CARLETON UNIVERSITY

SCHOOL OF MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Model-Based Estimation for a Relative Difference in Proportions

AUTHOR: Brianne Rogers

SUPERVISOR: Patrick Farrell

DATE: May 1, 2019

Contents

Abstract

The focus of this project lies in estimating the relative difference between proportions of a population using a logistic regression model. To do so, we use a multiple logistic regression model to gage the relationship between a set of independent explanatory variables and an indicator response variable, assuming the values 1 or 0 to represent a “success” or a “failure” respectively. A brief introduction to logistic regression will illustrate how we may study the contribution of a factor to the difference in proportions of a successful response variable by representing that factor as one of the explanatory variables in the regression model. Given values for this factor that may influence the response variable, we will manipulate the multiple logistic regression model to derive an expression for a relative difference in proportions that depends on the unknown parameters of the regression model. More specifically, we will consider how the exposure to some factor, measured as being absent or present, effects the response variable of interest and build a model for the relative difference between the proportion of the response that is a success given that the factor is present and the proportion of the response that is a success given that the factor is absent. Furthermore, we will fit the model for a relative difference in proportions so that it can be estimated. An expression for the bias of the estimate as well as an appropriate confidence interval for the true relative difference in proportions will be constructed and finally, we will generate data to test the suitability and limits of the model for a relative difference in proportions.

CHAPTER 1: Logistic Regression

1.1 Introduction to Simple Logistic Regression

Regression methods are often used to analyze the relationship between some response variable and one or more independent explanatory variables. In the event where the response variable of interest is discrete, assuming at least two possible values, logistic regression is particularly useful due to the mathematical advantages of the logistic distribution in terms of flexibility and easy implementation. Since the interest of this project lies in a dichotomous response, we will explore the case where the response variable has exactly two possible outcomes. Therefore, this can be represented as an indicator random variable taking on values 0 or 1.

First, consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = 0,1$$

where the response variable Y_i assumes the value either 0 or 1. Here, expected response $E(Y_i)$ is very intuitive for our application of logistic regression. Since $E(\varepsilon_i) = 0$, we have:

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{1.1}$$

Furthermore, we can assume Y_i is a Bernoulli distributed random variable such that:

$$P(Y_i = 1) = \pi_i \tag{1.2}$$

$$P(Y_i = 0) = 1 - \pi_i$$

Then, by the definition of the expected value of a random variable, we can derive:

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{1.3}$$

It follows from (1.1) and (1.3) that:

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

Therefore, when the response variable, Y_i, is an indicator variable that can assume exactly two values (0 or 1), the mean response is equivalent to the probability that $Y_i = 1$. It is easily shown that under these conditions, the error terms ε_i, have unequal variances and are not normally distributed since they only take on two possible values – for $Y_i = 0,1$. In addition, the mean response of Y_i is bounded between 0 and 1. As a result, a simple linear regression model is not feasible under these conditions since it does not respect these bounds and requires the assumptions of normality and constant variance. Rather, a model such as the logistic regression model in which the S-shaped distribution curve approaches the probabilities 0 and 1 asymptotically is more appropriate.

We can now consider the simple logistic regression model for a Bernoulli response variable Y_i with parameter $E(Y_i) = \pi_i$ which is given by:

$$Y_i = \mathrm{E}(Y_i) + \varepsilon_i$$

Indeed, ε_i may assume one of two possible values as previously mentioned. If $Y_i = 1$, we have $\varepsilon_i = 1 - \pi_i$ with probability π_i and if $Y_i = 0$ then $\varepsilon_i = -\pi_i$ with probability $1 - \pi_i$. Therefore, ε_i is a binomial distributed random variable with parameter π_i. Note that the error term ε_i depends on the Bernoulli distributed response variable Y_i and so, the simple logistic regression model may be rewritten as:

$$\mathrm{E}(Y_i) = \pi_i = \pi_i(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+ e^{\beta_0+\beta_1 x_i}} \tag{1.4}$$

where $Y_i \sim Bernoulli(\pi_i)$ are independent and the observations x_i are given constants.

It is also worth recognizing the logit transformation of π_i which represents the log odds that the response variable Y_i is equal to 1 given some value x_i. Observe that for some event that occurs with probability p, the odds of this event occurring is defined as the probability of success divided by the probability of failure. That is:

$$odds\ of\ an\ event = \frac{p}{1-p}$$

From equation (1.2), we know that Y_i is equal to 1 with probability π_i and so it follows that the odds that Y_i is equal to 1 given some value x_i is given by:

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

where $1 - \pi_i = \frac{1}{1+ e^{\beta_0+\beta_1 x_i}}$. Thus, the logit transformation may be written as:

$$\mathrm{logit}(x_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \ln\left(e^{\beta_0+\beta_1 x_i}\right) = \beta_0 + \beta_1 x_i \tag{1.5}$$

This transformation is a value that will be useful for finding appropriate estimators when fitting the simple logistic regression model.

1.2 Fitting the Simple Logistic Regression Model

Suppose we have a sample of n independent observations $(x_i, y_i), i = 1, \dots, n$ where y_i denotes the value of a Bernoulli response variable and x_i denotes the value of the explanatory variable for the i^{th} observation. For application purposes, we may assume the values of the outcome variable, 0 and 1, represent the absence or presence of some characteristic, respectively. As with any regression model, the unknown parameters must be estimated in order to fit the model to a set of data. To fit the logistic regression model in (1.4), we estimate the values for β_0 and β_1 using the maximum likelihood method of estimation. This method effectively manages the restrictions we encountered in Section 1.1 that arise when the response variable is an indicator whereas a method of estimation such as ordinary least squares, used for simple linear regression models, does not account for these conditions. The application of maximum likelihood begins with the construction of the likelihood function.

Using that each Y_i is a Bernoulli distributed random variable that assumes the values 0 and 1 with probabilities derived in (1.2), the probability mass function for Y_i is given by:

$$p_i(Y_i = y_i) = {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i} \qquad \mathrm{i} = 1, \dots, \mathrm{n}$$

This can be interpreted as the contribution of the pair (x_i, y_i) to the likelihood function and so, it follows that the likelihood function is given by their joint probability function. By independence, this is equivalent to the product of the individual contribution of the n observed pairs given by:

$$l(\beta) = \prod_{i=1}^{n} p_i(y_i) = \prod_{i=1}^{n} {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i} \tag{1.6}$$

where $\beta' = (\beta_0, \beta_1)$. The estimate of β is the value which maximizes the likelihood function in equation (1.6). However, it is mathematically simpler to evaluate the natural log of the likelihood function which yields the log-likelihood function defined as:

$$\begin{aligned} \mathrm{L}(\beta) = \ln[l(\beta)] &= \sum_{i=1}^{n}[\, y_i \ln(\pi_i) + (1 - y_i)\ln(1 - \pi_i)] \\ &= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1 - \pi_i) \end{aligned}$$

Using the logit transformation defined in equation (1.5) and the simplification $1 - \pi_i = (1 + e^{\beta_0 + \beta_1 x_i})^{-1}$, we can rewrite the above log-likelihood as:

$$\mathrm{L}(\beta) = \sum_{i=1}^{n} y_i\,(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0 + \beta_1 x_i}) \tag{1.7}$$

To find the maximum likelihood estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we differentiate equation (1.7) with respect to β_j for $j = 1,2$ as follows:

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

The values of β_0 and β_1 which maximize the two resulting expressions after differentiation with respect to both parameters are the maximum likelihood estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Therefore, after differentiating, we set both expressions equal to zero which yields the equations known as the likelihood equations:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}] = \sum_{i=1}^{n}[y_i - \pi_i(x_i)] = 0 \qquad (1.8)$$

and

$$\frac{\partial L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n}[y_i x_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} x_i] = \sum_{i=1}^{n} x_i[y_i - \pi_i(x_i)] = 0 \qquad (1.9)$$

Since equations (1.8) and (1.9) are nonlinear in the parameters β_0 and β_1, iterative methods are required to obtain the maximum likelihood estimates $\widehat{\beta'} = (\widehat{\beta}_0, \widehat{\beta}_1)$ which are included in most statistical software packages. One method to do so is the Newton-Raphson algorithm. To begin, this method uses the likelihood equations (1.8) and (1.9) to define:

$$q' = \left(\frac{\partial L(\beta)}{\partial \beta_0}, \frac{\partial L(\beta)}{\partial \beta_1}\right) = (\sum_{i=1}^{n}(y_i - \pi_i),\ \sum_{i=1}^{n} x_i(y_i - \pi_i))$$

and

$$H = \begin{bmatrix} \dfrac{\partial^2 L(\beta)}{\partial {\beta_0}^2} & \dfrac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_1} \\ \dfrac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_1} & \dfrac{\partial^2 L(\beta)}{\partial {\beta_1}^2} \end{bmatrix} = \begin{bmatrix} -\sum_{i=1}^{n} \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_i \pi_i(1-\pi_i) \\ -\sum_{i=1}^{n} x_i \pi_i(1-\pi_i) & -\sum_{i=1}^{n} {x_i}^2 \pi_i(1-\pi_i) \end{bmatrix}$$

To estimate β an initial estimate is made, say $\beta^{(0)}$ and an iterative procedure commences. At the j^th iteration of the procedure, we obtain $\beta^{(j+1)}$ using:

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

where $q^{(j)}$ and $H^{(j)}$ are the values of q and H evaluated at $\beta^{(j)}$, the jth approximation for β. When successive estimates of β converge, the algorithm terminates. If the termination occurs at iteration J, then $\widehat{\beta} = \beta^{(J)}$ is the maximum likelihood estimate for β.

Thus, now we can obtain the fitted value for π_i using $\widehat{\beta}$ given by:

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

It is also important to note that $-\mathrm{H}^{-1}$ evaluated at $\hat{\beta}$ will be an estimated asymptotic variance-covariance matrix that will be useful in later applications.

1.3 Multiple Logistic Regression

Now that we have a foundation in simple logistic regression, the case for more than one predictor variable is a simple extension. Indeed, several predictor variables are often needed in logistic regression to obtain intuitive predictions as there are usually many factors that influence our response of interest. When using a multiple logistic regression model, these predictor variables are flexible in terms of the characteristics they represent. For example, they might represent curvature or interaction effects and may be quantitative or qualitative indicator variables.

Let us consider a set of p independent predictor variables appended by a constant 1, denoted by the vector $X_i{}' = (1, X_{i1}, X_{i2}, \dots, X_{ip})$. Then, we can improve the Bernoulli distributed random variable with parameter π_i, which motivated the model for simple logistic regression, as follows:

$$Y_i = \mathrm{X}'\beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad Y_i = 0{,}1$$

where $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ is a vector of p+1 unknowns.

Note that $P(Y_i = 1) = E(Y_i) = \pi_i$ still holds and so the simple logistic regression model in expression (1.4) may be extended to a multiple logistic regression model as follows:

$$P(Y_i = 1) = \pi_i(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}} \tag{1.10}$$

where $Y_i \sim Bernoulli(\pi_i)$ are independent and the observations $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ip})$ are known constants by assumption. Similarly, the logit transformation in expression (1.5) can be easily extended to this (p+1)-variate model and is given by:

$$\mathrm{logit}(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Just as in simple logistic regression, we will utilize the maximum likelihood method of estimation to find appropriate estimates for the unknown p+1 parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.

