CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Modelling the Probability of Making an Insurance Claim Based on Logistic Regression

AUTHOR: Danielle Chartier

SUPERVISOR: Patrick Farrell

DATE: August 15, 2020

Table of Contents

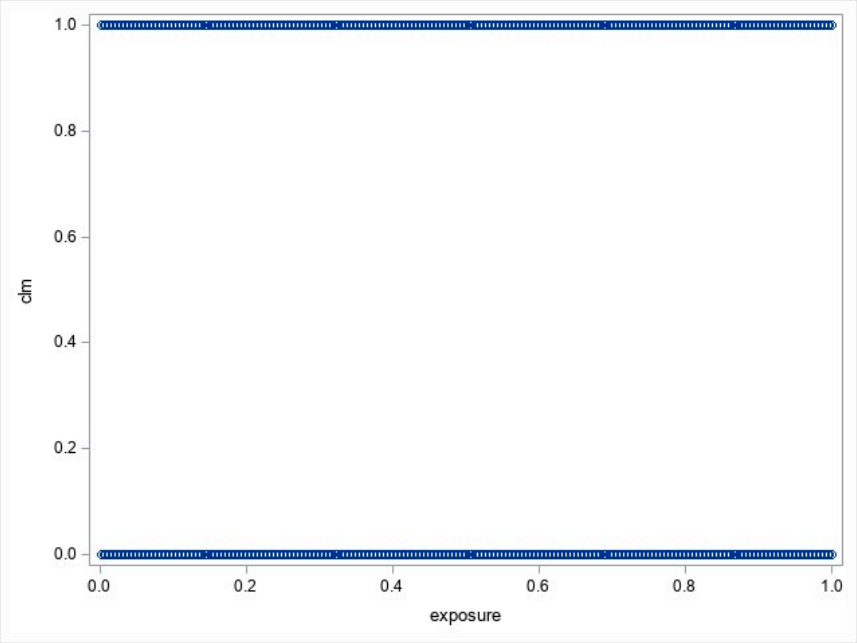
Abstract

This paper will explore the categorical logistic relationship between quantifiable independent variables against a categorical dependent Bernoulli random variable in an insurance claim dataset. In particular, this paper studies independent variables such as exposure, vehicle type, vehicle age, gender, geographic area, and driver age, as well as a binary dependent Bernoulli random variable of whether or not a given person made a car insurance claim. Using logistic regression, it will be determined which of the above variables have a significant effect in determining the probability of a given claimant making a claim. Additionally, it will explore relationships between independent variables, determined by adding an interaction term to the model.

Chapter 1- Logistic Regression

1.1 Introduction to Logistic Regression

As with any statistic model-building technique, logistic regression aims to find the most fitting model to describe or predict the relationships between a dependent variable and one or more independent variables. Logistic regression, however, is more suitable for datasets whose dependent variable is binary and independent variable(s) are categorical. Producing scatterplots of data set with binary response variables will look similar to the following:



(1.1.1)

It is clear to see that from the figure 1.1.1, a linear regression model would not be appropriate as the model does not follow a linear trend. As such, logistic regression is a common and useful statistical technique to fit appropriate models to binary response variables.

The scatterplot produced in figure 1.1.1 above represents a single, continuous, and independent variable; exposure (from car insurance data, this example is explored further in chapter 3), plotted against a binary dependent variable; insurance claims. Here, a simple linear

model cannot accurately predict the probability that a given exposure value will have a claim value of 0 or a claim value of 1, whereas simple logistic regression model will be able to predict these probabilities with more certainty.

Comparable to linear regression, there can be one or more independent variables, so logistic regression splits into two models; simple logistic regression and multiple logistic regression, which will be discussed in sections 1.2 and 1.3, respectively. Beginning with an understanding of simple logistic regression is essential to set the foundation to further build upon in multiple logistic regression.

1.2 Simple Logistic Regression

Simple logistic regression is generally an appropriate model to use when simple linear regression does not fit under the above conditions, since logistic regression shares many ideal characteristics with linear regression, in the sense that the logistic regression model is linear within its parameters and may range over all the real numbers depending on its domain. These ideal characteristics simplify the logistic model, making it easier to work with and fit in real-world applications.

Since logistic regression models fit best when the dependent variable is binary, it is important to recognize the significance of the distribution of the dependent variable. A binary dependent variable Y follows a Bernoulli distribution with parameter π with probability of success, $P(Y = 1) = \pi$, and probability of failure, $P(Y = 0) = (1 - \pi)$. Thus, the probability density function of Y is

$$f(y) = \pi^{y} (1 - \pi)^{1-y}$$

We can assume that for $i = 1, \dots, n,$ where n is the number of independent observations, then the Y_i's are independent with joint probability density function

$$f(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} f(y_i)$$

$$= \prod_{i=1}^{n} \pi_i{}^{y_i} (1 - \pi_i)^{1-y_i} \qquad (1.2.1)$$

Where

- π_i *is the parameter for* $Y_i, i = 1, \dots, n$;
- $\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$; and
- β_0, β_1 *are estimated using maximum likelihood method*

As noted above, β_0 and β_1 are estimated using the maximum likelihood method. Whereas in simple linear regression β_0 and β_1 are estimated using the least squares method. However, the least squares estimates for β_0 and β_1 do not share the ideal properties under logistic regression. As such, the maximum likelihood method provides more accurate estimations of β_0 and β_1 in that the estimates are more appropriate in applications of binary data. To use the maximum likelihood method to produce values for β_0 and β_1, one tries to maximize the likelihood function for values β_0 and β_1. First the natural logarithm of equation 1.2.1 is applied to both sides to produce:

$$\ln\big(f(y_1, y_2, \dots, y_n)\big) = \ln\left(\prod_{i=1}^{n} \pi_i{}^{y_i} (1 - \pi_i)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} \ln\left(\pi_i{}^{y_i} (1 - \pi_i)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - \pi_i) \qquad (1.2.2)$$

Next, equation 1.2.2 is then derived with respect to β_0, and then with respect to β_1, producing the following 2 equations (note the substitution of π_i in the second line as defined above), called the likelihood equations:

$$\frac{\partial(\ln(f(y_1, y_2, \dots, y_n)))}{\partial(\beta_0)} = \frac{\partial(\sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n}(1-y_i)\ln(1-\pi_i))}{\partial(\beta_0)}$$

$$= \frac{\partial[(\sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)) + (\sum_{i=1}^{n}(1-y_i)\ln(1-\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)))]}{\partial(\beta_0)}$$

$$= \frac{\partial(\sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right))}{\partial(\beta_0)} + \frac{\partial(\sum_{i=1}^{n}(1-y_i)\ln(1-\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)))}{\partial(\beta_0)}$$

$$= \frac{\sum_{i=1}^{n} y_i \partial(\ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right))}{\partial(\beta_0)} + \frac{\sum_{i=1}^{n}(1-y_i)\partial(\ln(1-\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)))}{\partial(\beta_0)}$$

$$= \sum_{i=1}^{n} y_i \left(\frac{\partial(\ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right))}{\partial(\beta_0)}\right) + \sum_{i=1}^{n}(1-y_i)\left(\frac{\partial(\ln(1-\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)))}{\partial(\beta_0)}\right)$$

$$= \sum_{i=1}^{n} y_i \left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right) - \sum_{i=1}^{n}(1-y_i)\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

$$= \sum_{i=1}^{n} y_i(1-\pi_i) - \sum_{i=1}^{n}(1-y_i)(\pi_i)$$

$$= \sum_{i=1}^{n} y_i - y_i\pi_i - \pi_i + y_i\pi_i \quad ; and$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i \qquad (1.2.3)$$

$$\frac{\partial(\ln(f(y_1, y_2, \dots, y_n)))}{\partial(\beta_1)} = \frac{\partial(\sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n}(1-y_i)\ln(1-\pi_i))}{\partial(\beta_1)}$$

$$= \sum_{i=1}^{n} y_i \left(\frac{\partial(\ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right))}{\partial(\beta_1)}\right) + \sum_{i=1}^{n}(1-y_i)\left(\frac{\partial(\ln(1-\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)))}{\partial(\beta_1)}\right)$$

$$= \sum_{i=1}^{n} y_i \left(\frac{x_i}{1+e^{\beta_0+\beta_1 x_i}}\right) - \sum_{i=1}^{n}(1-y_i)\left(\frac{e^{\beta_0+\beta_1 x_i} x_i}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

$$= \sum_{i=1}^{n} y_i(x_i)(1-\pi_i) - \sum_{i=1}^{n}(1-y_i)(x_i)(\pi_i)$$

$$= \sum_{i=1}^{n} y_i x_i - y_i x_i \pi_i - \pi_i x_i + y_i x_i \pi_i$$

$$= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \pi_i x_i \qquad (1.2.4)$$

Now all that remains is to set equations 1.2.3 and 1.2.4 equal to zero to solve for the unknown values of β_0 and β_1. Solving for the values β_0 and $\beta_1$requires an iterative procedure that is difficult and time consuming to do by hand. One common method included in most statistical software to calculate the values of β_0 and β_1 is the Newton Raphson method.

The Newton Raphson method uses the following equations:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0 \; ; and$$

$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \pi_i x_i = 0$$

to define the vector $\boldsymbol{q}$, where $\boldsymbol{q}' = (\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0 \, , \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \pi_i x_i = 0)$. This allows for the formation of the Hessian matrix, H, as defined below:

$$H = \begin{bmatrix} \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial \beta_0^2} & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial \beta_1^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i)}{\partial \beta_0} & \frac{\partial(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \pi_i x_i)}{\partial \beta_0} \\ \frac{\partial(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i)}{\partial \beta_1} & \frac{\partial(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \pi_i x_i)}{\partial \beta_1} \end{bmatrix}$$

$$= \begin{bmatrix} -\sum_{i=1}^{n} \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_i \pi_i(1-\pi_i) \\ -\sum_{i=1}^{n} x_i \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_i^2 \pi_i(1-\pi_i) \end{bmatrix}$$

The Newton Raphson iterative procedure aims to solve for the unknown vector $\boldsymbol{\beta} = (\beta_0, \beta_1)$ by first selecting an initial estimate for $\boldsymbol{\beta}$, which is denoted by $\boldsymbol{\beta}^{(0)}$. The procedure then uses the first two terms of the Taylor Series polynomial at $\boldsymbol{\beta}^{(0)}$ to produce a "better" solution; in that the estimate $\boldsymbol{\beta}^{(k)}$ will be closer to the actual solution as k gets larger. During the process, the $\boldsymbol{\beta}^{(k)th}$ estimate is given by

$$\boldsymbol{\beta}^{k} = \boldsymbol{\beta}^{(k-1)} - [\left(H^{(k-1)}\right)^{-1} * \boldsymbol{q}^{(k-1)}]$$

where $H^{(k-1)}$ and $\boldsymbol{q}^{(k-1)}$ denote the Hessian matrix H and the vector **q** evaluated at the prior estimate $\boldsymbol{\beta}^{(k-1)}$, respectively. The process continues in this manner until the estimates converge, which is the Newton Raphson estimate $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$. Having an estimate for $\boldsymbol{\beta}$ now allows for an estimate of π_i; specifically,

$$\hat{\pi}_i = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}}$$

1.3 Multiple Logistic Regression

Now that the theory of simple logistic regression has been discussed, multiple logistic regression may be understood more clearly. Whereas the simple logistic regression model uses one independent variable and one categorical dependent variable, multiple logistic regression can form models for cases where there are two or more independent variables that contribute to the discrete response variable.

