multivariable case, this changes slightly as x_i+1 would imply that 1 is added to all the variables in the vector x_i'. The odds ratio for a variable in logistic regression represents how the odds change if a single variable is increased by 1 holding all other variables constant. Thus, for the multivariable case one must compare x_{ij} and $x_{ij}+1$ for any $j\in[0, p]$. Note that the odds that $Y_i=1$ given x_{ij} is

$$O(x_{ij}) = \frac{\pi(x_{ij})}{1-\pi(x_{ij})} = \frac{\frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}}}} = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}}$$

and the odds that $Y_i=1$ given $x_{ij}+1$ is

$$O(x_{ij}+1) = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j(x_{ij}+1)+\cdots+\beta_p x_{ip}} = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}} e^{\beta_j}.$$

Taking the above equations, the odds ratio for any given x_{ij} is

$$\text{Odds Ratio} = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}} e^{\beta_j}}{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}}} = e^{\beta_j}.$$

Consequently, the log odds for any given x_{ij} is

$$\ln(\text{Odds Ratio}) = \beta_j.$$

This result becomes increasingly important as the number of variables increases as it allows the user to clearly see the effect one variable has on the odds of the dependent variable's success.

Equally important to the odds ratio is testing the significance of the variables. In the single variable case, only the G test statistic is needed. In the multivariate case, more tests are needed. One must be able to assess the fit of the overall model, test the significance of one variable in the overall model, and test the significance of a subset of variables in the overall model. Assessing the overall fit of the model follows from the chi-square hypothesis test (G test statistic), as seen earlier in this chapter. This test is based on the following hypothesis test

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ vs. } H_a: \text{at least one } \beta_i \text{ is nonzero.}$$

Recall, the G statistic is equal to

$$G = -2\ln\left(\frac{(likelihood\ without\ the\ variables)}{(likelihood\ with\ the\ variables)}\right) = -2\ln\left(\frac{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^{\sum_{i=1}^{n} y_i}\left(\frac{n-\sum_{i=1}^{n} y_i}{n}\right)^{n-\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} \widehat{\pi_i}^{y_i}(1-\widehat{\pi_i})^{1-y_i}}\right)$$

.

Since the likelihood without the variables remains the same in the multiple variable case, the G-statistic equation also remains the same. The only change from one case to the other is the degrees of freedom. The degree of freedom is equal to the number of variables, so in this case df = p. Thus, if $P[\chi^2(p)>G]$ is less than a specified alpha, then at least one independent variable is significant.

Next, one must be able to test the significance of a single variable in the model. To answer the question if a variable is significance, the following hypothesis must be tested

$$H_0: \beta_i = 0 \text{ vs. } H_a: \beta_i \neq 0$$

for any i=1,...,p. This test is done with a likelihood ratio chi-square test. To perform this, first one must calculate the G statistic for the full model. This statistic will be called G_{Full}. Secondly, the model is refitted without the X_i variable and the G statistic is calculated for the refitted model (called $G_{refitted}$). $G_{Refitted}$ will have p-1 degrees of freedom. After both G statistics are calculated, one must calculate $\Delta G = G_{Full} - G_{Refitted}$. This new statistic's degrees of freedom is equal to the difference in degrees of freedom between the two models, thus $\Delta df = p-(p-1) = 1$. Lastly, there is the test of the significance of a subset of variables in the model. This test is an expansion of the previous model with $G_{Refitted}$ is the G statistic of the fitted model without the subset of k variables thus, this model has $\Delta df = p-k$.

In summary, the logistic regression equation has been derived along with how to test for the significance of variables and the odds ratio. These points are critical in being able to fully interpret and understand the model created with or without software. In this project, Minitab will be used to calculate the regression equation and odds ratio. This information will be used later in the project to interpret the equation found to estimate the probability in making an insurance claim.

Chapter 2: Generating Data

Using logistic regression, the project will predict whether or not a person will make a claim based on a variety of factors. These factors are age, gender, and marital status. This information is vital in the insurance industry as insurance companies are always encouraging actuaries to find ways to predict the chance of a driver making a claim. Those with higher probabilities of making a claim will either be charged a higher premium or in extreme cases be turned away. Unfortunately, car insurance data is protected therefore the data has to be created. Existing probabilities found in Tomberlin's article "Predicting Accident Frequencies for Drivers Classified by Two Factors" will be used to aid in the creation of data. This chapter explains the details of how the data set was created using existing probabilities of claims and random generation.

Using the information for Territory 1 in Table 2 and 3 from Tomberlin's article, a table was created that became the bases of the information used for data generation.

Table 1: Information from Tomberlin's Article

	Total Population	Probability of a Claim	Number of People that Made a Claim
Married Middle-Aged Males	9852	0.0705	695
Single Middle-Aged Males	4682	0.0933	437
Married Middle-Aged Females	9830	0.0396	389
Single Middle-Aged Females	4110	0.0606	249
Old	2243	0.0531	119
Young Married Males	588	0.0850	50
Young Single Males	2981	0.1362	406
Young Females	3103	0.0719	223

The data from Tomberlin's article does not give enough information (on marital status and gender) for those in the old category so those over the age of 65 were ignored. For the young females, marital status was not given so it needed to be estimated. Assuming the ratio of married to single young males is the same as the married to single young females, an estimate for the number of married and single young females can be found.

Number of Young Married Females = 3103(588/3479) = 511

Number of Young Single Females = 3103 – 511 = 2592

Then the individual accident rates must be calculated. Knowing the overall probability is 7.19%, then the below equation must hold true

$$p_{married}\left(\frac{511}{3103}\right) + p_{single}\left(\frac{2592}{3103}\right) = 7.19\%$$

where $p_{married}$ and p_{single} represent the probability of an accident of a young female with the given marital status. To solve this equation, another equation that indicates the relationship between the two accident rates was needed. Based on the rates of the two middle aged female categories, there is about a 2% difference between the married and single female accident rates. Looking at the male categories, there seems to be a greater difference between the accident rates in the younger categories than the middle aged. Thus, an assumed difference between $p_{married}$ and p_{single} of 2.5% was made. Therefore,

$$p_{single} = p_{married} + 2.5\%.$$

Finally, using the last two equations to solve for the two accident rates yields

$$p_{married} = 5.10\% \quad p_{single} = 7.60\%.$$

Table 2 shows the updated information.