The log-likelihood function given in expression (1.7) can be broadened in multiple logistic regression to obtain:

$$\mathrm{L}(\beta) = \sum_{i=1}^{n} y_i\,(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \sum_{i=1}^{n} \ln(1 + \ e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}})$$

where y_i and $x_i' = \left(1, x_{i1}, x_{i2}, \dots, x_{ip}\right)$ are given and $\beta' = \left(\beta_0, \beta_1, \beta_2, \dots, \beta_p\right)$.

Thus, the maximum likelihood estimates of the unknown parameters are obtained by differentiating the above log-likelihood function with respect to each parameter and solving the resulting p+1 expressions after they are set equal to 0. In doing so, we solve the following likelihood equations:

$$\sum_{i=1}^{n}[\,y_i - \pi_i(x_i)] = 0$$

$$\sum_{i=1}^{n} x_{ij}[\,y_i - \pi_i(x_i)] = 0$$

for $j = 1, \dots, p$.

As in the univariate case, at this point it becomes difficult to manually calculate the estimated coefficients due to the likelihood equations being nonlinear in the unknown parameters. Thus, we revisit the Newton-Raphson algorithm to find the maximum likelihood estimate for β. Just as the method for finding estimates for the simple logistic regression model, we use the likelihood equations to define:

$$\mathrm{q}' = \left(\frac{\partial \mathrm{L}(\beta)}{\partial \beta_0}, \frac{\partial \mathrm{L}(\beta)}{\partial \beta_1}, \dots, \frac{\partial \mathrm{L}(\beta)}{\partial \beta_p}\right) = \left(\sum_{i=1}^{n}(y_i - \pi_i)\,,\ \sum_{i=1}^{n} x_{i1}(\,y_i - \pi_i), \dots\,,\ \sum_{i=1}^{n} x_{ip}(y_i - \pi_i)\right)$$

and

$$\mathrm{H} = \begin{bmatrix} \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial {\beta_0}^2} & \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial \beta_0 \partial \beta_p} \\ \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial \beta_0 \partial \beta_1} & \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial {\beta_1}^2} & \cdots & \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial \beta_0 \partial \beta_p} & \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial \beta_1 \partial \beta_1} & \cdots & \dfrac{\partial^2 \mathrm{L}(\beta)}{\partial {\beta_p}^2} \end{bmatrix}$$

After simplifying each entry of H, we have:

$$\mathrm{H} = \begin{bmatrix} -\sum\limits_{i=1}^{n} \pi_i(1-\pi_i) & -\sum\limits_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \cdots & -\sum\limits_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) \\ -\sum\limits_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & -\sum\limits_{i=1}^{n} {x_{i1}}^2\pi_i(1-\pi_i) & \cdots & -\sum\limits_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) \\ \vdots & \vdots & \ddots & \vdots \\ -\sum\limits_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) & -\sum\limits_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) & \cdots & -\sum\limits_{i=1}^{n} {x_{ip}}^2\pi_i(1-\pi_i) \end{bmatrix} \tag{1.11}$$

The iterative procedure that follows is parallel to that of the univariate case. After making an initial guess for β, say $\beta^{(0)}$, the j^th iteration of the algorithm computes:

$$\beta^{(j+1)} = \beta^{(j)} - (\mathrm{H}^{(j)})^{-1}\mathrm{q}^{(j)}$$

where $\mathrm{q}^{(j)}$ and $\mathrm{H}^{(j)}$ are the values of q and H evaluated at $\beta^{(j)}$, the j^th approximation for β. The algorithm continues this calculation in each iteration until successive estimates of β converge, say at iteration J, at which point we have the maximum likelihood estimate $\hat{\beta} = \beta^{(J)}$. Using this estimate, we can estimate π_i as follows:

$$\hat{\pi}_i = \hat{\pi}_i(x_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_p x_{ip}}}{1+ e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_p x_{ip}}} = \frac{e^{x_i{}'\hat{\beta}}}{1+ e^{x_i{}'\hat{\beta}}}$$

As mentioned in the univariate case, a transformation of the matrix H evaluated at $\hat{\beta}$ will be a useful estimated asymptotic variance-covariance matrix. We will now elaborate on this idea as the multivariate case will be useful when estimating a relative difference in proportions. Consider the $(p + 1) \times (p + 1)$ matrix obtained from negating every entry of H. This matrix is called the observed information matrix, denoted $\mathrm{I} = -\mathrm{H}$. Indeed, the entries of this matrix are functions of β and it is well-known that taking the inverse of this information matrix yields the covariance matrix of the maximum likelihood estimates $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$. Therefore, evaluating I^{-1} at $\hat{\beta}$ provides an estimated covariance matrix for the maximum likelihood estimates given by:

$$\mathrm{I}^{-1}(\hat{\beta}) = -\mathrm{H}^{-1}(\hat{\beta}) = \begin{bmatrix} \hat{V}(\hat{\beta}_0) & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_p) \\ \widehat{cov}(\hat{\beta}_1, \hat{\beta}_0) & \hat{V}(\hat{\beta}_1) & \cdots & \widehat{cov}(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{cov}(\hat{\beta}_p, \hat{\beta}_0) & \widehat{cov}(\hat{\beta}_p, \hat{\beta}_1) & \cdots & \hat{V}(\hat{\beta}_p) \end{bmatrix}$$

where the i^{th} diagonal entry holds an estimate for the variance of $\hat{\beta}_{i-1}, i = 1, \dots, p + 1$, and the off-diagonal $(i, j)^{th}$ entry is the covariance of $\hat{\beta}_{i-1}$ and $\hat{\beta}_{j-1}$ for $i, j = 1, \dots, p + 1, i \neq j$.

CHAPTER 2: Estimating a Relative Difference in Proportions

2.1 Modelling a Relative Difference in Proportions

As seen in the previous chapter, a logistic regression model holds many properties that are desirable when studying the relationship between a dichotomous response variable, which assumes two possible outcomes, and a set of p independent explanatory variables. Thus, such a model will be appropriate for our analysis of the relationship between a categorical response variable, specified by presence or absence of some characteristic, and p independent factors that may influence the presence of such a characteristic. Furthermore, the multiple logistic regression model will allow us to study the case when a discrete explanatory variable is considered and we wish to investigate the relative difference between proportions of the response variable that are present given various values of the discrete explanatory variable. That is, we will be able to model the difference between the proportion of the response variable that is present given that an explanatory variable assumes some value and the proportion of the response variable that is present given that the explanatory variable assumes some other value, all taken relative to one of the previously mentioned proportions.

To see this, consider the Bernoulli distributed response variable Y_i and p independent explanatory variables augmented by a constant 1, $X_i{}' = (1, X_{i1}, X_{i2}, \dots, X_{ip})$, as defined in Section 1.3. When the values $X_{ij} = x_{ij}$ are given constants, the relationship between the Bernoulli response and the explanatory variables can be shown using the multiple logistic regression model π_i defined in equation (1.10). In addition, assume one of the explanatory variables, X_{ij} for $1 \leq j \leq p$, is discrete and assumes two possible values. Then, X_{ij} can be represented as an indicator variable which takes on the values 1 or 0, representing “success” or “failure” of some event, respectively. Let us define π_0 to be the probability that $Y_i = 1$ when $X_{ij} = 0$ given by:

$$\pi_0(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}{1+ e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}} \tag{2.1}$$

We can also define π_1 to be the probability that $Y_i = 1$ when $X_{ij} = 1$ given by:

$$\pi_1(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_j+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}{1+ e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_j+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}} \tag{2.2}$$

Notice that π_0 and π_1 only differ by a single term, the j^{th} power of the exponential. Both expressions will be useful in implementation when there is an interest in the relative difference between the probability that $Y_i = 1$ given that $X_{ij} = 1$ and the probability that $Y_i = 1$ given that $X_{ij} = 0$. Going forward, we will denote this relative difference by η. In practice, the final expression for η will depend on which proportion, π_0 or π_1, serves as a useful comparison relative to the difference being modelled. That is, if we wish to take η relative to the probability that $Y_i = 1$ given that $X_{ij} = 0$, it can be expressed as:

$$\eta(x_i) = \frac{\pi_1(x_i) - \pi_0(x_i)}{\pi_0(x_i)}$$

Similarly, if the informative comparison is with the probability that $Y_i = 1$ given that $X_{ij} = 1$, we can express η as done above, with the exception that the denominator is replaced by $\pi_1(x_i)$. This expression for η can be expanded and simplified as follows:

$$\eta(x_i) = \frac{\dfrac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_j+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_j+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}} - \dfrac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}}{\dfrac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}}$$

$$= \frac{e^{\beta_j}+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_j+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_{j-1}x_{i(j-1)}+\beta_j+\beta_{j+1}x_{i(j+1)}+\cdots+\beta_p x_{ip}}} - 1 \tag{2.3}$$

Thus, the multiple logistic regression model π_i evaluated at two values of a discrete explanatory variable of interest allows us to model the relative difference between the probability that a dichotomous response variable is present given that the explanatory variable is a “success” and the probability that the response is present given that the explanatory variable is a “failure”.

As a motivating example, let us consider the relationship between lung cancer, measured as being present or absent, and two independent health factors, smoking and age. More specifically, we study the proportion of individuals from a particular age group that have lung cancer given that they are a smoker in comparison to the proportion of individuals from the same age group that have lung cancer given that they do not smoke. Observe that for each individual, smoking is measured as a discrete variable assuming two possible values, smoker or non-smoker, while age is a continuous variable, on some specified interval, that is independent of smoking habits. Furthermore, we can assume that the explanatory variable which represents smoking habits is represented as an indicator variable that assumes the value 1 when an individual is a smoker or smoking is a “success” and assumes the value 0 when an individual is a non-smoker or smoking is a “failure”. Indeed, we are interested in the difference in the probability of an individual having lung cancer when exposure to a risk factor, smoking, is either present or absent, and the age of the individual is also taken into consideration. For an analysis such as this, the relative difference in probability will be an informative value to compare these probabilities. Thus, we will derive an expression using multiple logistic regression that is parallel to equation (2.3) which will model the relative difference between the proportion of individuals who have lung cancer given that they are smokers and the proportion of individuals who have lung cancer given that they are non-smokers.

Let us assume a sample size of n. Then, for $i = 1,2, \ldots, n$ we define:

$$Y_i = \begin{cases} 1, & \text{if individual } i \text{ has lung cancer} \\ 0, & \text{otherwise} \end{cases}$$

and

$$X_{i1} = \begin{cases} 1, & \text{if individual } i \text{ is a smoker} \\ 0, & \text{otherwise} \end{cases}$$

where Y_i and X_{i1} are indicator variables for which the value 1 represents presence and the value 0 represents absence of their corresponding characteristic. In addition, let X_{i2} represent the age of the i^{th} individual. We may consider Y_i to be independent Bernoulli distributed random variables with expected values $E(Y_i) = P(Y_i = 1) = \pi_i$. Thus, given a set of n observations $x_i' = (1, x_{i1}, x_{i2})$ we can model this relationship using the multiple logistic regression model in equation (1.10) to obtain:

$$\mathrm{P}(Y_i = 1) = \pi_i(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}{1+ e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}$$

for $i = 1,2, \ldots, n$.