Since it was already discussed in section 1.2 that the logistic regression model fits best when the dependent variable is binary and where Y follows a Bernoulli distribution with parameter π and probability of success, $P(Y = 1) = \pi$, many of the conclusions gathered still hold true as more independent variables are added. As such, the probability density function defined in equation 1.2.1 remains the same, with some minor adjustments in π_i.

For simplicity, it is assumed that there are p independent variables, $p = 1,2,3, \dots$, in a sample of n independent observations. This way, the vector $\boldsymbol{x_i}$ can be defined by a set of the independent variables, first augmented by constant 1, where

$$\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip})$$

such that the unknown vector $\boldsymbol{\beta}$ as in simple logistic regression is now defined as

$$\boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_p)$$

where there are now $p + 1$ parameters to solve for.

Specifically, the joint probability density function remains

$$f(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} {\pi_i}^{y_i} \, (1 - \pi_i)^{1-y_i}$$

Where now

- $\boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_p)$ *and* $\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip})$ *replace* $\boldsymbol{\beta} = (\beta_0, \beta_1)$ *and* x_i *from section* 1.2, *respectively*;
- $\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}$

  $= \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}$; *and*
- $\boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_p)$ *is estimated using maximum likelihood method*

Again, the maximum likelihood method is used to estimate the unknown vector $\boldsymbol{\beta}$. To recall, this is done by first applying the natural logarithm of both sides of equation 1.2.1 to produce equation 1.2.2, then deriving equation 1.2.2 with respect to β_0 then β_1, and finally setting these equal to zero to estimate- using statistical software- the values of β_0 and β_1. This is easily extended to the multiple logistic regression case by deriving equation 1.2.2 with respect to each of the unknowns $\beta_j, j = 1, \dots, p$, setting each equation to zero, and estimating for each unknown. Doing so produces the equations

$$\frac{\partial(\ln\left(f\left(y_1, y_2, \dots, y_n\right)\right))}{\partial\left(\beta_j\right)} = \begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0\,, & if\ j = 0\ ; and \\ \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \pi_i x_{ij} = 0, & if\ j = 1, \dots, p \end{cases}$$

As in the single variable case, the Newton Raphson method is used to estimate $\boldsymbol{\beta}$. However in multiple logistic regression, **q** is instead defined as

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

Which then allows for the formation of the new Hessian matrix, H:

$$H = \begin{bmatrix} \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_0^2} & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_0\partial\beta_1} & \cdots & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_0\partial\beta_p} \\ \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_1\partial\beta_0} & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_1^2} & \cdots & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_1\partial\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_p\partial\beta_0} & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_p\partial\beta_1} & \cdots & \frac{\partial^2(\ln(f(y_1, y_2, \dots, y_n)))}{\partial\beta_p^2} \end{bmatrix}$$

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

The process that the Newton Raphson method uses is the exact same from simple logistic regression in that is solves for $\boldsymbol{\beta}$ by first selecting an initial estimate, $\boldsymbol{\beta}^{(0)}$, and using an iterative process defined by $\boldsymbol{\beta}^{k} = \boldsymbol{\beta}^{(k-1)} - [\left(H^{(k-1)}\right)^{-1} * \boldsymbol{q}^{(k-1)}]$ to produce a better estimate, and again

$H^{(k-1)}$ and $\boldsymbol{q}^{(k-1)}$ represent the H and **q** evaluated at the prior estimate $\boldsymbol{\beta}^{(k-1)}$, respectively.

Once the process reaches convergence, the final estimate is the Newton Raphson estimate, $\widehat{\boldsymbol{\beta}}$**.**

The value for π_i can now be estimated as $\hat{\pi}_i = \frac{e^{(x_i'\widehat{\boldsymbol{\beta}})}}{1+ e^{(x_i'\widehat{\boldsymbol{\beta}})}}$.

Chapter 2- Assessing Significance in Logistic Regression

2.1 Assessing Significance

Although the model has provided an estimate for $\boldsymbol{\beta}$, the estimate may not be reliable. Thus, assessing the significance of the coefficients is useful to determine if the estimate is probable in application. Many significance tests can be constructed to determine if-and which-independent variables have a substantial impact on the dependent response variable.

It is first important to note that the above Hessian matrices in sections 1.2 and 1.3 are very useful for finding the standard error of $\widehat{\boldsymbol{\beta}}$, since the square roots of the diagonal elements of $-(H)^{-1}$ provide the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Ideally, these standard errors will be quite low. However, if these standard errors are high, the estimate $\widehat{\boldsymbol{\beta}}$ may not be reliable.

Furthermore, recall from linear regression that the P-Value is a good indicator of whether or not a specific independent variable has a significant effect on the dependent response variable. P-values can be used to determine whether the null hypothesis is rejected or not; the null hypothesis generally being that a specific independent variable has no significant effect on the response variable, which is compared to the alternate hypothesis that the specific independent variable does have a significant effect on the response variable. In this way, if the calculated P-Values are larger than a given, or required, significance level α, $(p > \alpha)$, then the null hypothesis that the independent variable in question does not have any significant effect on the response variable should not be rejected. However, if the P-Value is small $(p < \alpha)$, then the null hypothesis should be rejected and it can be concluded that the independent variable in question has a significant effect on the response variable in the model. For any given level of α (commonly $\alpha = 0.01, \alpha = 0.05, or\ \alpha = 0.10$), the P-Value represents the confidence of a relationship amongst a given independent variable and the dependent response variable. The

calculation of P-values and their interpretation are identical within the logistic regression model as from linear regression. As such, the P-Value test will not be discussed further.

Another important test for assessing significance of independent variables is the Chi-Square Hypothesis test. The Chi-Square Hypothesis test for logistic regression is comparable to the F-Test for linear regression in its ability to test multiple variables simultaneously to determine if at least one has a significant effect on the model. The F-Test in linear regression focusses on the difference between the sum of squares of the model including the variable and the sum of squares of the model not including the variable. In this way, a small F value shows no significance of the independent variable on the model, whereas a large F value shows that the alternate hypothesis being tested has a high significance on the dependent variable. Similarly, the Chi-Square Hypothesis test is used for logistic regression since it functions best on categorical data. Similar to the F-Test, a large Chi-Square value shows evidence of the alternate hypothesis that the independent variable has a significant effect in the model, whereas a small Chi-Square value shows that the null hypothesis should not be rejected. Exact "rules" of when to accept and reject the null hypothesis based on the Chi-Square Test will be examined further in sections 2.2 and 2.3.

The $Odds$ and $Odds\ Ratio$ are also useful in determining the likelihood of the success of an event and will be discussed in their relation towards the significance of the independent variables in the model. The $Odds$ and $Odds\ Ratio$ are important tools for assessing the change in the "log-odds" of the dependent response variable occurring, $Y_i = 1$, for a one unit increase in the independent variable x_i. In general, the $Odds$ of an event occurring can be defined as:

$$O = \frac{p}{1-p}, where\ p\ is\ the\ probability\ of\ success$$

Similarly, since the goal is to assess the change in the log-odds for a one unit increase in x_i, the odds ratio is defined as

$$Odds\ Ratio = \frac{Odds\ at\ (x_i + 1)}{Odds\ at\ x_i}$$

In this way, taking the natural logarithm of the odds ratio allows for an assessment of how an independent variable effects the dependent variable.

Finally, the Wald Test and Score Test are important in statistical software packages to assess significance of independent variables. The Wald Test gives an idea of how far the independent variable(s) are from the value in the null hypothesis (often 0). Using the Wald Test in statistical software packages for multiple logistic regression allows for multiple independent variables to be assessed concurrently. The Wald Test uses a similar concept as the Chi-Square Test in that it aims to estimate the value of the G statistic. However, in direct calculation of the G statistic, it is often required to calculate the difference between two models (discussed in section 2.3), whereas the Wald Test is able to approximate the G statistic using only one model. The Score Test also only uses one model, but instead it measures the improvement in the model if additional independent variables are added. Also often referred to as the Lagrange Multiplier Test, the test's value is calculated using the slope of the likelihood function.

2.2 Significance Tests in Simple Logistic Regression

To use the Chi-Square statistic to assess goodness of fit in simple logistic regression, it is first important to note that the deviance, $D = -2\sum_{i=1}^{n}[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)]$, in the likelihood ratio test for logistic regression is exactly equivalent to the sum of squares due to error (SSE) from linear regression. Thus, the G statistic for the Chi-Square hypothesis test can be

defined by calculating the difference between the deviance with and without the independent variable of interest. For the case of simple logistic regression, G can be defined as:

$$G = D(model\ without\ the\ variable) - D(model\ with\ the\ variable)$$
$$= -2\ln\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right] \quad (2.2.1)$$

When testing the null hypothesis of no significance of a single independent variable against the alternate hypothesis that the independent variable has significance, it is easy to see that if x_i is not significant in the model then the alternate hypothesis ($\beta_1 = 0$) is not rejected, such that

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

which is equivalent for any value of i and thus $\pi_i = \pi$.

Since the derivative of $\ln(\pi_i)$ with respect to β_0 is $\pi(1-\pi)$, it is clear that the likelihood function for β_0 is identical whether or not the variable x_i is included in the model. Thus, equation 1.2.3 can be recalled and redefined as:

$$\frac{\partial(\ln(f(y_1, y_2, \dots, y_n)))}{\partial(\beta_0)} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0$$
$$\Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi = 0$$
$$\Rightarrow \sum_{i=1}^{n} y_i - n\left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right) = 0$$
$$\Rightarrow \sum_{i=1}^{n} y_i = n\left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right)$$

Which allows for a simple substitution of β_0 above with $\hat{\beta}_0$ and rearranging the equation to solve for the estimate $\hat{\beta}_0$ as follows:

$$\hat{\beta}_0 = \ln\left(\frac{\sum_{i=1}^{n} y_i}{n - \sum_{i=1}^{n} y_i}\right)$$

Having an estimate for β_0 now allows for a proper definition of G. Equation 2.2.1 can be extended, using $\hat{\beta}_0$, to solve for G:

$$G = -2\ln\left[\frac{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^{\sum_{i=1}^{n} y_i}\left(\frac{n - \sum_{i=1}^{n} y_i}{n}\right)^{n-\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}\right] \tag{2.2.2}$$

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

When testing the null hypothesis against the alternate hypothesis $\beta_1 = 0$, G follows a chi-square distribution with 1 degree of freedom. In this way, a large G statistic (relative to a specified level alpha) indicates that the independent variable has a significant effect on the dependent response variable. Specifically, the following relation applies making it simple to analyze significance;

If, (for $\alpha, a\ given\ constant$):

$$\begin{cases} [\mathbb{P}(\chi^2(1) > G)] < \alpha \implies \text{the independent variable is significant.} \\ [\mathbb{P}(\chi^2(1) > G)] > \alpha \implies \text{the independent variable is not significant.} \end{cases} \tag{2.2.3}$$

Next recall that in chapter one, it was discussed that logistic regression uses a dependent Bernoulli distributed variable Y_i and the probability of success was then defined as $P(Y = 1) = \pi$. In this way, the $Odds$ in simple logistic regression can be expressed as:

$$O = \frac{\pi}{1-\pi}$$

Specifically, the $Odds$ at x_i is:

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

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

$$= e^{\beta_0+\beta_1 x_i}$$

Similarly, the $Odds$ at $(x_i + 1)$ is:

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

$$= e^{\beta_0+\beta_1 x_i+\beta_1}$$

$$= e^{\beta_0+\beta_1 x_i} e^{\beta_1}$$

So the $Odds\ Ratio$ is:

$$Odds\ Ratio = \frac{O(x_i + 1)}{O(x_i)}$$

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

$$= e^{\beta_1}$$

Taking the natural logarithm of each side yields the "log-odds":

$$\ln(Odds\ Ratio) = \beta_1$$

In application, this means that the unknown variable β_1 measures the change in the logarithmic odds of success for a one unit increase in x_i.