Table 2: Updated Information from Tomberlin's Article

	Total Population	Probability of a Claim	Number of People that Made a Claim
Married Middle-Aged Males	9852	0.0705	695
Single Middle-Aged Males	4682	0.0933	437
Married Middle-Aged Females	9830	0.0396	389
Single Middle-Aged Females	4110	0.0606	249
Young Married Males	588	0.0850	50
Young Single Males	2981	0.1362	406
Young Married Females	511	0.0510	26
Young Single Females	2592	0.0760	197

From the data in Table 2, it is clear that the probability of an accident goes down with age and if a driver is married their chance of making a claim is lower. Also, males in any category compared to their female counterpart are more likely to make a claim. With this in mind, an assumption of a decreasing exponential distribution of the ages for those that had an accident was made. Using this assumption, the first attempt at creating data was done. Quickly, it became clear that the current way of breaking down the categories would not work. Separating by age category was creating what looked to be a spike in accidents at age 26 which given the data was not true. Thus, the model had to be adjusted.

After identifying, having two different age categories for each marital status/gender grouping was causing problems, a new table was created where age wasn't a factor, only marital status and gender. The new totals added the young and middle-aged drivers together. The probability of an accident was calculated by dividing the cumulative number of people that had a claim by the cumulative population of people. See Table 3.

Table 3: Population Information

	Total Population	Probability of a Claim	Number of People that Made a Claim
Married Females	10341	0.04013	415
Single Females	6702	0.06655	446
Married Males	10440	0.07136	745
Single Males	7663	0.1101	843

Given the data in Table 3, ages were randomly created for each category. The ages were generated using a uniform distribution in Minitab that spanned from 16 to 64 years old. A uniform distribution was chosen to generate ages based on the Canadian Census. It showed that the ages of Canada's population were relatively uniform for all ages and gender. In order to better illustrate the difference between the probability of an accident for a driver that is 16 years old and one that is 64 years old, a logistic distribution was used to find the probability of a person making a claim. For each category, a single variable logistic equation was found based on a chosen lower and upper bound for p, the probability of an accident. Based on the assumption that the younger a person is the higher their chance of a claim, the upper bound for p should be reached when x=16 and the lower bound should be reached when x = 64. With this intention, to find the coefficients of each equation the following equations must be solved;

$$Upper\ Bound = UB = \frac{e^{\beta_0+16\beta_1}}{1+e^{\beta_0+16\beta_1}} \qquad Lower\ Bound = LB = \frac{e^{\beta_0+64\beta_1}}{1+e^{\beta_0+64\beta_1}}.$$

Rearranging these equations yields

$$\beta_0 + 16\beta_1 = \ln(\frac{UB}{1-UB}) \qquad \beta_0 + 64\beta_1 = \ln(\frac{LB}{1-LB}).$$

Thus, one can solve for β_1 by substituting $\beta_0 = \ln\left(\frac{UB}{1-UB}\right) - 16\beta_1$ into $\beta_0 + 64\beta_1 = \ln(\frac{LB}{1-LB})$ to get

$$48\beta_1 = \ln\left(\frac{LB * (1-UB)}{(1-LB) * UB}\right).$$

In order to get the best coefficients for the logistic equation a process of trial and error had to occur. Upper and lower bounds for p were estimated then, β_0 and β_1 were found using these estimations. These coefficients were put into Minitab to generate probabilities based on the ages that were already generated. From there, a macro text file was called to randomly assign 0's or 1's based on a Bernoulli distribution. This macro code can be found in Appendix 1. This generated whether the person had made a claim. The average of the claim column was then calculated to see if it equaled the probability of a claim for that group found in Table 3. If they didn't match, another estimate for p was made and the process was repeated until they were roughly equal; within 0.002 of the actual.

To illustrate the steps taken, the process for finding the proper coefficients for the Single Males will be shown. Before the trial and error process could begin, the ages of the single males were randomly generated based on a uniform distribution. Recall, for this sample, the probability of an accident should be roughly 11%. The initial estimate for the range of probabilities of an accident was (0.02, 0.25). Thus,

$$\text{UB} = 0.25 \quad \text{and} \quad \text{LB} = 0.02.$$

Using the above bounds and $48\beta_1 = \ln\left(\frac{LB*(1-UB)}{(1-LB)*UB}\right)$, $\widehat{\beta_1}$ was calculated to be

$$\widehat{\beta_1} = -0.058191834.$$

From here, an estimate for β_0 was calculated by putting $\widehat{\beta_1}$ and UB into $\beta_0 + 16\beta_1 = \ln(\frac{UB}{1-UB})$ and rearranging. This yielded

$$\widehat{\beta_0} = -0.167543.$$

Using these coefficients, the probability of a claim was calculated for each generated single male based on their age. The equation used to find these probabilities is an equation explored in chapter 1,

$$\pi_i = \pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}},$$

with x_i representing the age of the single male. From here, the macro text file was called upon to randomly generate 0's and 1's given the calculated probabilities. To decide if an individual had made a claim, it was represented by a 1, or an individual has not had an accident, it was represented by a 0. To determine if the coefficients were a good fit, the

average of the accident column was taken. The column had a probability of an accident of about 9% which is low, so the process was repeated. This time the bounds were chosen to be (0.02, 0.30). Repeating the calculations above, estimates for β_0 and β_1 were found to be

$$\widehat{\beta_0} = 0.167543$$
$$\widehat{\beta_1} = -0.063428.$$

Again, the probability of an accident for each single male was calculated and the macro text file was called to assign accidents based on these probabilities and a Bernoulli distribution. However, this time when the probability of an accident was taken, it was found to be 11.105311% and these coefficients were kept.

The trial and error process was repeated for all four categories until coefficients that fit the data were found for each and the accidents for each group were generated. The upper and lower bounds and the probability of an accident generated is shown in Table 4.

Table 4: Generated Probability of an Accident

	Lower Bound	**Upper Bound**	**Probability of an Accident**
Married Females	0.01	0.10	0.04023
Single Females	0.01	0.20	0.06491
Married Males	0.015	0.20	0.07117
Single Males	0.02	0.30	0.11105

These generated probabilities are very close to the true probabilities of the population; therefore, this generated data was kept. Now that the data has been created, it is ready to be sampled.

Chapter 3: Sampling Data

Sampling data allows a statistician to make inferences about the entire population while minimizing costs of the project. Project costs can cover anything from money spent to time and manpower used. Although in this project, costs could be considered little to nothing, sampling was done to illustrate the power of proper sampling. Since the data for this project was already broken up into independent groups, stratified random sampling was the obvious choice of sampling method. If a population can be separated into nonoverlapping groups, called strata, then stratified sampling is a great choice of sampling method as in many instances it increases the quantity of information for a given cost when compared to more simple methods, like simple random sampling. In this chapter, the steps of stratified sampling will be explored using the population data generated in the previous chapter.