Under this model, we may use equation (2.1) to define the probability that a non-smoker of age x_{i2} has lung cancer to be $\pi_0(x_{i2})$ given by:

$$\pi_0(x_{i2}) = \frac{e^{\beta_0+\beta_2 x_{i2}}}{1+ e^{\beta_0+\beta_2 x_{i2}}}$$

Similarly, we use equation (2.2) to define the probability that a smoker of age x_{i2} has lung cancer to be $\pi_1(x_{i2})$ given by:

$$\pi_1(x_{i2}) = \frac{e^{\beta_0+\beta_1+\beta_2 x_{i2}}}{1+ e^{\beta_0+\beta_1+\beta_2 x_{i2}}}$$

Therefore, we may represent the relative difference between the proportion of individuals who have lung cancer given that they are smokers and the proportion of individuals who have lung cancer given that they are non-smokers as:

$$\eta(x_{i2}) = \frac{\pi_1(x_{i2}) - \pi_0(x_{i2})}{\pi_0(x_{i2})}$$

Notice that we represent this difference in probability relative to the probability of having lung cancer for non-smokers, $\pi_0(x_{i2})$, since this is the interesting comparison when considering a difference in smoking habits. Therefore, using this model we will eventually be able to estimate the difference in the probability of having lung cancer between smokers and non-smoker relative to the probability of having lung cancer for a non-smoker. The above expression for η can be expanded and simplified as done in equation (2.3), to find an expression that will be simpler to estimate. Doing so yields:

$$\begin{aligned}
\eta(x_{i2}) &= \frac{\pi_1(x_{i2}) - \pi_0(x_{i2})}{\pi_0(x_{i2})} \\
&= \frac{\frac{e^{\beta_0+\beta_1+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_1+\beta_2 x_{i2}}} - \frac{e^{\beta_0+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_2 x_{i2}}}}{\frac{e^{\beta_0+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_2 x_{i2}}}} \\
&= \left(\frac{e^{\beta_0+\beta_1+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_1+\beta_2 x_{i2}}} - \frac{e^{\beta_0+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_2 x_{i2}}}\right)\left(\frac{1+e^{\beta_0+\beta_2 x_{i2}}}{e^{\beta_0+\beta_2 x_{i2}}}\right) \\
&= \frac{e^{\beta_0+\beta_1+\beta_2 x_{i2}} + e^{\beta_0+\beta_1+\beta_2 x_{i2}}\, e^{\beta_0+\beta_2 x_{i2}}}{(1+e^{\beta_0+\beta_1+\beta_2 x_{i2}})e^{\beta_0+\beta_2 x_{i2}}} - 1 \\
&= \frac{e^{\beta_0+\beta_2 x_{i2}}\left(e^{\beta_1} + e^{\beta_0+\beta_1+\beta_2 x_{i2}}\right)}{(1+e^{\beta_0+\beta_1+\beta_2 x_{i2}})e^{\beta_0+\beta_2 x_{i2}}} - 1 \\
&= \frac{e^{\beta_1} + e^{\beta_0+\beta_1+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_1+\beta_2 x_{i2}}} - 1 \qquad (2.4)
\end{aligned}$$

Using multiple logistic regression, a similar approach can be taken to represent the relationship between any dichotomous response variable of interest and some set of independent explanatory variables that have a relevant influence on the response. More importantly, when interest lies in the relative difference in the probability of the response variable being "present" or equivalently, assuming the value 1 when exposed to certain factors that are represented as explanatory variables, an expression parallel to the value we derived for η in equation (2.3) may be useful and can be found using the multiple logistic regression model. Thus, this model allows us to study how exposure to some explanatory variable, possibly a risk factor with regards to the response variable of interest, effects the probability of the response being present relative to the probability of the response being present when there is no risk or exposure. A similar study can be carried out where the effects of the explanatory variable on the probability of the response being present is taken relative to the probability of the response being present when there is an exposure to, or risk caused by, the explanatory variable. Although we explore the simple case where an indicator explanatory variable is used, minor adjustments can be made to extend this idea to include a discrete explanatory variable assuming any number of values, two of which we wish to analyze their effect on the response variable and the relative difference in the success of the response variable given those two values of the explanatory variable.

2.2 Fitting the Model and Constructing Confidence Intervals

In the previous section, we derived an expression, η, to model the relative difference in the probability of an indicator random variable assuming the value 1 when some influential factor is either present or absent. Since this model is a function of the multiple logistic regression model π_i evaluated at specific values of an explanatory variable, the unknown parameters of η are simply the unknown parameters of π_i. As seen in Section 1.3, these parameters can be estimated using the Newton-Raphson algorithm which finds the maximum likelihood estimate for $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$. To see this, we propose the following estimate for η:

$$\hat{\eta}(x_i) = \frac{\widehat{\pi_1}(x_i) - \widehat{\pi_0}(x_i)}{\widehat{\pi_0}(x_i)}$$

Indeed, π_0 and π_1 were defined using the multiple logistic regression model evaluated at some value of an indicator explanatory variable and so, they may be estimated using the maximum likelihood estimates from π_i. Using the simplified expression (2.3) that was derived for η, we have the following estimate for the relative difference between two proportions:

$$\hat{\eta}(x_i) = \frac{e^{\hat{\beta}_j} + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_j + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_j + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip}}} - 1 \qquad (2.5)$$

where $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ are the maximum likelihood estimates for the model π_i. With that said, the work to fit this model is done since we have already discussed how to find these parameter estimates and the values $x_{ij}, i = 1,2,..,n, j = 1,2,\dots,p,$ are known constants.

Although we have an estimate, $\hat{\eta}$, for a relative difference in proportions based on a sample of n observations, it is unlikely that the sample estimate will be exactly equal to the true population value η. Thus, we are also interested in constructing a confidence interval for the true difference in proportions. In particular, we will construct a 95% Wald interval for η which leans on the central limit theorem to justify the use of Gaussian quantiles. Recall that the multiple logistic regression model that we used to define a relative difference in probability is given in expression (1.10) as follows:

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

By the central limit theorem, we have that for a large sample size n, the maximum likelihood estimate $\hat{\pi}_i$ tends to a normal distribution with mean $\pi_i = E(Y_i)$ and variance $\frac{\pi_i(1-\pi_i)}{n}$. Therefore, using the linear properties of independent normal random variables as well as that our estimate for a relative difference in proportions $\hat{\eta}$ depends on the maximum likelihood estimate $\hat{\pi}_i$, it is easy to show that as the sample size n approaches infinity, the estimate $\hat{\eta}$ tends to a normal distribution with mean η. It follows that:

$$\frac{\hat{\eta}-\eta}{\sqrt{Var(\hat{\eta})}} \sim N(0,1)$$

where $N(0,1)$ denotes the standard normal distribution. This is the result of the central limit theorem that warrants the use of a Wald confidence interval. That is, given a significance level α, we can construct the following $(1-\alpha)\%$ Wald interval for η:

$$\hat{\eta} \pm z_{\alpha/2}\sqrt{Var(\hat{\eta})} \tag{2.6}$$

where $z_{\alpha/2}$ is the $\left(\frac{\alpha}{2}\right) th$ quantile of the standard normal distribution. Therefore, when we specify the significance level $\alpha = 0.05$ we have our desired 95% Wald interval for a relative difference in proportions.

Prior to computing this confidence interval, we must be able to find the variance of the estimate $\hat{\eta}$. It will also be useful to derive the expected value of $\hat{\eta}$ since these two values will be intuitive for studying our model for a relative difference in proportions as well as the accuracy of the estimate for the model. That is, the variance of $\hat{\eta}$ will provide a confidence interval for the relative difference in proportions while an expression for the expected value of $\hat{\eta}$ will allow us to measure the bias of our estimate of the relative difference in proportions as follows:

$$E(\hat{\eta}) - \eta$$

Indeed, if our estimate $\hat{\eta}$ were to be unbiased, the above expression would equate to zero. This would tell us that for a large sample size, we can expect the estimate for a relative difference in proportions, $\hat{\eta}$, to be equal to the true population proportion η and have no bias. That is, we would have $E(\hat{\eta}) = \eta$ and so, our measure for the bias of the estimate becomes $E(\hat{\eta}) - \eta = 0$. However, it may be that our estimate is biased, in which case we will want this bias to be small. Otherwise, when the bias is large this tells us that the expected value of our estimate $\hat{\eta}$ greatly differs from the true relative difference in proportions, invalidating the estimate. With the complexity of the expression we have for $\hat{\eta}$ given in equation (2.5), it will be difficult to directly compute its mean and variance. Since $\hat{\eta}$ is nonlinear in its parameters it will be more convenient to linearize the expression before performing any further computations. We will do so using the delta method for a function of a random vector.

Let us first derive the delta method in general terms so that we may easily apply its result to the model for a relative difference in proportions. We denote the statistic of interest by T_n, where the subscript expresses a dependence on the sample size n. Suppose that for a large sample size n, T_n converges in distribution to a normal distribution with mean μ and standard error $\sigma/\sqrt{n}$. This limiting distribution can alternatively be expressed as:

$$\sqrt{n}(T_n - \mu) \overset{d}{\to} N(0, \sigma^2)$$

Thus, for a function g we can derive the limiting distribution of $g(T_n)$ using a second order Taylor series expansion of $g(t)$ in a neighbourhood μ, for some μ^* between t and μ. Assuming g can be differentiated at least twice at μ, we have:

$$\begin{aligned} g(t) &= g(\mu) + (t - \mu)g'(\mu) + \frac{1}{2}(t - \mu)^2 g''(\mu^*) \\ &= g(\mu) + (t - \mu)g'(\mu) + O(|t - \mu|^2) \end{aligned}$$

Substituting the random variable T_n for t yields:

$$\sqrt{n}[g(T_n) - g(\mu)] = \sqrt{n}(T_n - \mu)g'(\mu) + O_p(n^{-\frac{1}{2}})$$

However, as n approaches infinity, the additive term $O_p(n^{-\frac{1}{2}})$ is asymptotically negligible and so, $\sqrt{n}[g(T_n) - g(\mu)]$ has the same limiting distribution as $\sqrt{n}(T_n - \mu)g'(\mu)$. Therefore, we have that:

$$\sqrt{n}[g(T_n) - g(\mu)] \xrightarrow{d} N(0, \sigma^2[g'(\mu)]^2)$$

Since $\sigma^2 = \sigma^2(\mu)$ and $g'(\mu)$ depend on μ, the asymptotic variance is unknown. However, when $\sigma = \sigma(\cdot)$ and $g'(\cdot)$ are continuous at μ then $\sigma(T_n)g'(T_n)$ is a consistent estimator of $\sigma(\cdot)g'(\cdot)$.

Since the estimate for a relative difference in proportions depends on the maximum likelihood estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$, we will be interested in deriving a Taylor series expansion of $\hat{\eta}$ about $\hat{\beta} = \beta$. The ideas of the delta method for a function of a random variable translate to a function of a random vector without much additional effort and so I will omit the full derivation. Essentially, we suppose that T_n is asymptotically multivariate normal with mean μ and covariance matrix Σ/n and that $g(t_1, \dots, t_N)$ has a nonzero differential $\phi = (\phi_1, \dots, \phi_N)^T$ at μ where:

$$\phi_i = \frac{\partial g}{\partial t_i}|_{t=\mu}$$

A second order Taylor series expansion about μ yields the following expression:

$$g(T_n) - g(\mu) = (T_n - \mu)^T\phi + o(||T_n - \mu||)$$

Therefore, for a large sample size n, $g(T_n)$ is multivariate normal with mean $g(\mu)$ and variance $\phi^T\Sigma\phi/n$. Again, the little-o term in the expansion is asymptotically insignificant and can be ignored for a large sample size. Clearly, the Taylor series expansion of $g(T_n)$ is linear in its parameters which is what we aimed to accomplish for our estimate of a relative difference in proportions. Thus, taking $g(T_n)$ to be our estimate $\hat{\eta}$ as a function of the maximum likelihood estimates $\widehat{\beta'} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ will allow us to use the delta method to linearize $\hat{\eta}$ without losing its properties. Indeed, the delta method is valid for an expansion about the maximum likelihood estimators $\hat{\beta}$ since they follow a large-sample normal distribution with covariance matrix given by the inverse of the information matrix that was discussed in Section 1.3.