2.3 Significance Tests in Multiple Logistic Regression

The Chi-Square Hypothesis test and Odds Ratios can be extended into the case of multiple independent variables. Recall that under multiple logistic regression, with a set of p independent variables, then the following can be defined:

$$Set\ of\ p\ independent\ variables: \boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip})$$

$$Unknown\ vector\ value: \boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_p)$$

$$Density\ function: f(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} {\pi_i}^{y_i} (1 - \pi_i)^{1-y_i}; and$$

$$Paramater\ for\ Y_i: \pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}} = \frac{e^{\boldsymbol{x_i'\beta}}}{1+e^{\boldsymbol{x_i'\beta}}}$$

The Chi-Square Hypothesis test for multiple logistic regression tests the null hypothesis that none of the independent variables have a significant effect on the response variable against the alternate hypothesis that *at least* one of the independent variables is significant. In this way, the Chi-Square Hypothesis test shares many similarities with the F test from linear regression. As in simple logistic regression, the likelihood function is identical whether $\boldsymbol{x_i'}$ is included in the model or not, and so the estimate for β_0, $\hat{\beta}_0$, remains independent of $\boldsymbol{x_i'}$. Thus, the G-Statistic for multiple logistic regression is identical to the G statistic as defined in equation 2.2.2:

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

However, now when testing the null hypothesis against the alternate hypothesis $\beta_a = 0$, G follows a chi-square distribution with p degrees of freedom, where p is the number of independent variables. Thus, the significance of the independent variables can be determined by the relation of the probability of a chi-square distribution with p degrees of freedom being larger

than the G statistic compared to the specified level of alpha. Specifically, relation 2.2.3 can be easily extended into the multiple variables case where if, (for $\alpha, a\ given\ constant$):

$$\begin{cases} [\mathbb{P}(\chi^2(p) > G)] < \alpha \implies at\ least \text{ one independent variable is significant.} \\ [\mathbb{P}(\chi^2(p) > G)] > \alpha \implies \text{none of the independent variables are significant.} \end{cases}$$

Furthermore, it is often of interest which specific independent variable is significant in the model, so the G-Statistic can be refitted to assess the significance of a single variable. To do so, the null and alternate hypotheses must be redefined. After selecting a specific independent variable to assess, the likelihood ratio chi-square test compares the null hypothesis that the specific independent variable is not significant versus the alternate hypothesis that the specific independent variable is significant. To do so, a new G-statistic is defined as the difference between the G-statistic for the full model and the G-statistic for the "refitted" model. The G-statistic for the full model is identical to G as defined in equation 2.2.2, and is now denoted G_{full} The G-statistic is then recalculated for the model without the independent variable of interest, and is denoted $G_{refitted}$. In this way, the G-statistic of interest is the difference between G_{full} and $G_{refitted}$, denoted by $\Delta G = G_{full} - G_{refitted}$. When testing the null hypothesis that the variable is not significant against the alternate hypothesis that the variable is significant, ΔG follows a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between G_{full} and $G_{refitted}$. Recall G_{full} has p degrees of freedom, while $G_{refitted}$ has $(p-1)$ degrees of freedom. Then ΔG has $\Delta df = p - (p-1) = 1$ degrees of freedom. Thus, ΔG follows a chi-square distribution with 1 degree of freedom. In this case, the strategies for assessing significance of the independent variable of interest is exactly equivalent to the strategies from the single variable case defined in relation 2.2.3.

It is also often of interest to test the significance of a subset of the p independent variables. To do so, $G_{refitted}$ is redefined by calculating the G-statistic without the subset of independent variables of interest. Here, $\Delta df = p - k$, where k is the number of independent variables in the subset of interest. Thus, when testing the null hypothesis that none of the k independent variables are significant against the alternate hypothesis that *at least* one of the k independent variables are significant, ΔG follows a chi-square distribution with $p - k$ degrees of freedom. Here, to assess significance, the following relation holds;

If, (for $\alpha, a\ given\ constant$):

$$\begin{cases} [\mathbb{P}(\chi^2(p-k) > G)] < \alpha \implies at\ least \text{ one independent variable is significant.} \\ [\mathbb{P}(\chi^2(p-k) > G)] > \alpha \implies \text{none of the independent variables are significant.} \end{cases}$$

Finally, the $Odds$ and $Odds\ Ratio$ can be extended to the case of multiple logistic regression. Here, notation changes slightly as the odds ratio assesses a unit change in a single independent variable, holding all other independent variables constant. In this way, the $Odds$ at $x_{ij}, j \in [0, p]$, is:

$$O(x_{ij}) = \frac{\left(\frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1+\ e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}\right)}{\left(1-\frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1+\ e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}\right)}$$

$$= e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}$$

Similarly, the $Odds$ at $(x_{ij}+1)$ is:

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

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

$$= e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_j x_{ij}+\cdots+\beta_p x_{ip}} e^{\beta_j}$$

So the $Odds\ Ratio$ for multiple logistic regression is:

$$Odds\ Ratio = \frac{O(x_{ij}+1)}{O(x_{ij})}$$

$$= \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}} e^{\beta_j}}{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}$$

$$= e^{\beta_j}$$

Taking the natural logarithm of each side yields the "log-odds":

$$\ln(Odds\ Ratio) = \beta_j$$

Thus, similar to the log-odds in simple logistic regression, in multiple logistic regression the unknown variable β_j measures the change in the logarithmic odds of success for a one unit increase in the independent variable x_{ij}.

Chapter 3- Data and SAS

3.1 Introduction of Dataset

Next, to exemplify the theory and assessment from the previous two chapters, it is applied to a dataset of insurance claims. The raw dataset contains 67,856 observations, each of which has 10 variables. These 10 variables are:

Claim: a binary variable (0 or 1); 0 if no claim was made, 1 if at least one claim is made

Numclms: the number of claims made; 0 if no claim was made, integer values up to 4

Amount: value of claim; 0 if no claim was made, integer values

Exposure: the claimant's exposure; continuous between 0 and 1.

Value: value of car making claim, in $10,000's; continuous values

Vehicle: vehicle body; 13 categories (bus, convertible, coupe, hatchback, hardtop, motorized caravan, minibus, panel van, roadster, sedan, station wagon, truck, and utility)

Vehage: vehicle age; 1 (youngest), up to 4

Gender: gender of claimant; male or female

Area: claimant's geographical area of residence; categorical from A to F

Age: claimant's age; 1 (youngest), up to 6

Now, using this information, it is possible to set up a model to determine which variables have a significant effect on whether or not a claim is made. In this way, the dependent variable is the binary variable "Claim." This will allow the formation of a logistic regression model as discussed in chapter 1. However, including "Numclms" and "Amount" is redundant since these variables are just extensions of the dependent variable, including constraints. So the model will

be built using “Claim” as the dependent variable, and using the remaining 7 variables as the independent variables.

To get an idea of the variables and their ranges, a snapshot of the first 10 rows of the dataset is as follows:

Obs	value	exposure	clm	numclms	amount	vehicle	vehage	gender	area	age
1	1.0600	0.30390	0	0	0.00	HBACK	3	F	C	2
2	1.0300	0.64887	0	0	0.00	HBACK	2	F	A	4
3	3.2600	0.56947	0	0	0.00	UTE	2	F	E	2
4	4.1400	0.31759	0	0	0.00	STNWG	2	F	D	2
5	0.7200	0.64887	0	0	0.00	HBACK	4	F	C	2
6	2.0100	0.85421	0	0	0.00	HDTOP	3	M	C	4
7	1.6000	0.85421	0	0	0.00	PANVN	3	M	A	4
8	1.4700	0.55578	0	0	0.00	HBACK	2	M	B	6
9	0.5200	0.36140	0	0	0.00	HBACK	4	F	A	3
10	0.3800	0.52019	0	0	0.00	HBACK	4	F	B	4

Building the model in this way determines which (independent) variables have a significant effect in estimating the probability that a given claimant makes a claim or not. The model will be built by analyzing each variable one at a time, selecting and keeping the best one (or selecting none if none increase the G^2 statistic), and repeating this process until a final model is reached. Additionally, the interaction between certain independent variables will be analyzed in their significance in the model.

To use SAS to find the most appropriate model, each independent variable should be tested individually against the dependent “Claim” variable. In this way, the main portion of the code is as follows:

```
data claims;
infile 'p:\insurance1.dat';
input value exposure clm numclms amount vehicle $ vehage gender $ area $ age $ @@;
proc logistic descending;
class numclms vehicle vehage gender area age / param=ref ref=last;
model clm = exposure / lackfit covb;
run;
```

The first line of the code ("data claims;") defines the data file to which will be referenced in future lines. The second line calls the insurance claim dataset from the P drive. The third line defines each of the variables from the insurance claim dataset. Variables with a '$' succeeding the variable name denotes that the preceding variable is categorical rather than numerical. The fourth line is a built in procedure in SAS to build a logistic model that models the probability that Y=1, in other words the probability of a claimant making a claim. This line will print out the model information, the model fit statistics, the chi-square value for each of the 3 tests of interested (described in chapter 2) and their corresponding p-values, the analysis of the maximum likelihood estimates, and the odds ratio estimates. The fifth line defines the categorical variables again and instructs SAS to use the last category in each variable as the reference variable. Finally, the sixth line builds the logistic model with the independent variables of interest on the right side, and will print out the estimated covariance matrix, the partition for the Hosmer and Lemeshow test, and the goodness of fit statistic to analyze how well the model fits. This sixth line is the only line of code that changes throughout the analysis, depending on which independent variable(s) are being tested. In the above excerpt of code, exposure is being tested individually first, and SAS produces the following output:

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1199.1909	1	<.0001
Score	1199.9503	1	<.0001
Wald	1138.5274	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.6018	0.0362	9897.5104	<.0001
exposure	1	1.8421	0.0546	1138.5274	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
exposure	6.309	5.669	7.022

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
102.1805	8	<.0001

3.2 Selecting the Appropriate Model

For exposure, the G^2 value (likelihood ratio chi-square value) is $G^2 = 1199.1909$ on one degree of freedom with an associated p-value of $p - val < 0.0001$. The large G^2 statistic and the small p-value indicate that exposure is significant in the model. However, the rest of the 6 independent variables must be tested in an identical manner to determine which independent variable is estimated to be *most* significant. Repeating this process for each of the other 6 independent variables yields G^2 values all less than 1199.1909 (refer to appendix B for SAS output for each of the 7 individual models). Thus, the independent variable "exposure" is added into the model, and each of the remaining 6 are tested *with* "exposure". That is, for example, the model statement when testing "exposure" and "age" would change to:

```
model clm = exposure age / lackfit covb;
```

So "exposure" would remain in each model tested, while "age" would change for each of the remaining 6 variables tested. Then the process of selecting and accepting the variable with the highest G^2 value continues. In this dataset, this process continues and first accepts "exposure" into the model, then "age," "vehicle," "vehage," "area," and finally "value." Each of the G^2 values (and their associated degrees of freedom and p-values) are available for reference in Appendices B through H.