To begin, the size of the sample must be calculated. This is chosen based on what bound, B, is chosen for the error of estimation. A smaller bound can lead to a smaller confidence interval around p, where p is the proportion of claims over the entire population. Thus, a small bound is ideal but it is important to keep in mind that this leads to larger sample size and increase costs. Therefore, if cost is an issue a larger bound may be needed. Also, it is important to note that the bound is equal to

$$B = 2\sqrt{V(\hat{p}_{st})}.$$

For the purposes of this project, the bound was chosen to be 0.01. Letting L = 4, where L represents the number of strata, the equation for calculating the sample size required to estimate p with a bound of B is

$$n = \frac{\sum_{i=1}^{L} N_i^2 p_i q_i / a_i}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i}$$

where a_i is the fraction of observations allocated to stratum i, p_i is the population proportion of claims made for stratum i, N_i is the number of people in strata i, and $D=B^2/4=(0.01)^2/4=0.000025$. In cases where the true population mean is not known, statistician will do preliminary sampling to come up with an estimate for p_i. For this project, the true proportion of the population that made claims for each stratum is known

so those averages were used. The variable a_i is the proportion of the sample needed for each stratum. The equation for this variable is

$$a_i = \frac{N_i\sqrt{p_i q_i / c_i}}{\sum_{i=1}^{L} N_i\sqrt{p_i q_i / c_i}}$$

where c_i is the cost of obtaining a single observation for strata i. Assuming $c_i = c$ for all I, a_i is

$$a_i = \frac{N_i\sqrt{p_i q_i}}{\sum_{i=1}^{L} N_i\sqrt{p_i q_i}}$$

Using the following information taken from Table 3 in Chapter 2,

Table 5: Shortened Table 3

	Total Number of People in the Strata (N_i)	**Probability of a Making a Claim (p_i)**
Married Females (i=1)	10341	0.04013
Single Females (i=2)	6702	0.06655
Married Males (i=3)	10440	0.07136
Single Males (i=4)	7663	0.1101

each a_i was calculated which is shown in Table 6.

Table 6: Calculations for a_i

	$N_i\sqrt{p_iq_i}$	$\sum_{i=1}^{L} N_i\sqrt{p_iq_i}$	a_i
Married Females (i=1)	2029.56669	8787.293628	0.230966072
Single Females (i=2)	1671.57846	8787.293628	0.190226767
Married Males (i=3)	2687.51943	8787.293628	0.305841541
Single Males (i=4)	2398.62905	8787.293628	0.272965619

From here, n was ready to be calculated. Table 7 shows the extra calculations done in Excel needed to calculate n.

Table 7: Calculations for $\mathbf{N_i^2 p_i q_i / a_i}$

	$N_i^2 p_i q_i / a_i$
Married Females (i=1)	17834398.5
Single Females (i=2)	14688650.8
Married Males (i=3)	23616022.4
Single Males (i=4)	21077457.7
$\sum_{i=1}^{L} N_i^2 p_i q_i / a_i$	**77216529.3**

Thus,

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 p_i q_i}{a_i}}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i}$$

is equal to

$$n = \frac{77216529.3}{35146^2(0.000025) + 8787.293628} = 2330.0858$$

after substituting in the proper summations and then simplifying. This sample would be rounded to the nearest whole number; 2330.

After calculating the overall sample size, there is only one more step before sampling. Since it was determined that stratified sampling was the best method for sampling the population, the sample size of each stratum must be determined as well. This is determined by

$$n_i = a_i * n.$$

These numbers will need to be rounded. The calculated n_i's are shown in the Table 8 below.

Table 8: Calculation for n_i

	a_i	n_i
Married Females (i=1)	0.230966072	538
Single Females (i=2)	0.190226767	443
Married Males (i=3)	0.305841541	713
Single Males (i=4)	0.272965619	636

Now that the size of each strata in the sample has been calculated, the data was now ready to be sampled.

In Minitab, there is function that easily allows the user to sample their data. For each stratum, this function was used to take a random sample of size n_i. Show on the next page in Table 9 are the first 10 observations of the sampled data.

Table 9: Sampled Data - First 10 Observations

MF S Age	MF S Y	SF S Age	SF S Y	MM S Age	MM S Y	SM S Age	SM S Y
41	0	46	0	59	0	51	0
48	0	28	0	25	1	32	1
42	0	29	0	32	0	56	0
58	0	20	1	35	0	50	0
61	0	24	0	50	0	54	0
27	0	59	0	19	0	36	0
56	0	22	0	25	0	26	0
18	0	59	0	42	0	36	0
42	0	42	0	30	0	52	0

where MF = Married Female, SF = Single Female, MM = Married Male, SM = Single Male, and the solo S in each category is to indicate that the data is the sampled data.

The sample was then analyzed to find its variance. The equation for the sample variance is

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1}\right)$$

where $\hat{p}_{st}$ is the overall probability of a claim for the sample and $\hat{p}_i$ is the probability of a claim for each stratum. For each strata the probability of a claim based on the sample is shown in Table 10 on the next page.

Table 10: Sample Data - Probabilities of Making a Claim

	Married Females	Single Females	Married Males	Single Males	Population
Probability of Making a Claim	0.040892193	0.06546275	0.064516129	0.11320755	0.06836213
95% Confidence Interval for the Probability	(0.0371, 0.0447)	(0.0596, 0.0713)	(0.0599, 0.0692)	(0.1063, 0.1201)	(0.0584, 0.0783

Using these probabilities, the variance was found to be

$$\hat{V}(\hat{p}_{st}) = 0.0000245548.$$

Recall, that the bound was chosen to be 0.01 and that the bound also equals $2\sqrt{V(\hat{p}_{st})}$. This relation can be used as a way to check if the calculations done for the survey sample size were correct. In this case,

$$B = 2\sqrt{V(\hat{p}_{st})} = 2\sqrt{0.0000245548} = 0.009910551 \approx 0.01$$

which confirms that the calculations done for finding the sample size was correct.

Now that bound has been calculated, a 95% confidence interval for the true probability of a claim for the overall data was found. When sampling, a 95% confidence interval is calculated by

$$\hat{p}_{st} \pm 2\sqrt{\hat{V}(\hat{p}_{st})}.$$

Therefore, for the current sample the 95% confidence interval is

$$0.06836213 \pm 0.009910551.$$

This confidence interval is also shown above in Table 10. Since the data in this project was generated using the true population mean of each category the validity of the confidence interval can be tested.

The overall population has a probability of a claim of 6.9701%. Thus, for the sample data, the true population mean is contained within its confidence interval. To test the validity of survey sampling and the equation for a 95% confidence interval, the process of sampling the data was repeated 500 times. Each time a sample was taken, the bound and the upper and lower limit of the confidence interval was recorded. Along with this, an indicator column was used to track whether the true population mean was contained in this interval. This process was automated using a macro text file in Minitab. A sample of the test code used can be found in Appendix 2. After running the test code, it was found that the interval covered the true population proportion in 478 of the 500; a proportion of 95.6% of the time.