The first step in finding the mean and variance of the estimated relative difference in proportions will be to linearize $\hat{\eta}$ using a second order Taylor series expansion about $\hat{\beta} = \beta$ that is parallel to the expansion derived in the delta method for a function of a random vector. Recall that the maximum likelihood estimate $\hat{\beta}$ was obtained using the Newton-Raphson method of estimation in which the matrix H in equation (1.11) contained the second partial derivatives, with respect to β, of expressions that were dependent on π_i. Therefore, we know that $\hat{\eta}$, which is also dependent on π_i, is at least twice differentiable at β. Furthermore, the second derivatives of $\hat{\eta}$ will be continuous and so $\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_j\partial\hat{\beta}_k}|_{\hat{\beta}=\beta} = \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_j\partial\hat{\beta}_k}|_{\hat{\beta}=\beta}$. For this reason, we can sum the mixed partial derivatives in the Taylor series expansion of $\hat{\eta}$, leaving them with a coefficient of 1 rather than $\frac{1}{2}$ as follows:

$$\hat{\eta}(x_i) \approx \hat{\eta}(x_i)|_{\hat{\beta}=\beta} + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}(\hat{\beta}_0-\beta_0) + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}(\hat{\beta}_1-\beta_1) + \cdots + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}(\hat{\beta}_p-\beta_p) +$$
$$\frac{1}{2}\left[\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0{}^2}|_{\hat{\beta}=\beta}(\hat{\beta}_0-\beta_0)^2 + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1{}^2}|_{\hat{\beta}=\beta}(\hat{\beta}_1-\beta_1)^2 + \cdots + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_p{}^2}|_{\hat{\beta}=\beta}(\hat{\beta}_p-\beta_p)^2\right] +$$
$$\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1) + \cdots + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}(\hat{\beta}_0-\beta_0)(\hat{\beta}_p-\beta_p) +$$
$$\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}(\hat{\beta}_1-\beta_1)(\hat{\beta}_2-\beta_2) + \cdots + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}(\hat{\beta}_1-\beta_1)(\hat{\beta}_p-\beta_p) + \cdots +$$
$$\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_{p-1}\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}(\hat{\beta}_{p-1}-\beta_{p-1})(\hat{\beta}_p-\beta_p)$$

Since this expression for $\hat{\eta}(x_i)$ is linear in its parameters, it will be much easier to find its expected value and variance. Taking the expected value of both sides of this Taylor series expansion will yield the expected value of $\hat{\eta}(x_i)$. By linearity of expectation, the right-hand side of this expression becomes a sum of the expected value of each term. Furthermore, the partial derivatives of $\hat{\eta}$ evaluated at $\hat{\beta} = \beta$ are constants and can be factored out of the expected value for each term. Using these properties, we have the following expression for the expected value of $\hat{\eta}(x_i)$:

$$E(\hat{\eta}(x_i)) \approx E(\hat{\eta}(x_i)|_{\hat{\beta}=\beta}) + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}E(\hat{\beta}_0-\beta_0) + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}E(\hat{\beta}_1-\beta_1) + \cdots +$$
$$\frac{\partial\hat{\eta}}{\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}E(\hat{\beta}_p-\beta_p) + \frac{1}{2}\left[\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0{}^2}|_{\hat{\beta}=\beta}E\left[(\hat{\beta}_0-\beta_0)^2\right] + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1{}^2}|_{\hat{\beta}=\beta}E\left[(\hat{\beta}_1-\beta_1)^2\right] + \cdots +\right.$$
$$\left.\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_p{}^2}|_{\hat{\beta}=\beta}E\left[(\hat{\beta}_p-\beta_p)^2\right]\right] + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)] + \cdots +$$
$$\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_p-\beta_p)] + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}E[(\hat{\beta}_1-\beta_1)(\hat{\beta}_2-\beta_2)] + \cdots +$$
$$\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}E[(\hat{\beta}_1-\beta_1)(\hat{\beta}_p-\beta_p)] + \cdots + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_{p-1}\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}E[(\hat{\beta}_{p-1}-\beta_{p-1})(\hat{\beta}_p-\beta_p)]$$

We will now recall a few definitions that will lead to a nice simplification for the expected value. First note that the estimate $\hat{\eta}$ evaluated at $\hat{\beta} = \beta$ is constant and equal to the true relative difference in proportions η. Thus, the first term in this expansion becomes the expected value of η which evaluates to η itself. Also, as conditioned in the delta method, $\hat{\beta}_j, j = 1, \dots, p$, converges to a normal distribution with mean β_j. This implies that the maximum likelihood estimates $\hat{\beta}_j$ are asymptotically unbiased and so, $E(\hat{\beta}_j) = \beta_j$ for a large sample size n. Therefore, we have:

$$E(\hat{\beta}_j - \beta_j) = E(\hat{\beta}_j) - \beta_j = 0$$

for all $j = 1, \dots, p$ as n approaches infinity. Also, recall that we may define the variance of a variable Y using its expected value as follows:

$$V(Y) = E\left[\left(\mathrm{Y} - E(\mathrm{Y})\right)^2\right]$$

In addition, the covariance between two variables, Y_1 and Y_2, is defined, using expectation, by the expression:

$$cov(Y_1, Y_2) = E\left[\left(Y_1 - \mathrm{E}(Y_1)\right)\left(Y_2 - \mathrm{E}(Y_2)\right)\right]$$

By the above definitions and the asymptotic unbiasedness of the maximum likelihood estimators we have:

$$E\left[\left(\hat{\beta}_j - \beta_j\right)^2\right] = E\left[\left(\hat{\beta}_j - E(\hat{\beta}_j)\right)^2\right] = V(\hat{\beta}_j)$$

and

$$E\left[(\hat{\beta}_j - \beta_j)(\hat{\beta}_k - \beta_k)\right] = E\left[\left(\hat{\beta}_j - E(\hat{\beta}_j)\right)\left(\hat{\beta}_k - E(\hat{\beta}_k)\right)\right] = cov(\hat{\beta}_j, \hat{\beta}_k)$$

for $j, k = 1, \dots, p, j \neq k$. Combining these results leads to a simplified expression for the expected value of the estimate for a relative difference in proportions given by:

$$E\left(\hat{\eta}(x_i)\right) \approx \eta(x_i) + \tfrac{1}{2}\textstyle\sum_{j=0}^{p} \frac{\partial^2 \hat{\eta}}{\partial \hat{\beta}_j^{\,2}} |_{\hat{\beta}=\beta} V(\hat{\beta}_j) + \textstyle\sum_{j=0}^{p-1} \sum_{k>j}^{p} \frac{\partial^2 \hat{\eta}}{\partial \hat{\beta}_j \partial \hat{\beta}_k} |_{\hat{\beta}=\beta} cov(\hat{\beta}_j, \hat{\beta}_k) \qquad (2.7)$$

All that remains is to derive the variance and covariance of the maximum likelihood estimates. However, as seen at the end of Section 1.3, we can obtain estimates for these values by taking the inverse of the observed information matrix I, derived in the Newton-Raphson algorithm for finding the maximum likelihood estimates, and evaluating it at the estimate $\hat{\beta}$. Finally, substituting the estimates, $\hat{V}(\hat{\beta}_j)$ and $\widehat{cov}(\hat{\beta}_j, \hat{\beta}_k), j, k = 1, \dots, p$, obtained from $\mathrm{I}^{-1}(\hat{\beta})$ into equation (2.7) yields the final expression for the expected value of an estimated relative difference in proportions. More importantly, subtracting $\eta(x_i)$ from both sides of equation (2.7) provides a measure for the bias of the estimate $\hat{\eta}$, as discussed earlier in this section.

Similarly, taking the variance of both sides of the second order Taylor series expansion of $\hat{\eta}$ gives us an expression for the variance of an estimated relative difference in proportions. However, when taking the variance of the right-hand side of the second order expansion, we will be left will the sum of the variances of each term plus the sum of the covariance between every pair of terms in the expansion. To evaluate this, it would require finding the variance of the squared maximum likelihood estimates as well as many other cross covariance terms that we do not have estimates for. Therefore, using the result of the delta method which showed that the second order terms of the Taylor series expansion are asymptotically insignificant, we will omit the second order terms and use a first order Taylor series expansion to derive the variance. Ignoring the second partial derivatives in our second order Taylor series expansion of $\hat{\eta}$ provides a first order expansion. Then, taking the variance of both sides yields:

$$V(\hat{\eta}(x_i)) \approx V(\hat{\eta}(x_i)|_{\hat{\beta}=\beta}) + \left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_0 - \beta_0) + \left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_1 - \beta_1) + \cdots +$$
$$\left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_p}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_p - \beta_p) + 2\sum_{j=0}^{p-1}\sum_{k>j}^{p}\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_j}|_{\hat{\beta}=\beta}\right)\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_k}|_{\hat{\beta}=\beta}\right) cov(\hat{\beta}_j - \beta_j, \hat{\beta}_k - \beta_k)$$

As previously mentioned, the estimate $\hat{\eta}$ evaluated at $\hat{\beta} = \beta$ is simply the true relative difference in proportions η. Since this is a constant in our expansion, its variance is zero. Also, using the definitions for variance and covariance as well as the asymptotic unbiasedness of the maximum likelihood estimators we have:

$$V(\hat{\beta}_j - \beta_j) = E\left[\left((\hat{\beta}_j - \beta_j) - E(\hat{\beta}_j - \beta_j)\right)^2\right] = E\left[\left(\hat{\beta}_j - E(\hat{\beta}_j)\right)^2\right] = V(\hat{\beta}_j)$$

and

$$cov(\hat{\beta}_j - \beta_j, \hat{\beta}_k - \beta_k) = E\left[\left(\hat{\beta}_j - E(\hat{\beta}_j)\right)\left(\hat{\beta}_k - E(\hat{\beta}_k)\right)\right] = cov(\hat{\beta}_j, \hat{\beta}_k)$$

for $j, k = 1, \ldots, p$. Making these substitutions yields the following expression for the variance of the estimated relative difference in proportions:

$$V(\hat{\eta}(x_i)) \approx \sum_{j=0}^{p}\left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_j}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_j) + 2\sum_{j=0}^{p-1}\sum_{k>j}^{p}\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_j}|_{\hat{\beta}=\beta}\right)\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_k}|_{\hat{\beta}=\beta}\right) cov(\hat{\beta}_j, \hat{\beta}_k) \quad (2.8)$$

Just as with the expected value, we may estimate the variance and covariance of the maximum likelihood estimates using the inverse of observed information matrix from Section 1.3.