Additionally, it is noted that "gender" is rejected from the model, as the last model tested as the most recent model has $G^2 = 1394.2100, d.f = 30\ and\ p - val < 0.0001$, and adding "gender" to the model increases the G^2 statistic minimally, with an increase in the degrees of freedom. Specifically, the model including "gender" has $G^2 = 1394.5643, d.f = 31\ and\ p - val < 0.0001$. The increase in G^2 of 0.3543 is determined to not be worth the increase in degrees of freedom, and so "gender" is not included in the final model.

3.3 Interaction

At each stage of selection, the new variable added into the model is also analyzed in terms of its interaction with the other independent variables already added into the model. This is done by adding an extra term in the model statement. The model statement would now look as follows:

```
model clm = exposure age exposure*age / lackfit covb;
```

Adding the term "exposure*age" now gives information on the interaction between the 2 variables. Specifically, under the output for logistic procedure, there is a table titled "Joint Tests" where the chi-square statistic (and its associated degrees of freedom and p-value) are calculated based on Wald inference. For the example where "age" was determined to be the next most significant independent variable after exposure and was included into the model with exposure, the interaction between them is described in the last line of the following table:

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
exposure	1	114.6986	<.0001
age	5	32.3949	<.0001
exposure*age	5	8.1204	0.1497

The p-value for the "exposure*age" interaction variable is $p - val = 0.1497 > 0.10 = \alpha$ and so it is determined that there is no evidence of a significant interaction between "exposure" and "age" in this model. Examining the interaction between each of the newly added variables with the already accepted variables continues at each step, and the exact value for the interaction term is available for reference in Appendix I.

Through this process it is found that "vehicle," "area," and "value" have no evidence of interaction either with each other or with "exposure" or "age," while "vehage" is found to have evidence of a significant interaction between it and "exposure. The Joint Tests for "vehage*exposure" produces a p-value less than $\alpha = 0.1$ and so this interaction is not rejected, and is thus included in the model.

Next, the interaction between another independent variable with the interaction "exposure*vehage" is analyzed and determined that "area" also interacts with "exposure" and "vehage." However, similar as to why "gender" was rejected from the model, the trade-off for the increase in the G^2 statistic is not worth the increase in degrees of freedom. Specifically, the model with only the interaction "exposure*vehage" has $G^2 = 1394.2100, d.f = 30 \; and \; p - val < 0.0001$, while the model including "exposure*vehage*area" has $G^2 = 1404.1696$, $d.f = 42 \; and \; p - val < 0.0001$. The increase in the G^2 of 9.9596 is not worth the increase of 12 degrees of freedom, so the model including the interaction term "exposure*vehage*area" is rejected.

Thus, the final model includes the variables (in order) "exposure," "age," "vehicle," "vehage," "area," and "value," along with the interaction term "exposure*vehage." That is, the final model statement is:

```
model clm = exposure age vehicle vehage area value exposure*vehage / lackfit covb;
```

and the output for this model is:

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1394.2100	30	<.0001
Score	1408.5095	30	<.0001
Wald	1329.5191	30	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.9134	0.1264	958.7735	<.0001
exposure		1	1.5281	0.1060	207.7601	<.0001
age	1	1	0.5033	0.0731	47.4620	<.0001
age	2	1	0.2941	0.0652	20.3641	<.0001
age	3	1	0.2466	0.0635	15.0922	0.0001
age	4	1	0.2146	0.0632	11.5442	0.0007
age	5	1	-0.00415	0.0689	0.0036	0.9520
vehicle	BUS	1	1.2744	0.3811	11.1840	0.0008
vehicle	CONVT	1	-0.5845	0.6097	0.9192	0.3377
vehicle	COUPE	1	0.5920	0.1446	16.7576	<.0001
vehicle	HBACK	1	0.1787	0.0746	5.7410	0.0166
vehicle	HDTOP	1	0.3475	0.1131	9.4386	0.0021
vehicle	MCARA	1	0.8033	0.2970	7.3162	0.0068
vehicle	MIBUS	1	0.1362	0.1722	0.6260	0.4288
vehicle	PANVN	1	0.2641	0.1499	3.1051	0.0780
vehicle	RDSTR	1	0.2188	0.7499	0.0851	0.7705
vehicle	SEDAN	1	0.2010	0.0719	7.8094	0.0052
vehicle	STNWG	1	0.2322	0.0726	10.2178	0.0014
vehicle	TRUCK	1	0.1575	0.1156	1.8575	0.1729
vehage	1	1	-0.4021	0.1179	11.6249	0.0007
vehage	2	1	-0.0356	0.1030	0.1195	0.7296
vehage	3	1	-0.0990	0.0974	1.0330	0.3095
area	A	1	-0.0781	0.0726	1.1576	0.2820
area	B	1	0.0131	0.0737	0.0317	0.8587
area	C	1	-0.0410	0.0711	0.3324	0.5643
area	D	1	-0.1650	0.0795	4.3106	0.0379
area	E	1	-0.0898	0.0832	1.1664	0.2801
value		1	0.0305	0.0188	2.6307	0.1048
exposure*vehage	1	1	0.8653	0.1696	26.0425	<.0001
exposure*vehage	2	1	0.3534	0.1501	5.5439	0.0185
exposure*vehage	3	1	0.2882	0.1464	3.8764	0.0490

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
exposure	1	207.7601	<.0001
age	5	83.5295	<.0001
vehicle	12	37.2261	0.0002
vehage	3	13.8114	0.0032
area	5	11.4420	0.0433
value	1	2.6307	0.1048
exposure*vehage	3	26.2640	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age 1 vs 6	1.654	1.434	1.909
age 2 vs 6	1.342	1.181	1.525
age 3 vs 6	1.280	1.130	1.449
age 4 vs 6	1.239	1.095	1.403
age 5 vs 6	0.996	0.870	1.140
vehicle BUS vs UTE	3.577	1.695	7.548
vehicle CONVT vs UTE	0.557	0.169	1.841
vehicle COUPE vs UTE	1.808	1.361	2.400
vehicle HBACK vs UTE	1.196	1.033	1.384
vehicle HDTOP vs UTE	1.415	1.134	1.767
vehicle MCARA vs UTE	2.233	1.248	3.997
vehicle MIBUS vs UTE	1.146	0.818	1.606
vehicle PANVN vs UTE	1.302	0.971	1.747
vehicle RDSTR vs UTE	1.245	0.286	5.412
vehicle SEDAN vs UTE	1.223	1.062	1.408
vehicle STNWG vs UTE	1.261	1.094	1.454
vehicle TRUCK vs UTE	1.171	0.933	1.468
area A vs F	0.925	0.802	1.066
area B vs F	1.013	0.877	1.171
area C vs F	0.960	0.835	1.103
area D vs F	0.848	0.726	0.991
area E vs F	0.914	0.777	1.076
value	1.031	0.994	1.070

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
72.6603	8	<.0001

Chapter 4- Analysis and Interpretation

4.1 Expressing the Appropriate Model

Now that the appropriate model has been constructed, it is important to recognize the implications of the chosen model. As seen in chapter 3, the most fitting model includes the following variables: "exposure," "age," "vehicle," "vehage," "value," and "area". This means that the model can be represented by:

$$\text{logit}[\pi(x_i)] = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \sum_{j=0}^{p} \beta_j x_{ij}$$

Where $\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}} = \frac{e^{x_i'\boldsymbol{\beta}}}{1+e^{x_i'\boldsymbol{\beta}}}$; so equation 4.1.1 is equivalent to:

$$\pi_i = \frac{e^{\sum_{j=0}^{p}\beta_j x_{ij}}}{1+e^{\sum_{j=0}^{p}\beta_j x_{ij}}} \text{ so } \hat{\pi}_i = \frac{e^{\sum_{j=0}^{p}\hat{\beta}_j x_{ij}}}{1+e^{\sum_{j=0}^{p}\hat{\beta}_j x_{ij}}}$$

So the equation for the probability that a given claimant makes a claim is:

$$\hat{\pi}_i = \frac{e^{x_i'\boldsymbol{\beta}}}{1+e^{x_i'\boldsymbol{\beta}}}; with$$

$$\begin{aligned} e^{x_i'\boldsymbol{\beta}} = exp(&\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5} + \hat{\beta}_6 x_{i6} + \hat{\beta}_7 x_{i7} + \hat{\beta}_8 x_{i8} + \hat{\beta}_9 x_{i9} + \\ &\hat{\beta}_{10} x_{i10} + \hat{\beta}_{11} x_{i11} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{14} x_{i14} + \hat{\beta}_{15} x_{i15} + \hat{\beta}_{16} x_{i16} + \hat{\beta}_{17} x_{i17} + \\ &\hat{\beta}_{18} x_{i18} + \hat{\beta}_{19} x_{i19} + \hat{\beta}_{20} x_{i20} + \hat{\beta}_{21} x_{i21} + \hat{\beta}_{22} x_{i22} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{24} x_{i24} + \hat{\beta}_{25} x_{i25} + \\ &\hat{\beta}_{26} x_{i26} + \hat{\beta}_{27} x_{i27} + \hat{\beta}_{28} x_{i1} x_{i19} + \hat{\beta}_{29} x_{i1} x_{i20} + \hat{\beta}_{30} x_{i1} x_{i21}) \end{aligned}$$

Where x_{i1} is the exposure of the claimant (between 0 and 1); $x_{i2} = 1$ if the claimant is in the first (youngest) age category, and 0 otherwise; $x_{i3} = 1$ if the claimant is in the second age category, and 0 otherwise; $x_{i4} = 1$ if the claimant is in the third age category, and 0 otherwise; $x_{i5} = 1$ if the claimant is in the fourth age category, and 0 otherwise; $x_{i6} = 1$ if the claimant is in the fifth age category, and 0 otherwise; $x_{i7} = 1$ if the claimant drives a bus, and 0 otherwise; $x_{i8} = 1$ if

the claimant drives a convertible, and 0 otherwise; $x_{i9} = 1$ if the claimant drives a coupe, and 0 otherwise; $x_{i10} = 1$ if the claimant drives a hatchback, and 0 otherwise; $x_{i11} = 1$ if the claimant drives a hardtop, and 0 otherwise; $x_{i12} = 1$ if the claimant drives a motorized caravan, and 0 otherwise; $x_{i13} = 1$ if the claimant drives a minibus, and 0 otherwise; $x_{i14} = 1$ if the claimant drives a panel van, and 0 otherwise; $x_{i15} = 1$ if the claimant drives a roadster, and 0 otherwise; $x_{i16} = 1$ if the claimant drives a sedan, and 0 otherwise; $x_{i17} = 1$ if the claimant drives a station wagon, and 0 otherwise; $x_{i18} = 1$ if the claimant drives a truck, and 0 otherwise; $x_{i19} = 1$ if the claimant's vehicle is in the first (youngest) age category, and 0 otherwise, $x_{i20} = 1$ if the claimant's vehicle is in the second age category, $x_{i21} = 1$ if the claimant's vehicle is in the third age category, and 0 otherwise; $x_{i22} = 1$ if the claimant lives in geographical area A, and 0 otherwise; $x_{i23} = 1$ if the claimant lives in geographical area B, and 0 otherwise; $x_{i24} = 1$ if the claimant lives in geographical area C, and 0 otherwise; $x_{i25} = 1$ if the claimant lives in geographical area D, and 0 otherwise; $x_{i26} = 1$ if the claimant lives in geographical area E, and 0 otherwise, and x_{i27} is the value of the claimant's vehicle (in \$10,000's);. Additionally, the last 3 terms are the interaction terms between exposure and claimant age.