As demonstrated, the stratified sampling method was an appropriate choice for sampling the insurance claim data. This method proved to be a logical choice because the insurance claim data was already available in categories or strata. The validity of the formulas used, and the calculations were periodically checked to confirm that everything was performing the way it should. Now that the data has been sampled and double-checked the sample can be used to find a logistic regression equation.

Chapter 4: Performing and Interpreting Logistic Regression

The goal of the project was to find an equation using logistic regression that will estimate the probability of person making a claim on their insurance based on their gender, age, and marital status. In review, the population data has been generated and sampled following general survey sampling guidelines. The generated sample data will be used to create the equation to estimate the probability of a person making a claim. Using the theory shown in Chapter 1 and the created equation, an analysis of a variety of variables will be done.

Before performing the regression analysis, one last transformation needed to be made on the data. Recall that the data was broken up by age and marital status/gender, however for the regression equation it should be categorized by gender, age, and marital status. Therefore, all the categories must be joined with indicators for gender and marital status. For this project, females were indicated by a 1 and males a 0. Also, those in the married category were marked with a 1 and those that are single were 0s. Using Minitab, the data was stacked by combining all the data from each stratum into two large variables of Y and Age with additional variables used as indicators for gender and marital status. Three more variables were included representing the relationships between variables. They were created by multiplying the two variables together. These variables show the effect that certain combinations of variables have on the probability of a claim. For example, if after regression, the gender/married variable has a negative coefficient then that would suggest that being a married female decreases the chance of making a claim even more than being just female or married. The first 10 people, who are all married females, are shown in Table 11.

Table 11: Sample Data

Y Sample	Age Sample	Gender Sample	Married Sample	Age/Gen Sample	Age/Mar Sample	Gen/Mar Sample
0	41	1	1	41	41	1
0	48	1	1	48	48	1
0	42	1	1	42	42	1
0	58	1	1	58	58	1
0	61	1	1	61	61	1
0	27	1	1	27	27	1
0	56	1	1	56	56	1
0	18	1	1	18	18	1
0	42	1	1	42	42	1

Once the data was in the preferred form of age, gender, and married, logistic regression analysis could be performed. To begin, the regression analysis was run with all variables. Step-by-step, the least significant variable was removed until all variables are significant. Variables are deemed significant if they have a p-value less than 0.05. With Minitab, the data with all variables was put through logistic regression. The entire output can be found in Appendix 3, but the Deviance Table is shown in Table 12.

Table 12: Deviance Table – All Variables

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	6	160.91	26.8188	160.91	0.000
Age Sample	1	81.03	81.0318	81.03	0.000
Gen Sample	1	2.61	2.6134	2.61	0.106
Mar Sample	1	10.14	10.1378	10.14	0.001
AG Sample	1	0.25	0.2506	0.25	0.617
AM Sample	1	3.85	3.8533	3.85	0.050
GM Sample	1	0.10	0.1033	0.10	0.748
Error	2323	1051.34	0.4526		
Total	2329	1212.25			

As shown in Table 12, the p-value of GM (Gender/Married) is the largest and thus, the least significant so it was removed.

After removing the GM variable from the sample, the logistic regression was run again. This time the output from Minitab showed that AG (Age/Gender) was insignificant. The Deviance Table is shown in Table 13. The full Minitab output can be found in Appendix 4.

Table 13: Deviance Table – GM Removed

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	5	160.81	32.1619	160.81	0.000
Age Sample	1	81.39	81.3943	81.39	0.000
Gen Sample	1	2.51	2.5121	2.51	0.113
Mar Sample	1	10.16	10.1610	10.16	0.001
AG Sample	1	0.28	0.2805	0.28	0.596
AM Sample	1	3.89	3.8940	3.89	0.048
Error	2324	1051.44	0.4524		
Total	2329	1212.25			

The sample data was put through logistic regression on more time where all variables were shown to be significant. The Deviance Table for the model is shown in Table 14. The rest of the Minitab output can be found in Appendix 5

Table 14: Deviance Table – Chosen Model

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	4	160.53	40.1323	160.53	0.000
Age Sample	1	97.88	97.8828	97.88	0.000
Gen Sample	1	10.11	10.1146	10.11	0.001
Mar Sample	1	10.10	10.1034	10.10	0.001
AM Sample	1	3.86	3.8615	3.86	0.049
Error	2325	1051.72	0.4524		
Total	2329	1212.25			

Since all variables left in the model are significant, this is the accepted model of logistic regression and further analysis was done. Recall from Chapter 1 that the equation for π_i

$$\pi_i = P(Y_i = 1) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}.$$

Letting $\widehat{W}=\widehat{\beta_0}+\widehat{\beta_1}x_{i1}+\cdots+\widehat{\beta_p}x_{ip}$, the output from Minitab shows that for the sample data, $\widehat{W}$ is

$\widehat{W}$ = 0.987 - 0.0904*Age - 0.557*Gen - 1.515*Mar + 0.0297*AgeMar.

Thus,

$$\hat{P}(Y_i=1)=\frac{e^{\widehat{W}}}{1+e^{\widehat{W}}}=\frac{e^{0.987-0.0904*\text{Age}-0.557*\text{Gen}-1.515*\text{Mar}+0.0297*\text{AgeMar}}}{1+e^{0.987-0.0904*\text{Age}-0.557*\text{Gen}-1.515*\text{Mar}+0.0297*\text{AgeMar}}}.$$

The equation for W alone offers important information about the relationship between making a claim and the variables. In fact, the negative coefficient for the age variable indicates that as a person gets older their likelihood of getting in an accident goes down. Also, the negative coefficients for the gender and marital status variables imply that when their values are 1, so when a person is female and/or married respectively, the probability of them making a claim goes down. Surprisingly, the positive coefficient for the age/marital status relationship shows that those that are married have a slightly higher chance of making a claim as they age.