Therefore, replacing these terms in equation (2.8) with their corresponding estimates $\hat{V}(\hat{\beta}_j)$ and $\widehat{cov}(\hat{\beta}_j, \hat{\beta}_k)$, $j, k = 1, \ldots, p$, obtained from $\mathrm{I}^{-1}(\hat{\beta})$ yields the final expression for the variance of an estimated relative difference in proportions. Finally, we can use this expression for variance along with the estimate $\hat{\eta}$ given in equation (2.5) to construct a 95% Wald confidence interval that is parallel to equation (2.6).

To illustrate this, let us return to the example in which we want to estimate the relative difference between the probability of an individual having lung cancer given they are a smoker and the probability of an individual having lung cancer given they are a non-smoker. Given the model we derived for this relative difference in expression (2.4), the estimated relative difference directly follows from equation (2.5) and is given by:

$$\hat{\eta}(x_{i2}) = \frac{e^{\hat{\beta}_1} + e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2 x_{i2}}}{1+ e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2 x_{i2}}} - 1 \tag{2.9}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the maximum likelihood estimators of β_0, β_1, and β_2 respectively. We may also find a 95% confidence interval to estimate the true relative difference in proportions of individuals with lung cancer, η, using the Wald interval given in expression (2.6). To do so, we will need to find the variance of the estimate $\hat{\eta}(x_{i2})$ using the result of the delta method. The corresponding Taylor series expansion for the estimate of this model is given by:

$$\begin{aligned}
\hat{\eta}(x_{i2}) \approx {} & \hat{\eta}(x_{i2})|_{\hat{\beta}=\beta} + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}\left(\hat{\beta}_0 - \beta_0\right) + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\left(\hat{\beta}_1 - \beta_1\right) + \frac{\partial\hat{\eta}}{\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\left(\hat{\beta}_2 - \beta_2\right) + \\
& \frac{1}{2}\left[\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0{}^2}|_{\hat{\beta}=\beta}\left(\hat{\beta}_0 - \beta_0\right)^2 + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1{}^2}|_{\hat{\beta}=\beta}\left(\hat{\beta}_1 - \beta_1\right)^2 + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_2{}^2}|_{\hat{\beta}=\beta}\left(\hat{\beta}_2 - \beta_2\right)^2\right] + \\
& \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_1 - \beta_1\right) + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_2 - \beta_2\right) + \\
& \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\left(\hat{\beta}_1 - \beta_1\right)\left(\hat{\beta}_2 - \beta_2\right)
\end{aligned}$$

Then, using equation (2.7), we can find an expression for the expected value of the estimated relative difference between proportions of individuals with lung cancer given their smoking status as follows:

$$\begin{aligned}
E\left(\hat{\eta}(x_{i2})\right) \approx {} & \eta(x_{i2}) + \frac{1}{2}\left[\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0{}^2}|_{\hat{\beta}=\beta}\hat{V}\left(\hat{\beta}_0\right) + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1{}^2}|_{\hat{\beta}=\beta}\hat{V}\left(\hat{\beta}_1\right) + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_2{}^2}|_{\hat{\beta}=\beta}\hat{V}\left(\hat{\beta}_2\right)\right] + \\
& \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\widehat{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_0\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\widehat{cov}\left(\hat{\beta}_0, \hat{\beta}_2\right) + \frac{\partial^2\hat{\eta}}{\partial\hat{\beta}_1\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\widehat{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)
\end{aligned}$$

where $\hat{V}(\hat{\beta}_j)$ and $\widehat{cov}(\hat{\beta}_j, \hat{\beta}_k)$, $j, k = 1,2,3$, are the estimates obtained from the inverse of the observed information matrix evaluated at $\hat{\beta}$. The partial derivatives in the Taylor series expansion are well-defined and can be derived as shown in Appendix I. Therefore, evaluating the partial derivatives at $\hat{\beta} = \beta$ yields the following expression for the bias of $\hat{\eta}$:

$$\begin{aligned} &E(\hat{\eta}(x_{i2})) - \eta(x_{i2}) \\ &\qquad \approx \frac{1}{2}\Big[\pi_1(1-\pi_1)(1-2\pi_1)(1-e^{\beta_1})\hat{V}(\hat{\beta}_0) + (1-\pi_1)[e^{\beta_1}(1-\pi_1) \\ &\qquad + \pi_1]\,(1-2\pi_1)\hat{V}(\hat{\beta}_1) + x_{i2}{}^2\pi_1(1-\pi_1)(1-2\pi_1)(1-e^{\beta_1})\hat{V}(\hat{\beta}_2)\Big] \\ &\qquad + \pi_1(1-\pi_1)[1-2\pi_1-2e^{\beta_1}(1-\pi_1)]\widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &\qquad + x_{i2}\pi_1(1-\pi_1)(1-2\pi_1)(1-e^{\beta_1})\widehat{cov}(\hat{\beta}_0, \hat{\beta}_2) \\ &\qquad + x_{i2}\pi_1(1-\pi_1)\big[1-2\pi_1-2e^{\beta_1}(1-\pi_1)\big]\widehat{cov}(\hat{\beta}_1, \hat{\beta}_2) \end{aligned}$$

Finally, collecting like terms, we may express the bias of the estimated relative difference in the probability of an individual having lung cancer as:

$$\begin{aligned} E(\hat{\eta}(x_{i2})) - \eta(x_{i2}) \approx{}& \pi_1(1-\pi_1)(1-2\pi_1)(1-e^{\beta_1})\left[\tfrac{1}{2}\left(\hat{V}(\hat{\beta}_0) + x_{i2}{}^2\hat{V}(\hat{\beta}_2)\right) + \right. \\ & \left. x_{i2}\widehat{cov}(\hat{\beta}_0, \hat{\beta}_2)\right] + \tfrac{1}{2}(1-\pi_1)\big[e^{\beta_1}(1-\pi_1) + \pi_1\big](1-2\pi_1)\hat{V}(\hat{\beta}_1) + \\ & \pi_1(1-\pi_1)[1-2\pi_1-2e^{\beta_1}(1-\pi_1)\big[\widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) + x_{i2}\widehat{cov}(\hat{\beta}_1, \hat{\beta}_2)\big] \end{aligned}$$

Similarly, we will use a first order Taylor series expansion of $\hat{\eta}$ about $\hat{\beta} = \beta$ to derive the variance of the estimated relative difference in the probability of having lung cancer given smoking status. Using equation (2.8) we have the following expression for this variance:

$$\begin{aligned} V(\hat{\eta}(x_{i2})) \approx{}& \left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_0) + \left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_1) + \left[\frac{\partial\hat{\eta}}{\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\right]^2 V(\hat{\beta}_2) + \\ & 2\left[\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}\right)\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\right)cov(\hat{\beta}_0, \hat{\beta}_1) + \left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_0}|_{\hat{\beta}=\beta}\right)\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\right)cov(\hat{\beta}_0, \hat{\beta}_2) + \right. \\ & \left.\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_1}|_{\hat{\beta}=\beta}\right)\left(\frac{\partial\hat{\eta}}{\partial\hat{\beta}_2}|_{\hat{\beta}=\beta}\right)cov(\hat{\beta}_1, \hat{\beta}_2)\right] \end{aligned}$$

where $\hat{V}(\hat{\beta}_j)$ and $\widehat{cov}(\hat{\beta}_j, \hat{\beta}_k)$, $j, k = 1,2,3$, are the estimates obtained from the inverse of the observed information matrix evaluated at $\hat{\beta}$. Recall that a first order expansion is valid due to the insignificance of the second order terms when assuming a large sample size. Once again, the partial derivatives of $\hat{\eta}$ in this expansion are well-defined and evaluating them yields:

$$
\begin{aligned}
V\big(\hat{\eta}(x_{i2})\big) \approx{} & \left[\pi_1(1-\pi_1)(1-e^{\beta_1})\right]^2 V\big(\hat{\beta}_0\big) + \left[e^{\beta_1}(1-\pi_1)^2 + \pi_1(1-\pi_1)\right]^2 V\big(\hat{\beta}_1\big) + \\
& \left[x_{i2}\pi_1(1-\pi_1)(1-e^{\beta_1})\right]^2 V\big(\hat{\beta}_2\big) + 2\Big[\pi_1(1-\pi_1)(1-e^{\beta_1})\Big(e^{\beta_1}(1-\pi_1)^2 + \\
& \pi_1(1-\pi_1)\Big) cov\big(\hat{\beta}_0, \hat{\beta}_1\big) + x_{i2}\pi_1{}^2(1-\pi_1)^2(1-e^{\beta_1})^2 cov\big(\hat{\beta}_0, \hat{\beta}_2\big) + \\
& \Big(e^{\beta_1}(1-\pi_1)^2 + \pi_1(1-\pi_1)\Big) x_{i2}\pi_1(1-\pi_1)(1-e^{\beta_1}) cov\big(\hat{\beta}_1, \hat{\beta}_2\big)\Big]
\end{aligned}
$$

Finally, collecting like terms, we may express the variance of the estimated relative difference in the probability of having lung cancer as:

$$
\begin{aligned}
V\big(\hat{\eta}(x_{i2})\big) \approx{} & \pi_1{}^2(1-\pi_1)^2\big(1-e^{\beta_1}\big)^2\Big[V\big(\hat{\beta}_0\big) + x_{i2}\Big(V\big(\hat{\beta}_2\big) + 2cov\big(\hat{\beta}_0, \hat{\beta}_2\big)\Big)\Big] + \\
& \left[e^{\beta_1}(1-\pi_1)^2 + \pi_1(1-\pi_1)\right]^2 V\big(\hat{\beta}_1\big) + 2\pi_1(1-\pi_1)\big(1-e^{\beta_1}\big)\Big(e^{\beta_1}(1-\pi_1)^2 + \\
& \pi_1(1-\pi_1)\Big)\left[cov\big(\hat{\beta}_0, \hat{\beta}_1\big) + x_{i2} cov\big(\hat{\beta}_1, \hat{\beta}_2\big)\right]
\end{aligned}
$$

Thus, we have the following 95% Wald confidence interval for the true relative difference between the proportion of individuals who have lung cancer given they are smokers and the proportion of individuals who have lung cancer given they are non-smokers:

$$
\hat{\eta}(x_{i2}) \pm 1.96\sqrt{Var(\hat{\eta}(x_{i2}))}
$$

where $z_{0.05/2} = 1.96$ is the $(0.025)th$ quantile of the standard normal distribution.

CHAPTER 3: Testing the Model

3.1 Generating Data

The final phase of this study on modelling a relative difference in proportions due to an effect of an explanatory variable is to test the model using simulations. Since the probabilities that we wish to compare are instances of a “success” for a Bernoulli distributed random response variable that can be modeled using multiple logistic regression, we can easily generate samples of the response rather than using observational data. In doing so, we can repeatedly estimate a relative difference in proportions while controlling the true values of the parameters for the model. This will allow us to analyze the effect that the choice of parameters has on the fit of the model in terms of the bias, variance, and accuracy of our estimates for the relative difference. We will use the lung cancer example that was developed in the previous chapters to carry out the simulations and analysis.