Thus, using the output given in chapter 3 and Appendix I, the an equation for $\hat{\pi}_i$ can be constructed by first substituting the SAS values for $\hat{\beta}_p$ $(p = 0,1,2, \dots ,30)$ into $e^{\boldsymbol{x_i'\beta}}$ producing:

$$e^{\boldsymbol{x_i'\beta}} = exp(-3.9134 + 1.5281x_{i1} + 0.5033x_{i2} + 0.2941x_{i3} + 0.2466x_{i4} + 0.2146x_{i5} - 0.00415x_{i6} + 1.2744x_{i7} - 0.5845x_{i8} + 0.5920x_{i9} + 0.1787x_{i10} + 0.3475x_{i11} + 0.8033x_{i12} + 0.1362x_{i13} + 0.2641x_{i14} + 0.2188x_{i15} + 0.2010x_{i16} + 0.2322x_{i17} + 0.1575x_{i18} - 0.4021x_{i19} - 0.0356x_{i20} - 0.0990x_{i21} - 0.0781x_{i22} + 0.0131x_{i23} - 0.0410x_{i24} - 0.1650x_{i25} - 0.0898x_{i26} + 0.0305x_{i27} + 0.8653x_{i1}x_{i19} + 0.3534x_{i1}x_{i20} + 0.2882x_{i1}x_{i21})$$

Now, the exact estimated equation for the probability that a given claimant makes a claim is:

$$\hat{\pi}_i = [exp(-3.9134 + 1.5281x_{i1} + 0.5033x_{i2} + 0.2941x_{i3} + 0.2466x_{i4} + 0.2146x_{i5} - 0.00415x_{i6} + 1.2744x_{i7} - 0.5845x_{i8} + 0.5920x_{i9} + 0.1787x_{i10} + 0.3475x_{i11} + 0.8033x_{i12} + 0.1362x_{i13} + 0.2641x_{i14} + 0.2188x_{i15} + 0.2010x_{i16} + 0.2322x_{i17} + 0.1575x_{i18} - 0.4021x_{i19} - 0.0356x_{i20} - 0.0990x_{i21} - 0.0781x_{i22} + 0.0131x_{i23} - 0.0410x_{i24} - 0.1650x_{i25} - 0.0898x_{i26} + 0.0305x_{i27} + 0.8653x_{i1}x_{i19} + 0.3534x_{i1}x_{i20} + 0.2882x_{i1}x_{i21})] \Big/ [1 + exp(-3.9134 + 1.5281x_{i1} + 0.5033x_{i2} + 0.2941x_{i3} + 0.2466x_{i4} + 0.2146x_{i5} - 0.00415x_{i6} + 1.2744x_{i7} - 0.5845x_{i8} + 0.5920x_{i9} + 0.1787x_{i10} + 0.3475x_{i11} + 0.8033x_{i12} + 0.1362x_{i13} + 0.2641x_{i14} + 0.2188x_{i15} + 0.2010x_{i16} + 0.2322x_{i17} + 0.1575x_{i18} - 0.4021x_{i19} - 0.0356x_{i20} - 0.0990x_{i21} - 0.0781x_{i22} + 0.0131x_{i23} - 0.0410x_{i24} - 0.1650x_{i25} - 0.0898x_{i26} + 0.0305x_{i27} + 0.8653x_{i1}x_{i19} + 0.3534x_{i1}x_{i20} + 0.2882x_{i1}x_{i21})] \quad (4.1.1)$$

Where each of the $x_{ij}{}'s$ are defined as above. For this model, a Hosmer-Lemeshow test shows no evidence of a lack of fit of the model (p-value < 0.0001)

4.2 Interpretation of Odds Ratios

Using SAS, it was found that 6 of the 7 variables are significant in estimating the probability that a given claimant makes a claim, as well as the interaction between exposure and vehicle age. The exposure of the claimant is the biggest determiner of whether or not a claim is made. However, the combination of exposure, claimant age, vehicle type, vehicle age, vehicle value, and geographical area are also significant in estimating this probability. Furthermore, SAS provides the Odds Ratio estimates for each variable. In particular, the following conclusions can be made regarding claimant age while keeping the other variables fixed (using figure 4.2.1 below):

- The odds of a claimant in the first (youngest) age category making a claim is estimated to be 1.654 times higher than the odds of a claimant in the last (oldest) age category. Additionally, it is noted that with 95% confidence, the odds of a claimant in the first (youngest) age category making a claim is anywhere between 1.434 and 1.909 times the odds of a claimant in the last (oldest) age category.
- The odds of a claimant in the second age category making a claim is estimated to be 1.342 times higher than the odds of a claimant in the last (oldest) age category. Additionally, it is noted that with 95% confidence, the odds of a claimant in the second age category making a claim is anywhere between 1.181 and 1.525 times the odds of a claimant in the last (oldest) age category.
- The odds of a claimant in the third age category making a claim is estimated to be 1.280 times higher than the odds of a claimant in the last (oldest) age category. Additionally, it is noted that with 95% confidence, the odds of a claimant in the third age category making a claim is anywhere between 1.130 and 1.449 times the odds of a claimant in the last (oldest) age category.
- The odds of a claimant in the fourth age category making a claim is estimated to be 1.239 times higher than the odds of a claimant in the last (oldest) age category. Additionally, it is noted that with 95% confidence, the odds of a claimant in the fourth age category making a claim is anywhere between 1.095 and 1.403 times the odds of a claimant in the last (oldest) age category.
- The odds of a claimant in the fifth age category making a claim is estimated to be 0.996 times the odds of a claimant in the last (oldest) age category. Additionally, it is noted that with 95% confidence, the odds of a claimant in the fifth age category

making a claim is anywhere between 0.870 and 1.140 times the odds of a claimant in the last (oldest) age category.

These conclusions indicate that a claimant in the first (youngest) age category is more likely to make a claim than a claimant in any of the other categories, while a claimant in the fifth age category is least likely to make a claim. This makes sense logically since new drivers typically have less experience, and old drivers are getting old in age as well and may be losing some motor functions or memory skills, but the drivers in the second to oldest age category are still functioning well cognitively while having a high amount of driving experience. Similar conclusions can be pulled for comparisons between different claimant ages, vehicle type, geographical area, and value, and are summarized in the following tables:

Odds Ratio Estimates for Claimant Age			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Age 1 vs Age 6	1.654	1.434	1.909
Age 2 vs Age 6	1.342	1.181	1.525
Age 3 vs Age 6	1.28	1.13	1.449
Age 4 vs Age 6	1.239	1.095	1.403
Age 5 vs Age 6	0.996	0.87	1.14
Age 1 vs Age 5	1.661	1.466	1.883
Age 2 vs Age 5	1.347	1.211	1.5
Age 3 vs Age 5	1.285	1.159	1.425
Age 4 vs Age 5	1.245	1.123	1.379
Age 1 vs Age 4	1.335	1.193	1.494
Age 2 vs Age 4	1.083	0.988	1.187
Age 3 vs Age 4	1.032	0.946	1.127
Age 1 vs Age 3	1.293	1.155	1.447
Age 2 vs Age 3	1.049	0.957	1.149
Age 1 vs Age 2	1.233	1.098	1.384

(4.2.1)

Odds Ratio Estimates for Vehicle Value			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Value	1.031	0.994	1.07

(4.2.2)

Odds Ratio Estimates for Geographical Area			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Area A vs Area F	0.925	0.802	1.066
Area B vs Area F	1.013	0.877	1.171
Area C vs Area F	0.96	0.835	1.103
Area D vs Area F	0.848	0.726	0.991
Area E v Area F	0.914	0.777	1.076
Area A vs Area E	1.012	0.895	1.144
Area B vs Area E	1.108	0.978	1.256
Area C vs Area E	1.05	0.932	1.183
Area D vs Area E	0.928	0.807	1.066
Area A vs Area D	1.091	0.975	1.22
Area B vs Area D	1.195	1.066	1.339
Area C vs Area D	1.132	1.016	1.261
Area A vs Area C	0.964	0.887	1.047
Area B vs Area C	1.056	0.969	1.15
Area A vs Area B	0.913	0.833	1

(4.2.3)

Odds Ratio Estimates for Vehicle Type			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Bus vs Utility	3.557	1.695	7.548
Convertible vs Utility	0.557	0.169	1.841
Coupe vs Utility	1.808	1.361	2.4
Hatchback vs Utility	1.196	1.033	1.384
Hardtop vs Utility	1.415	1.134	1.767
Motorized Caravan vs Utility	2.233	1.248	3.997
Minibus vs Utility	1.146	0.818	1.606
Panel Van vs Utility	1.302	0.971	1.747
Roadster vs Utility	1.245	0.286	5.412
Sedan vs Utility	1.223	1.062	1.408
Station Wagon vs Utility	1.261	1.094	1.454
Truck vs Utility	1.171	0.933	1.468
Bus vs Truck	3.055	1.429	6.531
Convertible vs Truck	0.476	0.143	1.584
Coupe vs Truck	1.544	1.125	2.119
Hatchback vs Truck	1.021	0.833	1.252
Hardtop vs Truck	1.209	0.931	1.571
Motorized Caravan vs Truck	1.908	1.049	3.469
Minibus vs Truck	0.979	0.679	1.411
Panel Van vs Truck	1.112	0.803	1.54
Roadster vs Truck	1.063	0.243	4.653
Sedan vs Truck	1.044	0.856	1.275
Station Wagon vs Truck	1.078	0.883	1.314
Bus vs Station Wagon	2.836	1.354	5.94
Convertible vs Station Wagon	0.442	0.135	1.446
Coupe vs Station Wagon	1.433	1.104	1.86
Hatchback vs Station Wagon	0.948	0.854	1.053
Hardtop vs Station Wagon	1.122	0.926	1.36
Motorized Caravan vs Station Wag	1.77	1.001	3.13
Minibus vs Station Wagon	0.909	0.661	1.249
Panel Van vs Station Wagon	1.032	0.785	1.358
Roadster vs Station Wagon	0.987	0.228	4.267
Sedan vs Station Wagon	0.969	0.883	1.064
Bus vs Sedan	2.925	1.397	6.125
Convertible vs Sedan	0.456	0.138	1.504
Coupe vs Sedan	1.478	1.139	1.918
Hatchback vs Sedan	0.978	0.902	1.06
Hardtop vs Sedan	1.158	0.954	1.404
Motorized Caravan vs Sedan	1.826	1.03	3.239
Minibus vs Sedan	0.937	0.681	1.289
Panel Van vs Sedan	1.065	0.813	1.395
Roadster vs Sedan	1.018	0.235	4.41
Bus vs Roadster	2.874	0.558	14.811
Convertible vs Roadster	0.448	0.069	2.923
Coupe vs Roadster	1.452	0.329	6.414
Hatchback vs Roadster	0.961	0.221	4.167
Hardtop vs Roadster	1.137	0.26	4.973
Motorized Caravan vs Roadster	1.794	0.374	8.611
Minibus vs Roadster	0.921	0.206	4.114
Panel Van vs Roadster	1.046	0.236	4.636
Bus vs Panel Van	2.747	1.256	6.005
Convertible vs Panel Van	0.428	0.126	1.449
Coupe vs Panel Van	1.388	0.963	2.001
Hatchback vs Panel Van	0.918	0.7	1.204
Hardtop vs Panel Van	1.087	0.788	1.499
Motorized Caravan vs Panel Van	1.715	0.915	3.213
Minibus vs Panel Van	0.88	0.585	1.325
Bus vs Minibus	3.121	1.402	6.947
Convertible vs Minibus	0.486	0.143	1.658
Coupe vs Minibus	1.577	1.057	2.353
Hatchback vs Minibus	1.043	0.756	1.44
Hardtop vs Minibus	1.235	0.861	1.773
Motorized Caravan vs Minibus	1.949	1.02	3.722
Bus vs Motorized Caravan	1.602	0.632	4.06
Convertible vs Motorized Caravan	0.25	0.067	0.923
Coupe vs Motorized Caravan	0.809	0.435	1.506
Hatchback vs Motorized Caravan	0.535	0.301	0.953
Hardtop vs Motorized Caravan	0.634	0.349	1.15
Bus vs Hardtop	2.527	1.184	5.394
Convertible vs Hardtop	0.394	0.119	1.308
Coupe vs Hardtop	1.277	0.935	1.744
Hatchback vs Hardtop	0.845	0.693	1.029
Bus vs Hatchback	2.991	1.427	6.27
Convertible vs Hatchback	0.466	0.141	1.543
Coupe vs Hatchback	1.512	1.161	1.969
Bus vs Coupe	1.979	0.908	4.312
Convertible vs Coupe	0.308	0.092	1.037
Bus vs Convertible	6.417	1.584	25.992