Aside from the coefficient information, the equation for $P(Y_i=1)$ gives the estimated probability of a person making a claim based on all variables. For example, a single 20-year-old male has a probability of making a claim of

$$\hat{P}(Y_i=1)=\frac{e^{0.987-0.0904*\text{Age}-0.557*\text{Gen}-1.515*\text{Mar}+0.0297*\text{AgeMar}}}{1+e^{0.987-0.0904*\text{Age}-0.557*\text{Gen}-1.515*\text{Mar}+0.0297*\text{AgeMar}}}$$

$$=\frac{e^{0.987-0.0904*(20)-0.557*(0)-1.515*(0)+0.0297*(20)(0)}}{1+e^{0.987-0.0904*(20)-0.557*(0)-1.515*(0)+0.0297*(20)(0)}}$$

$$=0.3055.$$

While the probability of a single 20-year-old female is

$$\hat{P}(Y_i=1)=\frac{e^{0.987-0.0904*\text{Age}-0.557*\text{Gen}-1.515*\text{Mar}+0.0297*\text{AgeMar}}}{1+e^{0.987-0.0904*\text{Age}-0.557*\text{Gen}-1.515*\text{Mar}+0.0297*\text{AgeMar}}}$$

$$=\frac{e^{0.987-0.0904*(20)-0.557*(1)-1.515*(0)+0.0297*(20)(0)}}{1+e^{0.987-0.0904*(20)-0.557*(1)-1.515*(0)+0.0297*(20)(0)}}$$

$$=0.2013.$$

Thus, a single 20-year-old male is 52% more likely to make a claim than a single female of the same age. Important findings such as this can be done for any marital status, age, and gender combination.

Taking advantage of the fact that the entire population of the data is known, the analysis of the regression can be taken a step further. Repeating the process of eliminating insignificant variables in the model for the population data, it can be shown that the same variables are found to be significant as those in the sample model. These steps will not be shown here but the Minitab printouts can be found in Appendix 6. The value of Y for the population model is

$\widehat{W}$ = 0.3760 - 0.07024 Age Pop - 0.6078 Gen Pop - 0.843 Mar Pop + 0.01078 AgeMar Pop.

Although the actual values of the coefficients are different, the signs are the same. Given that the logistic regression equation is non-linear, these equations are very close and thus, allowing the conclusion to be drawn that the sample model is a good fit for the data.

Moving on from the population model and back to the sample model, separate equations can be found for each category of data. Doing this can save time when calculating $\hat{P}(Y_i = 1)$ as these equations have only one variable, age. For example, the married females' category would have Gen = 1 and Mar = 1. Putting those numbers into the model yields

$\widehat{W}$ = 0.987 - 0.0904 Age - 0.557(1) - 1.515(1) + 0.0297(1) * Age.

Thus, putting $\widehat{W}$ into the logistic regression and simplifying, the logistic equation for married females that is only a factor of age.

$$\hat{P}(Y_i = 1) = \frac{e^{-1.085-0.0607*Age}}{1+e^{-1.085-0.0607*Age}}.$$

Repeating this for all categories returns the Table of Category Logistic Regression Equations shown in Table 15.

Table 15: Table of Category Logistic Regression Equations

	Value of Gender Variable (Gen)	Value of Marital Status Variable (Mar)	$P(Y_i = 1)$	$\widehat{P}(Y_i = 1)$
Married Females	1	1	$\frac{e^{(\beta_0+\beta_2+\beta_3)+(\beta_1+\beta_4)*Age}}{1+e^{(\beta_0+\beta_2+\beta_3)+(\beta_1+\beta_4)*Age}}$	$\frac{e^{-1.085-0.0607*Age}}{1+e^{-1.085-0.0607*Age}}$
Single Females	1	0	$\frac{e^{(\beta_0+\beta_2)+(\beta_1)*Age}}{1+e^{(\beta_0+\beta_2)+(\beta_1)*Age}}$	$\frac{e^{0.43-0.0904*Age}}{1+e^{0.43-0.0904*Age}}$
Married Males	0	1	$\frac{e^{(\beta_0+\beta_3)+(\beta_1+\beta_4)*Age}}{1+e^{(\beta_0+\beta_3)+(\beta_1+\beta_4)*Age}}$	$\frac{e^{-0.528-0.0607*Age}}{1+e^{-0.528-0.0607*Age}}$
Single Males	0	0	$\frac{e^{\beta_0+(\beta_1)*Age}}{1+e^{\beta_0+(\beta_1)*Age}}$	$\frac{e^{0.987-0.0904*Age}}{1+e^{0.987-0.0904*Age}}$

Using these equations, the odds and odds ratios for each category were calculated. Recall that in Chapter 1, for a single variable logistic equation, it was shown that the odds of Y_i=1 is

$$O(x_i) = \frac{\pi(x_i)}{1-\pi(x_i)} = \frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_i}}} = e^{\beta_0+\beta_1 x_i}$$

Also seen in chapter 1,

$$\text{Odds Ratio} = \frac{e^{\beta_0+\beta_1 x_i} e^{\beta_1}}{e^{\beta_0+\beta_1 x_i}} = e^{\beta_1}.$$

For this project, each category may have more than one coefficient in front of the variable, so the odds ratio may contain more than one coefficient. For example, the odds ratio for the married females of Age=X is equal to

$$\text{Odds Ratio} = \frac{e^{(\beta_0+\beta_2+\beta_3)+(\beta_1+\beta_4)*(x+1)}}{e^{(\beta_0+\beta_2+\beta_3)+(\beta_1+\beta_4)*x}} = e^{\beta_1+\beta_4}.$$

Shown in Table 16 is the odds of a person of age X making a claim and odds ratios for a one-year increase in age from X to X+1.

Table 16: Odd and Odds Ratios

	Odds of Y_i=1 Equation	**Odds Ratio Equation**	**Odds Ratio Value**
Married Females	$e^{(\beta_0+\beta_2+\beta_3)+(\beta_1+\beta_4)*x}$	$e^{\beta_1+\beta_4}$	0.941
Single Females	$e^{(\beta_0+\beta_2)+(\beta_1)*x}$	e^{β_1}	0.914
Married Males	$e^{(\beta_0+\beta_3)+(\beta_1+\beta_4)*x}$	$e^{\beta_1+\beta_4}$	0.941
Single Males	$e^{\beta_0+(\beta_1)*x}$	e^{β_1}	0.914

The odds ratio value provides important information on how aging effects a person's probability of making a claim. Every time a person, male or female, who is married ages their odds of making claim go down by a factor of 0.941. A person who is single decreases their odds of making a claim by a factor 0.914 for each year they age. In the insurance industry the above odds ratios can be very important in determining rates and can lead to reduced rates for older clients.

The odds ratios can be expanded to greater than a one year. In the case of an age increase of N years, it is more accurate to have a 95% confidence interval than one number. Since the odds ratios for males and females are the same when they share the same marital status, only two calculations needed to be done. When calculating the confidence interval of the odds ratio, it is easier to calculate the confidence interval of the ln[odds ratio] first. Using the general odds ratio equation,

$$\text{Odds Ratio} = \frac{O(x_{ij}+1)}{O(x_{ij})} = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}} e^{\beta_j}}{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}}} = e^{\beta_j},$$

the 95% confidence interval for ln[odds ratio] is

$$\hat{\beta}_j * N \pm 1.96 * \hat{\sigma}(\hat{\beta}_j * N).$$

Therefore, the confidence interval for those that are single is

$$\hat{\beta}_1 * N \pm 1.96 * \hat{\sigma}(\hat{\beta}_1 * N)$$

which is equal to

$$-0.0904 * N \pm 1.96 * N^2 * V(\hat{\beta}_1)$$

$$-0.0904 * N \pm 1.96 * N^2 * 0.0001126$$

$$-0.0904 * N \pm 0.000220696 * N^2.$$

The variance for β_l was found in the variance covariance matrix shown in Table 17. It was calculated using Minitab.