First, let's discuss the macro that was created to estimate the relative difference between the proportion of individuals who have lung cancer given they are smokers and the proportion of individuals who have lung cancer given they are non-smokers. For reference, the complete macro is given in Appendix II. To begin, the macro sets column names for the variables and quantities that will be generated including two explanatory variables, the indicator $x1$ and the continuous variable $x2$ representing smoking status and age respectively, as well as the response variable Y representing the presence or absence of lung cancer and EtaHat which holds the generated estimates for the relative difference. We will denote the number of simulations of the estimate $\hat{\eta}(x_{i2})$ by N and the number of observations per simulation by n. That is, n will denote the sample size of observed individuals who may or may not have lung cancer. After setting values for $\beta_0, \beta_1, \beta_2$, n, and the true age value x_{i2}, the macro proceeds to calculate the true relative difference in proportions of individuals with lung cancer using the given parameters and the expression $\eta(x_{i2})$ from equation (2.4). We then simulate n values for $x1$, half of which are 0 while the remaining $\frac{n}{2}$ are 1, as well as n values for $x2$ that are uniformly distributed between 30 and 70 which yields the observations $x_i' = (1, x1, x2), i = 1, \dots, n$. These observations remain the same throughout a single run of the macro, and so they are constant for every sample of the response variable that will be generated. To conclude the preliminary portion of the macro, we perform a calculation of the true probability of having lung cancer, $P(Y = 1)$, using the preset parameters and the multiple logistic regression model π_i. The iterative portion of the macro is essentially a nested loop which repeatedly generates a sample of n observations of the Bernoulli distributed response variable Y and fits the multiple logistic regression model π_i using the pairs $(x_i, Y_i), i = 1, \dots, n$, to obtain the maximum likelihood estimators $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$. Although we allow variation in the sample size n for different executions of the macro, the outer loop that dictates the estimates for η always iterates 200 times so that the large sample assumption is upheld. Therefore, we obtain 200 samples of the form $(x_i, Y_i), i = 1, \dots, n$, and 200 corresponding estimates for the relative difference in proportions of individuals with lung cancer given smoking status. When the nested loop terminates, the outer loop carries out procedures such as a goodness of fit test and regression on the parameters of the model to test their significance, however, we are not too concerned with these details. In addition, for each sample of size n we compute a single estimate for the relative difference between proportions of individuals with lung cancer given smoking status, EtaHat, using the model estimate $\hat{\eta}(x_{i2})$ given in equation (2.9), the values for the maximum likelihood estimates $\hat{\beta}$, and the true value for age x_{i2}. Furthermore, for each estimate we also calculate its variance, obtain a 95% confidence interval for the true relative difference in proportions, record whether the confidence interval covered the true relative difference η, and calculate the error of the estimate in terms of its distance from the true relative difference. Finally, when the outer loop terminates, we obtain values for the estimated bias of $\hat{\eta}(x_{i2})$ and the estimated coverage rate of the confidence intervals by taking the average of the 200 values for the error of each estimate and the average of the 200 values for the coverage of the confidence intervals, respectively. We also estimate the standard deviation of $\hat{\eta}(x_{i2})$ by taking the sample standard deviation of the estimates held in EtaHat, and lastly, calculate the mean of the variances of the estimates EtaHat and take the square root for comparison to the standard deviation estimate.

We are most interested in analyzing the four values that were estimated before termination of the macro – bias of $\hat{\eta}$, standard deviation of $\hat{\eta}$, average $\sqrt{Var(\hat{\eta})}$, and coverage rate – and how their values are influenced by varying the preset parameters $\beta_0, \beta_1, \beta_2, n$, and x_{i2}. More specifically, we want to investigate how changing the parameters of the true model for a relative difference in the proportions of individuals with lung cancer given smoking status will affect the accuracy of our estimate and the ability to construct a confidence interval that will cover the true relative difference. A summary of 35 executions of the macro using various values of $\beta_0, \beta_1, \beta_2, n$, and x_{i2} is given in the table below:

Sim. #	**N**	**n**	**β_0**	**β_1**	**β_2**	Age (x_{i2})	**η**	**Estimated Bias ($\hat{\eta}$)**	**Sample Std. Dev. of $\hat{\eta}$**	**Average of $\sqrt{\hat{V}(\hat{\eta})}$**	**Coverage rate**
1	200	200	-1	0.1	0.01	40	0.0655	0.0196	0.1962	0.20896	0.965
2	200	200	-1	0.1	0.01	60	0.0604	0.0150	0.1891	0.1895	0.945
3	200	40	-1	0.1	0.01	60	0.0604	0.0735	0.5319	0.6021	0.95
4	200	200	-1	0.5	0.01	40	0.3406	0.0043	0.2291	0.2481	0.945

5	200	200	-1	0.1	0.03	50	0.0373	0.0150	0.1145	0.1168	0.955
6	200	50	-1	0.1	0.03	50	0.0373	0.0274	0.2644	0.2595	0.93
7	200	200	-2	0.2	0.02	40	0.1619	0.0353	0.2925	0.2985	0.93
8	200	100	-2	0.2	0.01	60	0.1702	0.1377	0.6909	0.6842	0.935
9	200	200	-2	0.5	0.01	40	0.4867	0.0830	0.4385	0.4679	0.955
10	200	200	-2	0.5	0.01	60	0.4612	0.1018	0.4515	0.4681	0.97
11	200	80	-2	0.5	0.01	60	0.4612	0.1724	0.8579	0.9454	0.91
12	200	200	-2	1	0.02	50	0.8591	0.0777	0.3962	0.4038	0.97
13	200	200	-2	1	0.01	60	1.0287	0.1208	0.5983	0.6024	0.93
14	200	200	-2	1	0.05	60	0.2048	0.0023	0.0873	0.0862	0.93
15	200	60	-2	1	0.05	60	0.2048	0.0471	0.2308	0.2334	0.885
16	200	200	-3	0.5	0.01	40	0.5779	0.3037	1.0686	1.1936	0.905
17	200	200	-3	0.5	0.01	60	0.5643	0.3910	1.4038	1.6312	0.93
18	200	120	-3	0.5	0.01	60	0.5643	0.4657	1.3977	1.9438	0.915
19	200	200	-3	2	0.01	60	3.8251	0.8918	2.7599	3.0913	0.885
20	200	120	-3	2	0.01	60	3.8251	1.5285	3.9462	5.6212	0.95
21	200	200	-3	2	0.05	40	1.7183	0.1086	0.6185	0.6284	0.95
22	200	200	-3	2	0.05	60	0.7616	0.0311	0.2179	0.2129	0.95
23	200	200	-3	3	0.01	40	7.659	1.4672	5.970	7.0215	0.91
24	200	200	-3	3	0.02	50	5.1329	0.5946	2.1976	2.3071	0.95
25	200	200	-3	3	0.05	50	1.4478	0.0119	0.3791	0.3653	0.91
26	200	120	-3	3	0.05	50	1.4478	0.1981	0.6334	0.5773	0.96
27	200	200	-3	4	0.01	40	10.6026	3.6808	8.567	9.5922	0.955
28	200	200	-3	4	0.02	50	6.389	0.9542	2.8627	2.95295	0.93
29	200	120	-3	4	0.02	70	4.4579	1.5460	6.01138	6.6394	0.915
30	200	100	-3	4	0.02	70	4.4579	1.1524	4.6547	6.5748	0.90
31	200	200	-3	5	0.01	40	12.2607	4.2553	16.0715	20.0371	0.94
32	200	200	-4	0.1	0.05	50	0.0844	0.0769	0.3445	0.3701	0.94
33	200	60	-4	0.1	0.05	50	0.0844	0.2147	0.9283	1.0291	0.915
34	200	200	-4	2	0.05	50	2.4121	0.2018	0.9309	0.9469	0.94
35	200	80	-4	2	0.05	50	2.4121	0.4059	1.8258	1.9908	0.93

The first column, titled simulation number, simply orders the executions of the macro that were performed. Columns two and three hold the number of replications of the estimate $\hat{\eta}(x_{i2})$, denoted N, and the sample size of individuals used to obtain each estimate $\hat{\eta}$, denoted n. The remaining columns hold the choice of parameters, the true relative difference in proportions η, and the quantities that were estimated in each run of the macro. Thus, the columns that will be most interesting are the last four columns which hold the estimated values. As mentioned in Section 2.2, we are aiming for our estimate of the relative difference in proportions to have a small bias as this would suggest that we can expect the estimate to be close to the true relative difference η. In addition, taking the average of the root of the variance of each estimate as shown in the second column from the right should approximate the standard error of $\hat{\eta}$. Thus, we can also compare this column to the values for the sample standard deviation of $\hat{\eta}$ and expect them to be very similar. Most importantly, since we constructed 95% confidence intervals using each estimate for the relative difference in proportions, we can expect the coverage rate of the confidence intervals throughout each execution of the macro to be approximately 95%. This is the measure that will provide the most insight on the performance of the estimate of a relative difference in proportions. That is, when the desired coverage rate is not obtained it is likely due to the choice of parameters, such as the sample size n, or is a result of unsatisfactory values for the bias or variance of the estimate. Therefore, when combined, these three properties will underline the effectiveness of the estimate for a relative difference between proportions of individuals with lung cancer.

3.2 Analysis

Looking at the table containing simulations that were executed using the macro given in Appendix II, we can assess the patterns that arise when altering the parameters in the model for a relative difference in proportions, η, by observing the estimates that the model produces as well as the resulting bias, variance, and coverage obtained from those estimates. When choosing parameters to test, I aimed to choose values for $\beta_0, \beta_1,$ and β_2 that would yield reasonable probabilities of an individual having lung cancer given their smoking status based on the multiple logistic regression model π_i. I found it most reasonable to maintain the value for β_0 to be relatively small and negative in effort to counterbalance two positive values for β_1 and β_2. Otherwise, the probability of having lung cancer for smokers, π_1, and the probability of having lung cancer for non-smokers, π_0, tend to be unrealistically large. At a glance, the pattern for choosing which parameters to test will be quite evident.

As previously mentioned, the most intuitive estimate that we obtained from the macro is the coverage rate of the confidence intervals in each simulation. This is because the coverage rate is a measure of how well the relative difference in proportions can be estimated. That is, if we obtain a coverage rate of 95% then this implies that we successfully constructed a confidence interval that contained the true relative difference in proportions using 95% of the estimates made for η. However, since we obtain this estimate of the coverage rate through Monte Carlo simulation, successive runs of the same simulation will exhibit sampling variability and so, the estimate is subject to some degree of error. In fact, this error is known as the Monte Carlo error. If we let p denote a target quantity of interest and p_n denote the Monte Carlo estimate of p from a simulation with n replications, then the Monte Carlo error is defined as the standard deviation of p_n given by:

$$\mathrm{MCE}\left(p_n\right) = \sqrt{Var(p_n)}$$

Indeed, the target coverage rate taken across all repetitions of the estimated relative difference in proportions is $p = 95\%$ which implies a Monte Carlo error of:

$$\sqrt{\frac{p(1-p)}{N}}$$

Furthermore, since our confidence intervals depend on quatiles of the standard normal distribution, we may increase this error by a factor of 3 since:

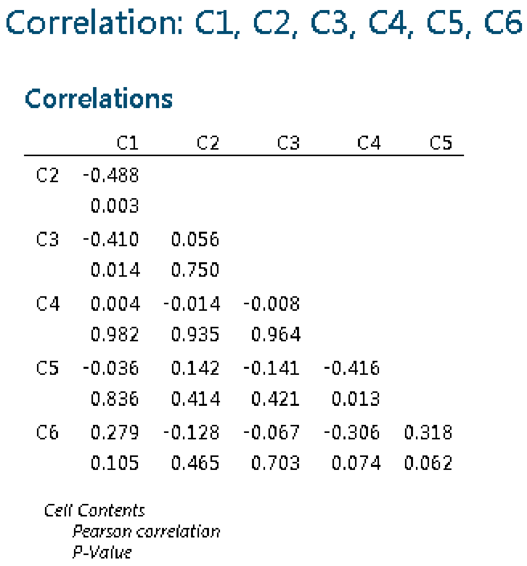
$$P(-3 \leq Z \leq 3) \approx 1$$

where Z follows a standard normal distribution. Therefore, the Monte Carlo error for the simulations which estimate the coverage rate becomes:

$$3\sqrt{\frac{p(1-p)}{N}} = 3\sqrt{\frac{0.95(0.05)}{200}} \approx 0.046$$

Thus, we will begin by analyzing the simulations which yield a coverage rate that is not within the range 0.95 ± 0.046.