(4.2.4)

Odds Ratio Estimates for Vehicle Age (Holding Exposure Constant at 0.22)			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Vehicle Age 1 vs Vehicle Age 4	1.235819113	0.939805879	1.531832348
Vehicle Age 2 vs Vehicle Age 4	0.958727878	0.698653837	1.21880192
Vehicle Age 3 vs Vehicle Age 4	1.036237122	0.787357955	1.285116289
Vehicle Age 1 vs Vehicle Age 3	0.489259625	0.327444114	0.651075136
Vehicle Age 2 vs Vehicle Age 3	0.379559466	0.202679457	0.556439474
Vehicle Age 1 vs Vehicle Age 2	1.289019691	1.13007608	1.447963302

Odds Ratio Estimates for Vehicle Age (Holding Exposure Constant at 0.71)			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Vehicle Age 1 vs Vehicle Age 4	0.808751968	0.359068562	1.258435374
Vehicle Age 2 vs Vehicle Age 4	0.806288226	0.409209656	1.203366795
Vehicle Age 3 vs Vehicle Age 4	0.899764695	0.515954498	1.283574892
Vehicle Age 1 vs Vehicle Age 3	0.368748293	0.260798148	0.476698439
Vehicle Age 2 vs Vehicle Age 3	0.367624957	0.153359234	0.581890679
Vehicle Age 1 vs Vehicle Age 2	1.003055659	0.901441989	1.104669329

Odds Ratio Estimates for Vehicle Age (Holding Exposure Constant at 0.45)			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Vehicle Age 1 vs Vehicle Age 4	1.012796179	0.645634919	1.37995744
Vehicle Age 2 vs Vehicle Age 4	0.883883511	0.560272642	1.20749438
Vehicle Age 3 vs Vehicle Age 4	0.969776157	0.658185592	1.281366722
Vehicle Age 1 vs Vehicle Age 3	0.428444102	0.308874302	0.548013902
Vehicle Age 2 vs Vehicle Age 3	0.373910057	0.188881074	0.55893904
Vehicle Age 1 vs Vehicle Age 2	1.14584803	1.032086026	1.259610033

(4.2.5)

Odds Ratio Estimates for Exposure for a 0.25 unit increase			
		95% Wald Comfidence Limits	
	Point Estimate	C.I. Lower Bound	C.I. Upper Bound
Vehicle Age Category 1	1.819114783	1.786017389	1.852212178
Vehicle Age Category 2	1.600594304	1.574036583	1.627152024
Vehicle Age Category 3	1.574716097	1.549404291	1.600027902
Vehicle Age Category 4	1.465248716	1.438745178	1.491752253

(4.2.6)

It is important to note that to make comparisons against categories that are included as the indicator variables, the code must be changed in the class statement. Such a change sets the reference variable to one of the current indicator variables, rather than just setting it to the last category.

Furthermore, SAS does not directly provide the Odds Ratio Estimates for Vehicle Age or for Exposure since there is an interaction amongst these variables included in the model. In this situation, the point estimates and 95% Wald Confidence Limits must be computed by hand. This is done so by using the $\hat{\beta}$ estimates provided from SAS along with the Covariance Matrix, also provided by SAS. To compute the Odds Ratios for vehicle age, exposure must be held constant at some arbitrary value. Here, the Odds Ratios for Vehicle Age are computed using exposure levels of the 25^{th}, 50^{th}, and 75^{th} percentiles. To compute the Odds Ratios for Exposure, generally the odds would be calculated at an arbitrary exposure level and then recalculated for a one unit increase. However, exposure is a continuous value between 0 and 1, so an increase of 1 unit would be out of range. Instead, the odds are recalculated for a 0.1 unit increase in exposure. Then conclusions can be made for each of the 6 independent variables in the model.

Specifically, a claimant that drives a bus is most likely to make a claim, while a claimant that drives a convertible is least likely to make a claim, as observed in figure 4.2.4. Again this is logical since busses are large vehicles that may be hard to manoeuvre, while convertibles are generally sports cars and tend to be well cared for. Additionally, claimants in geographical area B are most likely to make a claim, while claimants in geographical area D are least likely to make a claim, as observed from figure 4.2.3. This cannot be confirmed by logic since no information is available on which geographical areas are urban, suburban, or rural. Using logic, it can be estimated that geographical area B is an urban area with a higher volume

of traffic, while geographical area D is a rural area with low traffic volume. Also, the odds of a claimant making at least one claim increases by 1.031 times for every $10,000 increase in vehicle value, seen in figure 4.2.2. It is concluded with 95% confidence that the odds of a claimant making at least one claim is between 0.994 and 1.070 times for every $10,000 increase in vehicle value (figure 4.2.2). Using the median exposure, it is estimated that for a fixed exposure level of 0.45, the odds of a vehicle in the first age category making a claim is the highest, while the odds of a vehicle in the second age category is the lowest (figure 4.2.5). This is intuitive since claimants driving brand new cars or very old may have more disposable income to buy these vehicles.

4.3 Discarded Variables

It was determined through the model building process that the gender variable could be discarded from the model, as it has no significance in determining the probability of a claimant making a claim. This is a surprising result as it is widely believed within the car insurance industry that gender plays a significant role, specifically that males are more likely to make claims. However, using the 67,856 observations in the dataset, it is clear that males are not any more likely than females to make insurance claims.

Moreover, at the beginning of chapter 3 it was noted that the "number of claims" and "claim amount" variables were discarded as they were essentially the dependent variable with constraints. It is important to recognize that any of these 3 variables ("claim," "numclms," and "amount") could have been used as the dependent variable instead. The values under "numclms" and "amount" will be 0 if no claim is made, and all other values indicate a claim was made. Thus, a simple SAS program can be constructed to turn either the "numclms" or "amount"

variables to binary by setting all values greater than zero to 1. In this way, the "numclms" and "amount" variables will be identical to the "claim" variable used throughout this analysis.

Conclusion

This project aimed to build an appropriate logistic regression model to estimate the probability of a person making a car insurance claim based on a given exposure level, driver age, vehicle type, vehicle type, vehicle value, driver's geographical area, and driver's gender. This is a very practical analysis in the insurance industry as insurance prices generally fluctuate based on the estimated probability of a given person making a claim. That is, the higher the estimated probability of a person making a claim, the higher their insurance rate will be.

It was discussed in chapter 1 that logistic regression is an important statistical method to model a dependent binary variable against one or more independent variables of interest. The logistic model was derived for simple logistic regression as well as for multiple logistic regression. The likelihood equations were derived using maximum likelihood estimation, and using the Newton-Raphson method, were set to zero to solve for the estimates $\beta_i,\ i = 0, \dots p$.

Chapter 2 went on to discuss ways to assess significance of the estimates found from the theory in chapter 1. Specifically, many hypothesis tests were discussed, including the Chi-Square hypothesis test (likelihood hypothesis test), the Score hypothesis test, the Wald hypothesis test, and the P-value hypothesis test. Additionally, the Odds and Odds ratios were derived in their relation to a one unit increase in any specific variable to determine how the probability of making a claim may change as a single independent variable changes.

Next, chapter 3 introduced the dataset used throughout the analysis. The dataset contains 67,856 rows of information on the following 10 variables: exposure, claim, number of claims, amount of claim, value of vehicle, type of vehicle, geographical area, gender, claimant age, and vehicle age. SAS was utilized to produce output for analyzing and building the most fitting model for the given data. The model was built by testing one individual variable at a time,

and building the model upwards, selecting a new variable to accept into the model at each step. The interaction between specific independent variables was also discussed in its relation towards the model and variable significance.

Finally, chapter 4 combined all the information discussed within the previous 3 chapters to consider the real-world implications of the model. Specifically, the equation for the final model is stated using the estimates for β_i as given in the SAS output. Conclusions were made based on the model estimates and Odds Ratio estimates provided by SAS. Finally, the Odds Ratios to form confidence intervals for the two variables included in the interaction term must be computed by hand (and using excel spreadsheets) to compute the point estimates and 95% Wald Confidence limits for both vehicle age and exposure.

In conclusion, the final model for determining the probability of a given claimant making a claim is given by equation 4.1.1. It was found that the most significant independent variables in the model are exposure, claimant age, vehicle type, vehicle age, area, and value, in that order. The only variable discarded from the model was gender, which proves surprising considering how the insurance industry often charges males higher insurance rates than they charge females. It was determined that the youngest and oldest drivers are, unsurprisingly, most likely to make a claim, as well as claimants who drive a bus, compared to claimants in other age categories and claimants who drive other vehicles, respectively.

References

Czepiel, Scott A. “Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation.” *Saedsayad*, 2002, saedsayad.com/docs/mlelr.pdf.

Farrell, Patrick. STAT 5602: Analysis of Categorical Data, Week 6 Logistic Regression Part 1. 2020. https://culearn.carleton.ca/moodle/mod/lesson/view.php?id=1505497

“FAQ: How are the Likelihood Ratio, Wald, and Lagrange Multiplier (Score) Tests Different and/or Similar?” *UCLA Statistical Consulting Group*, (n.d.), stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-andor-similar/.

Hosmer, David W., and Stanley Lemeshow. “Applied Logistic Regression, Second Edition.” *Wiley Series in Probability and Statistics*, John Wiley & Sons Inc., 2000, resource.heartonline.cn/20150528/1_3kOQSTg.pdf.

“Setting Reference Levels for CLASS Predictor Variables.” *SAS Institute Inc.*, 2009, support.sas.com/kb/37/108.html.

Zhu, Ji, and Trevor Hastie. “Kernel Logistic Regression and the Import Vector Machine.” *Advances in Neural Information Processing Systems*, 2002, papers.nips.cc/paper/2059-kernel-logistic-regression-and-the-import-vector-machine.pdf.

Appendices

Appendix A:

Macro code (using SAS) used in analysis

```
data claims;
infile 'p:\insurance1.dat';
input value exposure clm numclms amount vehicle $ vehage gender $ area $ age $ @@;
proc logistic descending;
class numclms vehicle vehage gender area age / param=ref ref=last;
model clm = exposure age vehicle vehage area value exposure*vehage / lackfit covb;
run;
```

Note that this code represents the final model. The model statement began testing claim against each individual variable, then against 2 variables, and so on.