Table 17: Variance Covariance Matrix

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\hat{\beta}_0$	0.110742	-0.0032766	-0.0113545	-0.107213	0.0032795
$\hat{\beta}_1$	-0.003277	0.0001126	0.0000416	0.003264	-0.0001126
$\hat{\beta}_2$	-0.011354	0.0000416	0.0321572	0.001362	-0.0000501
$\hat{\beta}_3$	-0.107213	0.0032636	0.0013616	0.229066	-0.0067433
$\hat{\beta}_4$	0.003280	-0.0001126	-0.0000501	-0.006743	0.0002267

Calculating the ln[odds ratio] confidence interval for the married males and females was slightly more challenging as the odds ratio equation includes more than one equation. In this case, a 95% confidence interval for those that are married is

$$\left(\hat{\beta}_1 + \widehat{\beta_4}\right) * N \pm 1.96 * \hat{\sigma}(\left(\hat{\beta}_1 + \widehat{\beta_4}\right) * N)$$

which is equivalent to

$$-0.0607 * N \pm 1.96 * N^2 * (V(\hat{\beta}_1) + V(\hat{\beta}_4) + 2Cov(\hat{\beta}_1, \hat{\beta}_4))$$

$$-0.0607 * N \pm 1.96 * N^2 * (0.0001126 + 0.0002267 + 2 * (-0.0001126))$$

$$-0.0607 * N \pm 0.00022364 * N^2.$$

The variances and covariances were taken from the matrix in Table 17.

Now that 95% confidence intervals were created for the ln[odds ratio], confidence intervals for the odds ratio can be found. Using the upper and lower bounds of the ln[odds ratio], a 95% confidence interval was found using the following equation.

$$(e^{Lower\ Bound}, e^{Upper\ Bound})$$

This allows anyone evaluating insurance data access to an interval for the odds ratio over a given period of years. See Table 18. Also, see Table 19 for examples of the confidence intervals for given values of N.

Table 18: 95% Confidence Interval for the Odds Ratio

	Lower Bound	Upper Bound	95% Confidence Interval for the Odds Ratio
Married	$-0.0607 * N - 0.00022364 * N^2$	$-0.0607 * N + 0.00022364 * N^2$	$(e^{-0.0607*N-0.00022364*N^2}, e^{-0.0607*N+0.00022364*N^2})$
Single	$-0.0904 * N - 0.000220696 * N^2$	$-0.0904 * N + 0.000220696 * N^2$	$(e^{-0.0904*N-0.000220696*N^2}, e^{-0.0904*N+0.000220696*N^2})$

Table 19: Examples of the 95% Confidence Interval for the Odds Ratio

	95% Confidence Interval for the Odds Ratio	N=1	N=5	N=10
Married	$(e^{-0.0607*N-0.00022364*N^2}, e^{-0.0607*N+0.00022364*N^2})$	(0.9409, 0.9413)	(0.7341, 0.7424)	(0.5329, 0.5573)
Single	$(e^{-0.0904*N-0.000220696*N^2}, e^{-0.0904*N+0.000220696*N^2})$	(0.9134, 0.9138)	(0.6329, 0.6399)	(0.3961, 0.4140)

In summary, the final transformations of the sample data were completed, and a logistic regression model was found. This model was deemed a good fit because all variables were found to be significant and it was similar to the population model. The model included variables for age, gender, and marital status as well as a relationship variable between age and marital status. Analysis of these variables and their coefficients was completed to determine their effects on a person's probability of making a claim. Additional equations were found for each of the marital status/gender categories. Further analysis was completed by finding the odds ratio of each equation.

Conclusion

This project was able to demonstrate the importance of logistic regression and its use in the insurance industry. Using this important type of regression, the question of how likely a person is to make a claim was answered by way of a mathematical equation. The equation was taken a step further when equations were found for each stratum and analysis was done to determine their odds ratios.

In Chapter 1, the logistic regression equation was derived along with the Maximum Likelihood Equations for the coefficients. This chapter also looked into important topics such as Newton-Raphson Method for finding the coefficients, the chi-squared hypothesis tests, and the odds ratios. All of these topics were examined for both the single and multi-variable equation.

In Chapter 2, the insurance claim data found in Tomberlin's "*Predicting Accident Frequencies for Drivers Classified by Two Factors*" was used to find the probability of making a claim for each stratum (married females, single females, married males, and single males). These probabilities were then used to generate data based on logistic distribution while the ages were generated using a uniform distribution. A macro code was then created to randomly assign claims based on the probabilities and a Bernoulli equation.

In Chapter 3, the generated data was sampled to demonstrate how sampling can yield similar results at a lower cost to the statistician. The stratified sampling method was used since the population data was already categorized and stratified sampling increases the quantity of information for a given cost when compared to simple methods like simple random sampling. The sample size was then calculated using the stratified sampling equations, and a portion of the population was randomly taken as the sample accordingly. The mean and variance of the sample as well as a 95% confidence interval for the sample mean were found. Formulas and calculations in this chapter were checked periodically for their validity.

In the final chapter, Minitab was used to create a logistic regression equation for the probability of making a claim. After using p-values to eliminate insignificant variables, the equation found was

$$\hat{P}(Y_i = 1) = \frac{e^{\widehat{W}}}{1+e^{\widehat{W}}} = \frac{e^{0.987 - 0.0904*\text{Age} - 0.557*\text{Gen} - 1.515*\text{Mar} + 0.0297*\text{AgeMar}}}{1+e^{0.987 - 0.0904*\text{Age} - 0.557*\text{Gen} - 1.515*\text{Mar} + 0.0297*\text{AgeMar}}}.$$

Analysis of these variables and their coefficients was completed to determine their effects on a person's probability of making a claim. Further analysis was done by finding the equations for different stratum and determining their odd ratios. It was found that those that had the same marital status shared the same odds ratio regardless of gender.

The project findings confirmed some well-known beliefs while also having a few surprising results. The findings confirmed the belief that men, especially young men, are more likely to make a claim on their car insurance than female drivers of their age. It also confirmed the belief that younger drivers, regardless of gender, have a higher probability of making a claim. A surprising conclusion from the analysis was that people who are married have a slightly increased chance of making a claim as they age. Another interesting result was that the relationship combined variable for age and gender was found to be insignificant. Popular belief would have suggested that this variable would have a significant impact on the probability of making a claim.