Simulations number 15 and 30 both yield a coverage rate that is below what can be credited to Monte Carlo error. Naturally, there may be some other aspect of the simulation that did not permit the estimates to construct appropriate confidence intervals for the true relative difference in proportions of individuals with lung cancer. If we observe the simulations that immediately precede these, specifically simulations 14 and 29, the parameters chosen for these simulations only differ from 15 and 30, respectively, in terms of the sample size n. In fact, the sample size is increased in both simulations 14 and 29 and the resulting coverage rates of these simulations are within an acceptable distance from 0.95. Considering that all other simulations had reasonable coverage and the bias and variance of simulations 15 and 30 appear to be regular, their downfall is most likely the small sample size of individuals that was used to fit the model. Intuitively, a small sample size will decrease the accuracy of the estimates produced. Since these estimates are used to construct the confidence intervals which the coverage rate depends on, an inaccurate estimate will likely construct a confidence interval that is shifted away from the true relative difference in proportions and will not cover η, consequently skewing the coverage rate. To see this, we may find the correlation between the coverage rate and the parameters that we control in the macro, $\beta_0, \beta_1, \beta_2$, n, and x_{i2}, using the values in columns 3-7 and column 12 from the table of simulations. Doing so yields the following chart:



Correlation: C1, C2, C3, C4, C5, C6

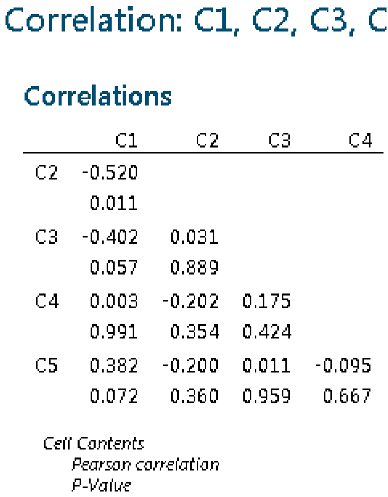
Correlations

	C1	C2	C3	C4	C5
C2	-0.488				
	0.003				
C3	-0.410	0.056			
	0.014	0.750			
C4	0.004	-0.014	-0.008		
	0.982	0.935	0.964		
C5	-0.036	0.142	-0.141	-0.416	
	0.836	0.414	0.421	0.013	
C6	0.279	-0.128	-0.067	-0.306	0.318
	0.105	0.465	0.703	0.074	0.062

Cell Contents
Pearson correlation
P-Value

where C1, C2, C3, C4, C5, and C6 represent $\beta_0, \beta_1, \beta_2, x_{i2}, n$, and coverage rate, respectively. The final row of the chart provides the p-value to test the significance of the correlation between coverage rate and the other five parameters. Although none of the parameters have a significant correlation with respect to the coverage rate at any commonly used significance level α, the most noteworthy p-value for a correlation with coverage rate is 0.062, obtained from C5 which represents n. Therefore, we can conclude that as the sample size of individuals increases, the coverage rate is likely to improve as seen in the simulations 15 to 14 and 30 to 29. This is also the case for many of the other simulations that were performed twice and only differed in the value for n such as simulations 10 and 11 as well as simulations 32 and 33.

It is also interesting to note that as the value for β_0 grows further from zero in the negative direction, keeping all other parameters constant, the probability of having lung cancer given an individual is a non-smoker, π_0, decreases. Therefore, the numerator of η will increase while the denominator of η decreases and so, as β_0 grows further from zero in the negative direction, the relative difference in proportions, η, increases. This can be seen through the simulations 4, 9, and 16, where β_1, β_2, x_{i2}, and n remain constant while β_0 strays away from 0. Another result of the decreasing behaviour of π_0 as β_0 becomes smaller is that it becomes difficult to fit the multiple logistic regression model for non-smokers when the sample size of individuals, n, is small. That is, when there are only a few observations in the sample being fit to the model π_i and the probability of having lung cancer for non-smokers is very small, it becomes unlikely to generate an observation where the individual is a non-smoker and has lung cancer. In fact, all non-smoking individuals in the sample may not have lung cancer and this lack of variation in the observations will not allow the multiple logistic regression model to be fit. Indeed, this occurred may times when running simulations where the value for β_0 was relatively far from zero and the sample size n was small. All observed non-smokers did not have lung cancer and the coefficients of π_i could not be estimated, causing the macro to terminate. If we look at any of the small sample simulations in the table which had a value $\beta_0 \leq -2$, the corresponding sample size n is the smallest sample size for which the macro would fully execute. For instance, simulations 18 and 20 were both run using a sample size of 50, 60, 80, and 100, none of which were executed without an error in fitting the regression model. This influence of β_0 is an indication that this coefficient has some significance to the model. Indeed, when omitting all simulations with a sample size $n < 200$ and reevaluating the correlation between coverage rate and the parameters $\beta_0, \beta_1, \beta_2$, and x_{i2} we have the following chart:



Correlation: C1, C2, C3, C4, C5

Correlations

	C1	C2	C3	C4
C2	-0.520			
	0.011			
C3	-0.402	0.031		
	0.057	0.889		
C4	0.003	-0.202	0.175	
	0.991	0.354	0.424	
C5	0.382	-0.200	0.011	-0.095
	0.072	0.360	0.959	0.667

Cell Contents
Pearson correlation
P-Value

where C1, C2, C3, C4, and C5 represent $\beta_0, \beta_1, \beta_2, x_{i2}$, and coverage rate, respectively. Again, the final row of this chart provides the p-value to test the significance of the correlation between coverage rate and the other four parameters. Although none of the parameters have a significant correlation with respect to coverage rate at any commonly used significance level α, the p-value for the correlation between C1 and C5, representing β_0 and the coverage rate, is 0.072 which is much smaller than that of any other parameter and indicates that the greatest effect on coverage rate is from β_0. Therefore, given a larger number of simulations with a constant, large sample of observations, n, it could be possible to estimate the coverage rate of the confidence intervals for a relative difference in proportions of individuals with lung cancer solely using the parameter β_0. That is, assuming a constant sample size, generating additional large sample simulations could improve the significance of the p-value for the correlation between β_0 and the coverage rate which would suggest performing a simple linear regression with the response variable as the coverage rate and a single explanatory variable β_0. This would allow us to predict the coverage of the confidence intervals using a single parameter.

In summary, it appears that the model for the relative difference between the probability of having lung cancer given an individual is a smoker and the probability of having lung cancer given an individual is a non-smoker performs as expected. When the simulations obey the large sample requirements of Monte Carlo simulation and the central limit theorem, the bias of the estimates produced is relatively small, the two estimates for the standard deviation of $\hat{\eta}$ are similar, and the desired coverage rate of the confidence intervals is achieved. However, the macro is limited in its estimation of the relative difference in proportions since we cannot expect as efficient of a performance when using a small sample of individuals. A small sample size inhibits the ability to obtain estimates for η since the lack of variation with certain parameters does not allow the multiple logistic regression model to be fit and so, the parameter estimates for η cannot be found. In addition, when the parameters are successfully estimated for a small sample size it is more likely that the resulting confidence interval does not cover the true relative difference in proportions. Thus, taking a large sample will yield the desired results as anticipated. Furthermore, having a large and constant sample size provides the possibility of estimating the mean coverage of the confidence intervals using far less parameters than required to fit a multiple logistic regression model.

Conclusion

The goal of this project was to illustrate how to model and estimate a relative difference in proportions using a multiple logistic regression model. In doing so, we explored simple logistic regression and saw the benefits of using such a model to represent the relationship between a dichotomous response variable and an appropriate explanatory variable. The introductory chapter also demonstrated how to fit the simple logistic regression model to find the maximum likelihood estimates required to estimate the probability of a success for the response variable. Furthermore, the attractiveness of the simple logistic regression model transferred to the multivariate case in which we used parallel techniques to model and estimate the probability of a success for a dichotomous response variable when p independent explanatory variables are considered. This multiple logistic regression model was the building block for the model of a relative difference in proportions. By representing the factor which may contribute to a difference in proportions as an explanatory variable in the multiple logistic regression model, π_i, we were able to build a model for a relative difference in proportions that depends on the unknown parameters of π_i. Thus, estimating the model for a relative difference in proportions becomes arbitrary once it is known how to fit the multiple logistic regression model and find the maximum likelihood estimates for its unknown parameters. The complexity of the expression for our estimate posed a challenge when finding its bias and constructing appropriate confidence intervals. However, linearizing the estimate for a relative difference in proportions using the results of the delta method made finding the expected value and variance of the estimate quite simple, yielding the desired expressions for bias and a 95% Wald confidence interval.

Throughout the project, the ideas that were developed were demonstrated by an example to estimate the relative difference between proportions of individuals with lung cancer given they are smokers and the proportion of individuals with lung cancer given they are non-smokers. The corresponding model, estimate, bias, and confidence interval for this example were obtained using the methods discussed in each chapter. The results were used to perform Minitab simulations to test the limits of the model. By generating data for hypothetical individuals that may or may not have lung cancer given their age and smoking status, the techniques developed throughout the project were used to find estimates for a relative difference in having lung cancer due to smoking. As anticipated, the model behaves well for large sample sizes, yielding estimates with a small bias and approximately a 95% coverage rate. However, smaller sample sizes proved to be far less efficient and less informative. Nevertheless, the goal to estimate a relative difference in proportions was accomplished and the ideas can be applied to many situations in which the effect of a set of independent explanatory variables on the presence of a response variable is of interest.