Also note that class statement was altered in the reference command to compare Odds Ratios against categories other than the last one.

Appendix B:

Output from first stage of selection (in reading order: age; exposure; gender; value; vehage; vehicle; area) where exposure was selected based on likelihood ratio statistic.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	71.1300	5	<.0001
Score	71.3330	5	<.0001
Wald	70.8826	5	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1199.1909	1	<.0001
Score	1199.9503	1	<.0001
Wald	1138.5274	1	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.2875	1	0.5918
Score	0.2873	1	0.5919
Wald	0.2868	1	0.5923

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.6800	1	<.0001
Score	23.0681	1	<.0001
Wald	23.1287	1	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	26.3086	3	<.0001
Score	26.5296	3	<.0001
Wald	26.4762	3	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	39.5472	12	<.0001
Score	43.3441	12	<.0001
Wald	42.0411	12	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.1173	5	0.0028
Score	18.1602	5	0.0028
Wald	18.1263	5	0.0028

Appendix C:

Output from second stage of selection (in reading order: exposure with age; exposure with area; exposure with gender; exposure with value; exposure with vehage; exposure with vehicle) where age was selected based on likelihood ratio statistic.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1288.0431	6	<.0001
Score	1286.5361	6	<.0001
Wald	1221.0615	6	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1214.3945	6	<.0001
Score	1215.0766	6	<.0001
Wald	1152.8891	6	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1200.3095	2	<.0001
Score	1200.9952	2	<.0001
Wald	1139.5153	2	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1221.8819	2	<.0001
Score	1223.2147	2	<.0001
Wald	1160.5893	2	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1227.0485	4	<.0001
Score	1227.1226	4	<.0001
Wald	1164.3664	4	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1235.1634	13	<.0001
Score	1238.7082	13	<.0001
Wald	1174.0394	13	<.0001

Appendix D:

Output from third stage of selection (in reading order: exposure and age with vehage; exposure and age with area; exposure and age with value; exposure and age with vehicle; exposure and age with gender) where vehicle was selected based on likelihood ratio statistic.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1313.8280	9	<.0001
Score	1311.7637	9	<.0001
Wald	1244.9353	9	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1300.1034	11	<.0001
Score	1298.3783	11	<.0001
Wald	1232.2240	11	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1307.3730	7	<.0001
Score	1306.4089	7	<.0001
Wald	1239.8685	7	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1325.0779	18	<.0001
Score	1326.1398	18	<.0001
Wald	1257.0406	18	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1288.4984	7	<.0001
Score	1286.9341	7	<.0001
Wald	1221.4209	7	<.0001

Appendix E:

Output from fourth stage of selection (in reading order: exposure, age, and vehicle with area; exposure, age, and vehicle with value; exposure, age, and vehicle with vehage; exposure, age, and vehicle with gender) where vehage was selected based on likelihood ratio statistic.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1336.6312	23	<.0001
Score	1337.5060	23	<.0001
Wald	1267.7326	23	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1343.4127	19	<.0001
Score	1344.6726	19	<.0001
Wald	1274.5044	19	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1353.6445	21	<.0001
Score	1353.7817	21	<.0001
Wald	1283.2100	21	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1325.3462	19	<.0001
Score	1326.4165	19	<.0001
Wald	1257.2967	19	<.0001

Appendix F:

Output from fifth stage of selection (in reading order: exposure, age, vehicle, and vehage with area; exposure, age, vehicle, and vehage with value; exposure, age, vehicle, and vehage with gender, including interaction term between exposure and vehage) where area was selected based on likelihood ratio statistic.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1391.6637	29	<.0001
Score	1405.8604	29	<.0001
Wald	1326.9390	29	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1382.6707	25	<.0001
Score	1397.1320	25	<.0001
Wald	1318.8253	25	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1380.3896	25	<.0001
Score	1394.7549	25	<.0001
Wald	1316.5087	25	<.0001

Appendix G:

Output from sixth stage of selection (in reading order: exposure, age, vehicle, vehage, and area with gender; exposure, age, vehicle, vehage, and area with value, including interaction term between exposure and vehage) where value was selected based on likelihood ratio statistic.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1391.9021	30	<.0001
Score	1406.1169	30	<.0001
Wald	1327.1744	30	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1394.2100	30	<.0001
Score	1408.5095	30	<.0001
Wald	1329.5191	30	<.0001

Appendix H:

Output from final stage of selection (exposure, age, vehicle, vehage, area, and value with gender, including interaction term between exposure and vehage) where gender was NOT selected based on likelihood ratio statistic and degrees of freedom.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1394.5643	31	<.0001
Score	1408.8828	31	<.0001
Wald	1329.8655	31	<.0001

Appendix I:

SAS complete output for final model (exposure, age, vehicle, vehage, area, and value, including interaction term between exposure and vehage)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1394.5643	31	<.0001
Score	1408.8828	31	<.0001
Wald	1329.8655	31	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.9172	0.1265	958.2694	<.0001
exposure		1	1.5280	0.1060	207.7487	<.0001
age	1	1	0.5010	0.0732	46.8947	<.0001
age	2	1	0.2912	0.0653	19.8529	<.0001
age	3	1	0.2437	0.0637	14.6606	0.0001
age	4	1	0.2124	0.0633	11.2640	0.0008
age	5	1	-0.00569	0.0689	0.0068	0.9342
vehicle	BUS	1	1.2727	0.3811	11.1506	0.0008
vehicle	CONVT	1	-0.5964	0.6101	0.9558	0.3283
vehicle	COUPE	1	0.5865	0.1449	16.3820	<.0001
vehicle	HBACK	1	0.1707	0.0758	5.0714	0.0243
vehicle	HDTOP	1	0.3438	0.1133	9.2108	0.0024
vehicle	MCARA	1	0.7993	0.2971	7.2394	0.0071
vehicle	MIBUS	1	0.1295	0.1725	0.5637	0.4528
vehicle	PANVN	1	0.2651	0.1499	3.1284	0.0769
vehicle	RDSTR	1	0.2170	0.7499	0.0837	0.7723
vehicle	SEDAN	1	0.1943	0.0728	7.1215	0.0076
vehicle	STNWG	1	0.2274	0.0731	9.6828	0.0019
vehicle	TRUCK	1	0.1586	0.1156	1.8820	0.1701
vehage	1	1	-0.4045	0.1180	11.7530	0.0006
vehage	2	1	-0.0377	0.1030	0.1342	0.7141
vehage	3	1	-0.1007	0.0975	1.0675	0.3015
area	A	1	-0.0770	0.0726	1.1230	0.2893
area	B	1	0.0141	0.0737	0.0367	0.8481
area	C	1	-0.0400	0.0711	0.3161	0.5739
area	D	1	-0.1653	0.0795	4.3244	0.0376
area	E	1	-0.0898	0.0832	1.1644	0.2806
value		1	0.0312	0.0188	2.7524	0.0971
gender	F	1	0.0194	0.0326	0.3540	0.5518
exposure*vehage	1	1	0.8658	0.1696	26.0774	<.0001
exposure*vehage	2	1	0.3538	0.1501	5.5565	0.0184
exposure*vehage	3	1	0.2888	0.1464	3.8908	0.0486

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age 1 vs 6	1.650	1.430	1.905
age 2 vs 6	1.338	1.177	1.521
age 3 vs 6	1.276	1.126	1.446
age 4 vs 6	1.237	1.092	1.400
age 5 vs 6	0.994	0.869	1.138
vehicle BUS vs UTE	3.571	1.692	7.536
vehicle CONVT vs UTE	0.551	0.167	1.821
vehicle COUPE vs UTE	1.798	1.353	2.388
vehicle HBACK vs UTE	1.186	1.022	1.376
vehicle HDTOP vs UTE	1.410	1.129	1.761
vehicle MCARA vs UTE	2.224	1.242	3.981
vehicle MIBUS vs UTE	1.138	0.812	1.596
vehicle PANVN vs UTE	1.304	0.972	1.749
vehicle RDSTR vs UTE	1.242	0.286	5.402
vehicle SEDAN vs UTE	1.214	1.053	1.401
vehicle STNWG vs UTE	1.255	1.088	1.449
vehicle TRUCK vs UTE	1.172	0.934	1.470
area A vs F	0.926	0.803	1.068
area B vs F	1.014	0.878	1.172
area C vs F	0.961	0.836	1.105
area D vs F	0.848	0.725	0.991
area E vs F	0.914	0.777	1.076
value	1.032	0.994	1.070
gender F vs M	1.020	0.956	1.087

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
exposure	1	207.7487	<.0001
age	5	82.6449	<.0001
vehicle	12	36.9348	0.0002
vehage	3	13.9078	0.0030
area	5	11.5478	0.0415
value	1	2.7524	0.0971
gender	1	0.3540	0.5518
exposure*vehage	3	26.2973	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
73.0686	8	<.0001