In conclusion, this project demonstrated that the findings of the odds ratios for each stratum could be greatly beneficial in determining rates of insurance. To further increase the accuracy and depth of the knowledge, this project could be extended by adding in other variables that could possibly have an effect on a person's probability of making a claim on their car insurance. Examples of other variables that could be used are type of car, postal code, and whether they did drivers education.

References

Canada Revenue Agency, *2016 Census*, www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=PR&Code1=01&Geo2=&Code2=&Data=Count&SearchText=Canada&SearchType=Begins&SearchPR=01&B1=All&TABID=1, June 2018

Foltz, B, *Statistics 101: Logistic Regression*, https://www.youtube.com/watch?v=zAULhNrnuL4&list=PLIeGtxpvyG-JmBQ9XoFD4rs-b3hkcX7Uu, May 2018

Hosmer, D.W., and Lemeshow, S. (1989), *Applied Logistic Regression*, John Wiley & Sons Inc., Canada

Scheaffer, R.L., Mendenhall III, W., and Ott, R.L. (2006), *Elementary Survey Sampling 6th Edition*, Thomson Brooks/Cole, Belmont, USA

Tomberlin, T.J., (1988), "*Predicting Accident Frequencies for Drivers Classified by Two Factors*", Journal of the American Statistical Association, Volume 83, Number 402, 309-321

Appendix

Appendix 1 – Macro Code – Probability of Making a Claim

```
gmacro
loop
do k1=1:n
let k2=c1(k1)
random 1 c3;
bernoulli k2.
let c2(k1)=c3(1)
erase c3
enddo
endmacro
```

Appendix 2 – Macro Code – Verifying a 95% Confidence Interval when Sampling

```
gmacro
strat
do k50=1:500
sample 538 c50 c51 c80 c81.
sample 443 c52 c53 c82 c83.
sample 713 c54 c55 c84 c85.
sample 636 c56 c57 c86 c87.
let k1=10341
let k2=6702
let k3=10440
let k4=7663
let k5=k1+k2+k3+k4
let k6=(k1*mean(c81)+k2*mean(c83)+k3*mean(c85)+k4*mean(c87))/k5
let k7=mean(c81)
let k8=mean(c83)
let k9=mean(c85)
let k10=mean(c87)
let k11=(k1*k1)*((k1-538)/k1)*((k7*(1-k7))/537)
let k12=(k2*k2)*((k2-443)/k2)*((k8*(1-k8))/442)
let k13=(k3*k3)*((k3-713)/k3)*((k9*(1-k9))/712)
let k14=(k4*k4)*((k4-636)/k4)*((k10*(1-k10))/635)
let k15=k11+k12+k13+k14
let k16=k15/(k5*k5)
let k17=sqrt(k16)
let k18=2*k17
let k19=k6-k18
let k20=k6+k18
let c90(k50)=k7
let c91(k50)=k8
let c92(k50)=k9
let c93(k50)=k10
let c94(k50)=k6
let c95(k50)=k18
let c96(k50)=k19
let c97(k50)=k20
enddo
let c98=(c96 le 0.069701 and c97 ge 0.069701)
endmacro
```

Appendix 3 – Minitab Output – Sample Data - All Variables

Binary Logistic Regression: Y Sample versus Age Sample, ... M Sample

Method

Link function	Logit
Rows used	2330

Response Information

Variable	Value	Count	
Y Sample	1	169	(Event)
	0	2161	
	Total	2330	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	6	160.91	26.8188	160.91	0.000
Age Sample	1	81.03	81.0318	81.03	0.000
Gen Sample	1	2.61	2.6134	2.61	0.106
Mar Sample	1	10.14	10.1378	10.14	0.001
AG Sample	1	0.25	0.2506	0.25	0.617
AM Sample	1	3.85	3.8533	3.85	0.050
GM Sample	1	0.10	0.1033	0.10	0.748
Error	2323	1051.34	0.4526		
Total	2329	1212.25			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
13.27%	12.78%	1065.34

Coefficients

Term	Coef	SE Coef	VIF
Constant	1.081	0.368	
Age Sample	-0.0930	0.0118	2.48
Gen Sample	-0.848	0.524	8.58
Mar Sample	-1.555	0.491	8.48
AG Sample	0.0081	0.0161	8.79

AM Sample	0.0297	0.0151	9.44
GM Sample	0.116	0.362	2.32

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age Sample	0.9112	(0.8903, 0.9326)
Gen Sample	0.4285	(0.1534, 1.1970)
Mar Sample	0.2112	(0.0807, 0.5528)
AG Sample	1.0081	(0.9768, 1.0404)
AM Sample	1.0301	(1.0001, 1.0610)
GM Sample	1.1235	(0.5526, 2.2841)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y ' = 1.081 - 0.0930 Age Sample - 0.848 Gen Sample - 1.555 Mar Sample
+ 0.0081 AG Sample
+ 0.0297 AM Sample + 0.116 GM Sample

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	2323	1051.34	1.000
Pearson	2323	2254.55	0.842
Hosmer-Lemeshow	8	4.00	0.857

Appendix 4 – Minitab Output – Sample Data – Without GM

Binary Logistic Regression: Y Sample versus Age Sample, ... M Sample
Method

Link function	Logit
Rows used	2330

Response Information

Variable	Value	Count	
Y Sample	1	169	(Event)
	0	2161	
	Total	2330	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	5	160.81	32.1619	160.81	0.000
Age Sample	1	81.39	81.3943	81.39	0.000
Gen Sample	1	2.51	2.5121	2.51	0.113
Mar Sample	1	10.16	10.1610	10.16	0.001
AG Sample	1	0.28	0.2805	0.28	0.596
AM Sample	1	3.89	3.8940	3.89	0.048
Error	2324	1051.44	0.4524		
Total	2329	1212.25			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
13.27%	12.85%	1063.44

Coefficients

Term	Coef	SE Coef	VIF
Constant	1.070	0.368	
Age Sample	-0.0932	0.0119	2.50
Gen Sample	-0.810	0.511	8.15
Mar Sample	-1.521	0.479	8.09
AG Sample	0.0085	0.0160	8.73
AM Sample	0.0298	0.0151	9.43

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age Sample	0.9111	(0.8901, 0.9325)
Gen Sample	0.4448	(0.1634, 1.2114)
Mar Sample	0.2184	(0.0854, 0.5588)
AG Sample	1.0086	(0.9774, 1.0408)
AM Sample	1.0303	(1.0003, 1.0612)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y ' = 1.070 - 0.0932 Age Sample - 0.810 Gen Sample - 1.521 Mar Sample
+ 0.0085 AG Sample
+ 0.0298 AM Sample