Appendix I

To evaluate the Taylor series expansion of $\hat{\eta}(x_{i2})$, we must compute the following partial derivatives:

$$\frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0}, \frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_1}, \frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_2}, \frac{\partial^2\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0{}^2}, \frac{\partial^2\hat{\eta}(x_{i2})}{\partial\hat{\beta}_1{}^2}, \frac{\partial^2\hat{\eta}(x_{i2})}{\partial\hat{\beta}_2{}^2}, \frac{\partial^2\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0\partial\hat{\beta}_1}, \frac{\partial^2\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0\partial\hat{\beta}_2}, \frac{\partial^2\hat{\eta}(x_{i2})}{\partial\hat{\beta}_1\partial\hat{\beta}_2}$$

First, let $f(\hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{i2}$. The following expressions will be useful for computation:

1) $\hat{\eta}(x_{i2}) = \frac{e^{\hat{\beta}_1} + e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}} - 1 = \frac{e^{\hat{\beta}_1}}{1+e^{f(\hat{\beta})}} + \frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}} - 1$

2) $\frac{e^{\hat{\beta}_1}}{1+e^{f(\hat{\beta})}} = e^{\hat{\beta}_1}\left(1+e^{f(\hat{\beta})}\right)^{-1}$

$$\Rightarrow \partial\left(\frac{e^{\hat{\beta}_1}}{1+e^{f(\hat{\beta})}}\right)\frac{1}{\partial\hat{\beta}_0} = -e^{\hat{\beta}_1}(1+e^{f(\hat{\beta})})^{-2}e^{f(\hat{\beta})}$$

3) $1-\pi_1(x_{i2}) = 1 - \frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}} = \frac{1+e^{f(\hat{\beta})} - e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}} = \frac{1}{1+e^{f(\hat{\beta})}}$

$$\Rightarrow \pi_1(x_{i2})[1-\pi_1(x_{i2})] = \frac{e^{f(\hat{\beta})}}{[1+e^{f(\hat{\beta})}]^2}$$

Returning to the partial derivatives:

$$\begin{aligned}
\frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0} &= \frac{\partial}{\partial\hat{\beta}_0}\left(\frac{e^{\hat{\beta}_1}}{1+e^{f(\hat{\beta})}} + \frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}} - 1\right) \\
&= \frac{e^{f(\hat{\beta})}\left(1+e^{f(\hat{\beta})}\right) - e^{f(\hat{\beta})}e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2} - \frac{e^{\hat{\beta}_1}e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2} \\
&= \frac{e^{f(\hat{\beta})}\left[\left(1+e^{f(\hat{\beta})}\right) - e^{f(\hat{\beta})}\right]}{\left[1+e^{f(\hat{\beta})}\right]^2} - \frac{e^{\hat{\beta}_1}e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2} \\
&= \left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2} \\
&= \pi_1(1-\pi_1)(1-e^{\hat{\beta}_1})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_1} &= \frac{\partial}{\partial\hat{\beta}_1}\left(\frac{e^{\hat{\beta}_1}+e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}} - 1\right) \\
&= \frac{(e^{\hat{\beta}_1}+e^{f(\hat{\beta})})\left(1+e^{f(\hat{\beta})}\right) - (e^{\hat{\beta}_1}+e^{f(\hat{\beta})})e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2} \\
&= \frac{e^{\hat{\beta}_1}+e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2} \\
&= e^{\hat{\beta}_1}(1-\pi_1)^2 + \pi_1(1-\pi_1)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \hat{\eta}(x_{i2})}{\partial \hat{\beta}_2} &= \frac{\partial}{\partial \hat{\beta}_2}\left(\frac{e^{\hat{\beta}_1}+e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}-1\right)\\
&= \frac{x_{i2}e^{f(\hat{\beta})}\left(1+e^{f(\hat{\beta})}\right)-\left(e^{\hat{\beta}_1}+e^{f(\hat{\beta})}\right)x_{i2}e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\\
&= x_{i2}\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\\
&= x_{i2}\pi_1(1-\pi_1)(1-e^{\hat{\beta}_1})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \hat{\eta}(x_{i2})}{\partial \hat{\beta}_0{}^2} &= \frac{\partial}{\partial \hat{\beta}_0}\left(\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\right)\\
&= \left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}\left[1+e^{f(\hat{\beta})}\right]^2-2e^{f(\hat{\beta})}\left[1+e^{f(\hat{\beta})}\right]e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^4}\\
&= \left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left(1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}\right)\\
&= \pi_1(1-\pi_1)(1-2\pi_1)(1-e^{\hat{\beta}_1})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \hat{\eta}(x_{i2})}{\partial \hat{\beta}_1{}^2} &= \frac{\partial}{\partial \hat{\beta}_1}\left(\frac{e^{\hat{\beta}_1}+e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\right)\\
&= \frac{\left(e^{\hat{\beta}_1}+e^{f(\hat{\beta})}\right)\left[1+e^{f(\hat{\beta})}\right]^2-2\left(e^{\hat{\beta}_1}+e^{f(\hat{\beta})}\right)\left(1+e^{f(\hat{\beta})}\right)e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^4}\\
&= \frac{e^{\hat{\beta}_1}+e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left(1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}\right)\\
&= (1-\pi_1)[e^{\hat{\beta}_1}(1-\pi_1)+\pi_1]\,(1-2\pi_1)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \hat{\eta}(x_{i2})}{\partial \hat{\beta}_2{}^2} &= \frac{\partial}{\partial \hat{\beta}_2}\left(x_{i2}\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\right)\\
&= x_{i2}\left(1-e^{\hat{\beta}_1}\right)\frac{x_{i2}e^{f(\hat{\beta})}\left[1+e^{f(\hat{\beta})}\right]^2-2e^{f(\hat{\beta})}(1+e^{f(\hat{\beta})})x_{i2}e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^4}\\
&= x_{i2}{}^2\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left(1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}\right)\\
&= x_{i2}{}^2\pi_1(1-\pi_1)(1-2\pi_1)(1-e^{\hat{\beta}_1})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0\partial\hat{\beta}_1} &= \frac{\partial}{\partial\hat{\beta}_0}\left(\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}+\frac{e^{\hat{\beta}_1}}{\left[1+e^{f(\hat{\beta})}\right]^2}\right) \\
&= \frac{e^{f(\hat{\beta})}\left[1+e^{f(\hat{\beta})}\right]^2-2e^{f(\hat{\beta})}\left(1+e^{f(\hat{\beta})}\right)e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^4}-2\frac{e^{\hat{\beta}_1}e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^3} \\
&= \frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left(1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}-2\frac{e^{\hat{\beta}_1}}{1+e^{f(\hat{\beta})}}\right) \\
&= \pi_1(1-\pi_1)[1-2\pi_1-2e^{\hat{\beta}_1}(1-\pi_1)]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_0\partial\hat{\beta}_2} &= \frac{\partial}{\partial\hat{\beta}_0}\left(x_{i2}\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\right) \\
&= x_{i2}\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}\left[1+e^{f(\hat{\beta})}\right]^2-2e^{f(\hat{\beta})}\left(1+e^{f(\hat{\beta})}\right)e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^4} \\
&= x_{i2}\left(1-e^{\hat{\beta}_1}\right)\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left(1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}\right) \\
&= x_{i2}\pi_1(1-\pi_1)(1-2\pi_1)\left(1-e^{\hat{\beta}_1}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial\hat{\eta}(x_{i2})}{\partial\hat{\beta}_1\partial\hat{\beta}_2} &= \frac{\partial}{\partial\hat{\beta}_1}\left(x_{i2}\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left(1-e^{\hat{\beta}_1}\right)\right) \\
&= x_{i2}\frac{e^{f(\hat{\beta})}\left[1+e^{f(\hat{\beta})}\right]^2-2e^{f(\hat{\beta})}\left(1+e^{f(\hat{\beta})}\right)e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^4}\left(1-e^{\hat{\beta}_1}\right)-\frac{e^{f(\hat{\beta})}e^{\hat{\beta}_1}}{\left[1+e^{f(\hat{\beta})}\right]^2} \\
&= x_{i2}\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left[\left(1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}\right)\left(1-e^{\hat{\beta}_1}\right)-e^{\hat{\beta}_1}\right] \\
&= x_{i2}\frac{e^{f(\hat{\beta})}}{\left[1+e^{f(\hat{\beta})}\right]^2}\left[1-2\frac{e^{f(\hat{\beta})}}{1+e^{f(\hat{\beta})}}-2\frac{e^{f(\hat{\beta})}e^{\hat{\beta}_1}}{1+e^{f(\hat{\beta})}}-2e^{\hat{\beta}_1}\right] \\
&= x_{i2}\pi_1(1-\pi_1)\left[1-2\pi_1-2e^{\hat{\beta}_1}(1-\pi_1)\right]
\end{aligned}
$$

Evaluating each derivative at $\hat{\beta} = \beta$ will yield the desired values for the second order Taylor series expansion.

Appendix II

The macro coded for Minitab to simulate the probability of having lung cancer for smokers and non-smokers and to estimate the relative difference between these probabilities is given below:

```
gmacro
honours
name c1 "x1"
name c2 "x2"
name c4 "Y"
name c5 "EtaHat"
name c6 "Var(EtaHat)"
name c7 "Lower"
name c8 "Upper"
name c9 "Cover"
name c10 "Error"
### k1 is the value of Beta_0
let k1=-3.0
let k2=0.5
let k3=0.01
let k4=100
let k7=2*k4
let k8=40
let k9=(exp(k2)-1)/(1+exp(k1+k2+k3*k8))
Set c1
1( 0 : 1 / 1 )k4
End.
Random k7 c2;
Uniform 30 70.
Let c2 = ROUND(c2,0)
let c3=exp(k1+k2*c1+k3*c2)/(1+exp(k1+k2*c1+k3*c2))
do k20=1:200
do k6=1:k7
let k5=c3(k6)
random 1 c50;
bernoulli k5.
let c4(k6)=c50(1)
enddo
erase c50
name C11 "COEF" M1 "XPWX".
Gzlm;
Nodefault;
REvent 1;
Response C4;
```

```
Continuous C1 C2;
Terms C1 C2;
Constant;
Binomial;
Logit;
TOdds;
Increment 1 1;
Unstandardized;
Tmethod;
Trinfo;
Tdeviance;
Tsummary;
Tcoefficients;
Tequation;
Tgoodness;
TDiagnostics 0;
Coefficients 'COEF';
Xpwxinverse 'XPWX'.
copy m1 c12-c14
let k10=exp(c11(2))-1
let k11=exp(c11(1)+c11(2)+c11(3)*k8)
let k12=1+k11
let k14=k11/k12
let c5(k20)=k10/k12
let k15=-k10*k14*(1-k14)
let k16=(exp(c11(2))*(1-k14)**2)+k14*(1-k14)
let k17=-k8*k10*k14*(1-k14)
let k18=k15*k15*c12(1)+k16*k16*c13(2)+k17*k17*c14(3)
let k19=2*(k15*k16*c12(2)+k15*k17*c12(3)+k16*k17*c13(3))
let c6(k20)=k18+k19
let c7(k20)=c5(k20)-1.96*sqrt(c6(k20))
let c8(k20)=c5(k20)+1.96*sqrt(c6(k20))
let c9(k20)=(k9 ge c7(k20) and k9 le c8(k20))
let c10(k20)=c5(k20)-k9
erase c11-c14
enddo
let k21=stde(c5)
let k22=mean(c6)
let k23=sqrt(k22)
let k24=mean(c10)
let k25=mean(c9)
print k9 k24 k21 k23 k25
endmacro
```

References

Agresti, Alan. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Farrell, P. STAT 5602: Logistic Regression 1 [Word document].

Farrell, P. STAT 5602: Logistic Regression 3 [Word document].

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Koehler, E. Brown, E., & Haneuse, S.J.-P.A. (2009). On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *The American Statistician, 63*(2), 155-156. Retrieved from https://www-jstor-org.proxy.library.carleton.ca/stable/25652244?pq-origsite=summon&seq=1#metadata_info_tab_contents

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). Available from http://people.math.carleton.ca/~awoodsid/69.354/Kutnertextbook.pdf

Newcombe, R. G. (2013). *Confidence Intervals for Proportions and Related Measures of Effect Size.* Available from https://ebookcentral-proquest-com.proxy.library.carleton.ca/lib/oculcarleton-ebooks/reader.action?docID=988751#