Parameter	Intercept	exposure	age1	age2	age3	age4	age5	vehicleBUS	vehicleCONVT	vehicleCOUPE	vehicleHBACK	vehicleHDTOP	vehicleMCARA	vehicleMIBUS	vehiclePANVN	vehicleRDSTR	vehicleSEDAN	vehicleSTNWG	vehicleTRUCK	vehage1	vehage2	vehage3	areaA	areaB	areaC	areaD	areaE	value	vehage1 exposure	vehage2 exposure	vehage3 exposure
Intercept	0.015973	-0.00965	-0.00344	-0.00349	-0.00327	-0.00318	-0.00306	-0.00457	-0.00077	-0.00399	-0.0043	-0.0042	-0.00411	-0.00399	-0.00359	-0.00255	-0.00424	-0.00386	-0.00391	-0.0038	-0.004	-0.00436	-0.00412	-0.00419	-0.00415	-0.00434	-0.00439	-0.00044	0.006942	0.006575	0.006582
exposure	-0.00965	0.011239	0.000112	0.000106	0.000052	0.000061	9.90E-06	-0.00024	0.000137	0.000382	-0.00002	-0.00015	0.000283	0.000217	-0.00072	-0.00004	-0.00005	-0.00006	-0.00024	0.006592	0.006597	0.006583	0.000061	0.000045	0.000048	0.000019	0.000079	-5.86E-06	-0.01123	-0.01124	-0.01121
age1	-0.00344	0.000112	0.005336	0.003049	0.003026	0.003014	0.003001	-0.00021	-0.00007	0.000083	0.000192	0.000142	0.000683	0.000204	-0.00007	0.000225	0.000424	0.000253	-0.00008	-0.00008	-0.00006	-0.00008	0.000205	0.000206	0.000166	0.000279	0.000288	-0.00002	-0.00004	-0.00001	0.000011
age2	-0.00349	0.000106	0.003049	0.004246	0.003046	0.003019	0.003005	0.000198	0.000128	-2.06E-06	0.000217	0.000032	0.000646	0.000442	0.000135	-0.00005	0.000366	0.000091	0.000035	0.000033	-1.52E-06	-0.00006	0.000286	0.000231	0.000228	0.000353	0.000332	-0.00002	-0.00007	-0.00005	-8.65E-06
age3	-0.00327	0.000052	0.003026	0.003046	0.004028	0.003012	0.003002	0.000042	0.000122	-0.00006	0.00018	0.00004	0.000612	-0.00022	0.00001	0.000131	0.000264	-1.24E-06	0.000035	0.000013	-8.22E-06	-0.00005	0.000167	0.000122	0.000138	0.000219	0.000207	-0.00002	-0.00002	-5.77E-07	-6.78E-06
age4	-0.00318	0.000061	0.003014	0.003019	0.003012	0.00399	0.002989	0.000038	-0.00002	-0.00012	0.000121	0.000034	0.000408	-0.00012	-0.00006	-0.00011	0.000206	0.000044	-0.00004	-0.00002	2.20E-07	-0.00003	0.000099	0.000082	0.000057	0.000163	0.000157	-0.00002	2.31E-06	-0.00003	-0.00004
age5	-0.00306	9.90E-06	0.003001	0.003005	0.003002	0.002989	0.004747	-0.00014	0.000106	-9.30E-06	0.00014	0.00006	0.000227	0.000115	-0.00002	-0.00004	0.000178	0.000058	-0.00004	0.000019	0.000018	-0.00003	0.00004	-7.37E-06	5.09E-06	0.000073	0.000397	-0.00003	-0.00005	-0.00003	-5.62E-06
vehicleBUS	-0.00457	-0.00024	-0.00021	0.000198	0.000042	0.000038	-0.00014	0.145223	0.003792	0.004077	0.004113	0.004169	0.004122	0.004119	0.004181	0.003824	0.004122	0.004095	0.004181	0.000099	-0.00025	-0.00006	0.000504	0.000523	0.000455	0.000556	0.000387	0.000043	0.0001	0.00059	0.000328
vehicleCONVT	-0.00077	0.000137	-0.00007	0.000128	0.000122	-0.00002	0.000106	0.003792	0.371716	0.004972	0.002116	0.004584	0.007359	0.004898	0.003471	0.008996	0.003073	0.005616	0.004511	0.004104	0.003055	0.001193	-0.0012	-0.00107	-0.00092	-0.00053	-0.00044	-0.00224	0.000054	-0.00016	0.000518
vehicleCOUPE	-0.00399	0.000382	0.000083	-2.06E-06	-0.00006	-0.00012	-9.30E-06	0.004077	0.004972	0.020912	0.004161	0.004197	0.004378	0.00444	0.004258	0.0045	0.004212	0.004253	0.004099	0.000536	0.000585	0.000429	-0.00067	-0.00062	-0.00058	-0.00027	-0.00011	-0.0001	-0.0001	-0.00032	-0.00018
vehicleHBACK	-0.0043	-0.00002	0.000192	0.000217	0.00018	0.000121	0.00014	0.004113	0.002116	0.004161	0.005561	0.004088	0.003697	0.0041	0.00442	0.003748	0.004514	0.003988	0.004057	-0.00095	-0.00072	-0.00039	-0.00052	-0.00044	-0.00048	-0.00018	-0.00007	0.000355	0.000024	-0.00008	-0.00003
vehicleHDTOP	-0.0042	-0.00015	0.000142	0.000032	0.00004	0.000034	0.00006	0.004169	0.004584	0.004197	0.004088	0.012791	0.004281	0.004217	0.004161	0.004253	0.004133	0.004212	0.004156	0.000179	0.000075	-1.18E-06	0.00009	0.000136	0.00012	0.00019	0.000295	-0.00007	0.000155	0.0001	0.000054
vehicleMCARA	-0.00411	0.000283	0.000683	0.000646	0.000612	0.000408	0.000227	0.004122	0.007359	0.004378	0.003697	0.004281	0.088208	0.004402	0.004024	0.005095	0.00398	0.004486	0.004238	0.001456	0.001082	0.000482	-0.00025	-0.00021	-0.00029	-0.00034	-0.00033	-0.00051	-0.00032	-0.00043	-0.00017
vehicleMIBUS	-0.00399	0.000217	0.000204	0.000042	-0.00022	-0.00012	0.000015	0.004119	0.004898	0.00444	0.0041	0.004217	0.004402	0.029637	0.004264	0.004313	0.004186	0.004256	0.004101	0.000843	0.0007	0.000444	-0.00051	-0.00055	-0.00059	-0.0002	-0.00018	-0.00011	-0.00019	-0.00025	-0.00012
vehiclePANVN	-0.00359	-0.00072	-0.00007	0.000135	0.00001	-0.00006	-0.00002	0.004181	0.003471	0.004258	0.00442	0.004161	0.004024	0.004264	0.022458	0.004004	0.004363	0.004104	0.004113	-0.00052	-0.00055	-0.00035	-0.00054	-0.00052	-0.00055	-0.00019	-0.0001	0.000131	0.000693	0.000668	0.000439
vehicleRDSTR	-0.00255	-0.00004	0.000225	-0.00005	0.000131	-0.00011	-0.00004	0.003824	0.008996	0.0045	0.003748	0.004253	0.005095	0.004313	0.004004	0.562359	0.003976	0.004699	0.004227	0.000131	0.000527	0.000328	-0.00098	-0.00084	-0.00063	-0.00038	-0.00014	-0.00069	0.000559	-0.00007	0.000178
vehicleSEDAN	-0.00424	-0.00005	0.000424	0.000366	0.000264	0.000206	0.000178	0.004122	0.003073	0.004212	0.004514	0.004133	0.00398	0.004186	0.004363	0.003976	0.005175	0.00409	0.00409	-0.00053	-0.00048	-0.00031	-0.00048	-0.00044	-0.00046	-0.0002	-0.00009	0.000196	0.000061	0.000035	0.000037
vehicleSTNWG	-0.00386	-0.00006	0.000253	0.000091	-1.24E-06	0.000044	0.000058	0.004095	0.005616	0.004253	0.003988	0.004212	0.004486	0.004256	0.004104	0.004699	0.00409	0.005276	0.004174	0.000301	0.000143	-0.00012	0.000057	-0.0001	-0.00005	-0.00008	0.000034	0.000115	-0.00022	0.000063	0.000109	0.000105
vehicleTRUCK	-0.00391	-0.00024	-0.00008	0.000035	0.000035	-0.00004	-0.00004	0.004181	0.004511	0.004099	0.004057	0.004156	0.004238	0.004101	0.004113	0.004227	0.00409	0.004174	0.013359	-0.0001	-0.00012	-0.00008	0.000087	0.000084	0.00008	-0.00005	-0.0001	-0.00006	0.000296	0.000297	0.000064
vehage1	-0.0038	0.006592	-0.00008	0.000033	0.000013	-0.00002	0.000019	0.000099	0.004104	0.000536	-0.00095	0.000179	0.001456	0.000843	-0.00052	0.000131	-0.00053	0.000301	-0.0001	0.013907	0.005843	0.005275	-0.00006	-0.00008	-0.00008	-0.00007	-0.00007	-0.0007	-0.01707	-0.00656	-0.00645
vehage2	-0.004	0.006597	-0.00006	-1.52E-06	-8.22E-06	2.20E-07	0.000018	-0.00025	0.003055	0.000585	-0.00072	0.000075	0.001082	0.0007	-0.00055	0.000527	-0.00048	0.000143	-0.00012	0.005843	0.010603	0.005149	-0.00006	-0.00009	-0.00008	-0.00006	-0.00004	-0.00051	-0.00656	-0.01344	-0.00649
vehage3	-0.00436	0.006583	-0.00008	-0.00006	-0.00005	-0.00003	-0.00003	-0.00006	0.001193	0.000429	-0.00039	-1.18E-06	0.000482	0.000444	-0.00035	0.000328	-0.00031	0.000057	-0.00008	0.005275	0.005149	0.009494	-9.71E-06	4.80E-06	-0.00002	-0.00004	-6.83E-06	-0.00024	-0.00656	-0.00658	-0.01271
areaA	-0.00412	0.000061	0.000205	0.000266	0.000167	0.000099	0.00004	0.000504	-0.0012	-0.00067	-0.00052	0.00009	-0.00025	-0.00051	-0.00054	-0.00098	-0.00048	-0.0001	0.000087	-0.00006	-0.00006	-9.71E-06	0.005272	0.00427	0.004268	0.004169	0.004131	0.000049	-0.00008	-0.00002	0.000023
areaB	-0.00419	0.000045	0.000206	0.000231	0.000122	0.000082	-7.37E-06	0.000523	-0.00107	-0.00062	-0.00044	0.000136	-0.00021	-0.00055	-0.00052	-0.00084	-0.00044	-0.00005	0.000094	-0.00008	-0.00009	4.80E-06	0.00427	0.005427	0.004279	0.004179	0.004141	0.000077	-0.00008	0.00002	6.72E-07
areaC	-0.00415	0.000048	0.000166	0.000228	0.000136	0.000057	5.09E-06	0.000455	-0.00092	-0.00058	-0.00048	0.00012	-0.00029	-0.00059	-0.00055	-0.00063	-0.00046	-0.00008	0.00008	-0.00008	-0.00008	-0.00002	0.004268	0.004279	0.005454	0.004175	0.004137	0.000072	-0.00007	-7.36E-06	0.00004
areaD	-0.00434	0.000019	0.000279	0.000353	0.000219	0.000163	0.000073	0.000556	-0.00053	-0.00027	-0.00018	0.00019	-0.00034	-0.00002	-0.00019	-0.00038	-0.0002	0.000034	-0.00005	-0.00007	-0.00006	-0.00004	0.004169	0.004179	0.004175	0.006319	0.004105	0.000059	-0.00006	0.00003	0.000079
areaE	-0.00439	0.000078	0.000288	0.000332	0.000207	0.000157	0.000097	0.000387	-0.00044	-0.00011	-0.00007	0.000295	-0.00003	-0.00018	-0.0001	-0.00014	-0.00009	0.000115	-0.0001	-0.00007	-0.00004	-6.83E-06	0.004131	0.004141	0.004137	0.004105	0.006921	0.000049	-0.00008	-0.00003	-0.00002
value	-0.00044	-5.86E-06	-0.00002	-0.00002	-0.00002	-0.00002	-0.00003	0.000043	-0.00224	-0.0001	0.000355	-0.00007	-0.00051	-0.00011	0.000131	-0.00069	0.000196	-0.00022	-0.00006	-0.0007	-0.00051	-0.00024	0.000049	0.000077	0.000072	0.000059	0.000049	0.000353	-0.00002	3.86E-06	-0.00006
vehage1 exposure	0.006942	-0.01123	-0.00004	-0.00007	-0.00002	2.31E-06	-0.00005	0.0001	0.000054	-0.0001	0.000024	0.000155	-0.00032	-0.00019	0.000693	0.000559	0.000061	0.000063	0.000296	-0.01707	-0.00656	-0.00656	-0.00008	-0.00008	-0.00007	-0.00006	-0.00008	-0.00002	0.028748	0.011226	0.011208
vehage2 exposure	0.006575	-0.01124	-0.00001	-0.00005	-5.77E-07	-0.00003	-0.00003	0.00059	-0.00016	-0.00032	-0.00008	0.0001	-0.00043	-0.00025	0.000668	-0.00007	0.000035	0.000109	0.000297	-0.00658	-0.01344	-0.00658	-0.00002	0.00002	-7.36E-06	0.000032	-0.00003	3.86E-06	0.011226	0.022526	0.011211
vehage3 exposur	0.006582	-0.01121	0.000011	-8.65E-06	-6.78E-06	-0.00004	-5.62E-06	0.000028	0.000518	-0.00018	-0.00003	0.000054	-0.00017	-0.00012	0.000439	0.000178	0.000037	0.000105	0.000064	-0.00645	-0.00649	-0.01271	0.000023	6.72E-07	0.00004	0.000079	-0.00002	-0.00006	0.011208	0.011211	0.021432