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	2324	1051.44	1.000
Pearson	2324	2261.88	0.818
Hosmer-Lemeshow	8	4.44	0.815

Appendix 5 – Minitab Output – Sample Data – Chosen Model

Binary Logistic Regression: Y Sample versus Age Sample, ... M Sample
Method

Link function	Logit
Rows used	2330

Response Information

Variable	Value	Count	
Y Sample	1	169	(Event)
	0	2161	
	Total	2330	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	4	160.53	40.1323	160.53	0.000
Age Sample	1	97.88	97.8828	97.88	0.000
Gen Sample	1	10.11	10.1146	10.11	0.001
Mar Sample	1	10.10	10.1034	10.10	0.001
AM Sample	1	3.86	3.8615	3.86	0.049
Error	2325	1051.72	0.4524		
Total	2329	1212.25			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
13.24%	12.91%	1061.72

Coefficients

Term	Coef	SE Coef	VIF
Constant	0.987	0.333	
Age Sample	-0.0904	0.0106	2.00
Gen Sample	-0.557	0.179	1.00
Mar Sample	-1.515	0.479	8.07
AM Sample	0.0297	0.0151	9.42

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age Sample	0.9136	(0.8948, 0.9328)
Gen Sample	0.5732	(0.4033, 0.8146)
Mar Sample	0.2198	(0.0860, 0.5615)
AM Sample	1.0301	(1.0002, 1.0610)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y ' = 0.987 - 0.0904 Age Sample - 0.557 Gen Sample - 1.515 Mar Sample + 0.0297 AM Sample

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	2325	1051.72	1.000
Pearson	2325	2290.47	0.691
Hosmer-Lemeshow	8	2.81	0.946

Appendix 6(a) – Minitab Output – Population Data – All Variables

Binary Logistic Regression: Y versus Age, Gender, ... n, AgeMS, GenMS

Method

Link function	Logit
Rows used	35146

Response Information

Variable	Value	Count	
Y	1	2445	(Event)
	0	32701	
	Total	35146	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	6	1848.7	308.115	1848.69	0.000
Age	1	689.5	689.550	689.55	0.000
Gender	1	27.0	26.974	26.97	0.000
MarStat	1	44.2	44.224	44.22	0.000
AgeGen	1	0.5	0.540	0.54	0.463
AgeMS	1	8.4	8.435	8.43	0.004
GenMS	1	0.0	0.023	0.02	0.878
Error	35139	15901.3	0.453		
Total	35145	17749.9			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
10.42%	10.38%	15915.26

Coefficients

Term	Coef	SE Coef	VIF
Constant	0.4045	0.0988	
Age	-0.07124	0.00299	2.61
Gender	-0.687	0.132	8.68
MarStat	-0.837	0.126	8.50
AgeGen	0.00282	0.00384	8.39

AgeMS	0.01073	0.00370	9.07
GenMS	-0.0138	0.0900	2.58

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age	0.9312	(0.9258, 0.9367)
Gender	0.5029	(0.3879, 0.6520)
MarStat	0.4329	(0.3381, 0.5543)
AgeGen	1.0028	(0.9953, 1.0104)
AgeMS	1.0108	(1.0035, 1.0181)
GenMS	0.9863	(0.8269, 1.1765)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y ' = 0.4045 - 0.07124 Age - 0.687 Gender - 0.837 MarStat + 0.00282 AgeGen + 0.01073 AgeMS - 0.0138 GenMS

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	35139	15901.26	1.000
Pearson	35139	34784.38	0.910
Hosmer-Lemeshow	8	12.40	0.134

Appendix 6(b) – Minitab Output – Population Data – GenMS Removed

Binary Logistic Regression: Y versus Age, Gender, ... t, AgeGen, AgeMS

Method

Link function	Logit
Rows used	35146

Response Information

Variable	Value	Count	
Y	1	2445	(Event)
	0	32701	
	Total	35146	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	5	1848.7	369.734	1848.67	0.000
Age	1	701.6	701.608	701.61	0.000
Gender	1	30.4	30.417	30.42	0.000
MarStat	1	48.2	48.221	48.22	0.000
AgeGen	1	0.5	0.533	0.53	0.465
AgeMS	1	8.4	8.443	8.44	0.004
Error	35140	15901.3	0.453		
Total	35145	17749.9			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
10.42%	10.39%	15913.28

Coefficients

Term	Coef	SE Coef	VIF
Constant	0.4069	0.0975	
Age	-0.07124	0.00298	2.61
Gender	-0.694	0.126	7.84
MarStat	-0.842	0.121	7.88
AgeGen	0.00281	0.00384	8.38
AgeMS	0.01074	0.00370	9.07

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age	0.9312	(0.9258, 0.9367)
Gender	0.4998	(0.3905, 0.6396)
MarStat	0.4307	(0.3394, 0.5465)
AgeGen	1.0028	(0.9953, 1.0104)
AgeMS	1.0108	(1.0035, 1.0181)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y ' = 0.4069 - 0.07124 Age - 0.694 Gender - 0.842 MarStat + 0.00281 AgeGen + 0.01074 AgeMS

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	35140	15901.28	1.000
Pearson	35140	34773.88	0.917
Hosmer-Lemeshow	8	12.28	0.139

Appendix 6(c) – Minitab Output – Population Data – Chosen Model

Binary Logistic Regression: Y versus Age, Gender, MarStat, AgeMS
Method

Link function	Logit
Rows used	35146

Response Information

Variable	Value	Count	
Y	1	2445	(Event)
	0	32701	
	Total	35146	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	4	1848.1	462.034	1848.14	0.000
Age	1	880.9	880.874	880.87	0.000
Gender	1	189.8	189.798	189.80	0.000
MarStat	1	48.4	48.373	48.37	0.000
AgeMS	1	8.5	8.506	8.51	0.004
Error	35141	15901.8	0.453		
Total	35145	17749.9			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
10.41%	10.39%	15911.81

Coefficients

Term	Coef	SE Coef	VIF
Constant	0.3760	0.0878	
Age	-0.07024	0.00265	2.05
Gender	-0.6078	0.0450	1.00
MarStat	-0.843	0.121	7.88
AgeMS	0.01078	0.00370	9.07

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age	0.9322	(0.9273, 0.9370)
Gender	0.5446	(0.4986, 0.5947)
MarStat	0.4303	(0.3392, 0.5459)
AgeMS	1.0108	(1.0035, 1.0182)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 0.3760 - 0.07024 Age - 0.6078 Gender - 0.843 MarStat + 0.01078 AgeMS

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	35141	15901.81	1.000
Pearson	35141	34917.80	0.800
Hosmer-Lemeshow	8	11.07	0.198