Predicting Insurance Status with Logistic Regression

Alan Wong

Intended Graduation: Fall 2022

Contents

1 Abstract

In this paper, we are going to explore the classification method of logistic regression and how it is applied and analyzed. We are going to use a data set containing multiple attributes to perform an exploratory data analysis. We will be mainly focusing on education, wage, and health insurance status. The goal of this capstone is to learn about logistic regression and multiple logistic regression methods as a tool to use for classification and the mathematics behind it. Then we will apply it to our data set to predict the outcomes and compare it to the original results of the data set. In our predictions, we will be looking to predict whether an individual has health insurance or not based on wage and education status separately.

2 Introduction

Predictive analysis and classification is something that we see in our day to day life. Whenever we apply for a loan or a credit card we are asked to provide many pieces of information in order to assess our eligibility. Some pieces of information have bigger impact on eligibility than others. The importance of this is that it allows us to collect valuable information to make business decision or improve machine learning and artificial intelligent. There are many methods of classification that each have a different purposes and use case. One method is known as Logistic regression. Logistic Regression is an important method used for classification because it is able to compute multiple variables and classify them into one of two outputs. To show how logistic regression is used, we first start with the data set we are applying it to and the technology needed.

2.1 Python Implementation

Python[9] will be our main tool in doing our analysis which includes the creation of our graphs, data set manipulation, and logistic regression implementation. The python library seaborn[1] and matplotlib[3] will be used to create our graphs and figures with labels. Pandas[5] is another library we will be using to import and manipulate our data set to where we want, so we can apply logistic regression. Lastly we will be using statsmodels[2] to create our models for logistic regression.

3 Exploratory Data Analysis

In this section, we are going to describe our data set and find relationships between the data variables using graphs, tables and chi squared analysis.

3.1 Data Set Description

The title of the data set is Mid-Atlantic Wage Data[6]. The data was collected on March 2011 by Steve Miller through a population survey. This is a data set that contains a collection of male workers in the mid-Atlantic region between the years 2003- 2009.

The sample size is 3000. There are four binary variables: gender, job class, health status, and health insurance status. There are three additional categorical variables: martial status, race, and education. The two quantitative variables are age and wage. The data set variable and attributes are shown in Table 1.

Variable Name	Role	Type	Level of Measure-ments	Units	Range of Values
Age	Independent	Quantitative	Ratio	Year	18-100
Martial Status	Independent	Categorical	Nominal	N/A	Married, Never Married, Divorced,Separated, Widowed
Race	Independent	Categorical	Nominal	N/A	White,Black,Asian,Other
Education	Independent	Categorical	Ordinal	N/A	HSGrad,CollegeGrad, Some College, Advanced Degree, Not HS Grad
Job Class	Independent	Categorical	Nominal	N/A	Industrial, Information
Health Status	Independent	Categorical	Nominal	N/A	Good, Very Good
Wage	Independent	Quantitative	Ratio	Thousands of dollars	0-400
Health Insurance	Dependent	Categorical	Nominal	N/A	Yes / No

Table 1: Data Set Description

For education, we are looking at the level of education that the individual has reached. In the variable job class, industrial is defined as a job that produces goods through a manufacturing process. Information is defined as an individual who works behind a computer, doing desk and administrative work, or could be known as a white collar worker. In the variable health status, Very Good means that the individuals is completely healthy, in physical sense, free of diseases, and health issues. Good Health means that they have some health issue so that they cannot be classified under the category of ”Very Good”. This means that they are not completely healthy, free of diseases, and health issues. If the individual has some sort of health issue regardless of what the issue is they will be grouped as Good Health status. Our target variable will be Insurance Status. A target variable a dependent variable that we want to predict using other variables in the data set.

Before we start the classification of our data set, we must look into our data set and understand the patterns and qualities of the data set in an exploratory data analysis. For the analysis, in this section we are going to put our focus in analyzing wage, education, and insurance status. We chose these varaibles because we assume that individuals with higher wage and higher education are more likely to have health insurance. In the next section, we will begin the analysis of each variables.

3.2 Univariate Analysis

In this section, we going to describe each individual variables using various tables and graphs to understand how they are represented in the data set.

As shown in Table 2 we see that the mean age is 42.4 and mean wage is $111,700 .

Label	Wage($1000)
Mean	111.7
Standard Deviation	41.7
Minimum	20.1
25% Quartile	85.4
50% Quartile	104.9
75% Quartile	128.7
Maximum	318.3

Table 2: Summary of Quantitative Variables

In Figure 1, we see that the wage distribution is right skewed and that the majority of the individuals earn between the $50,000 to $150,000. In Figure 1, we see that there are many outliers from wages $200,000 and greater. There could be many outliers since the histogram is right skewed.

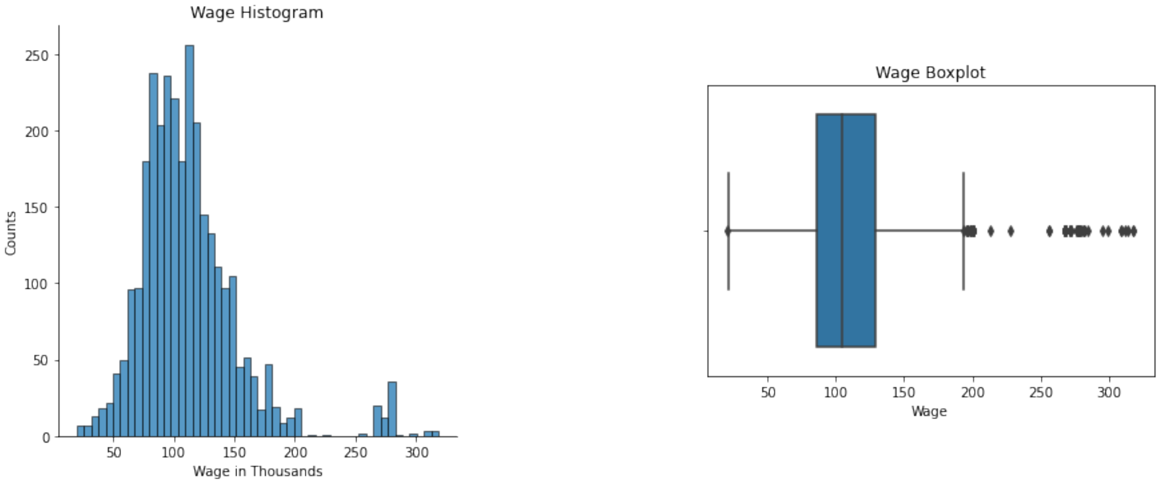

Figure 1: Wage Histogram and Box Plot

In Table 2, we see that most of the individuals had at least graduated from high school. About 59 percent of the individuals have attended college and 23 percent of the individuals graduated with a college degree and 14 percent graduated with an advance degree. Only 9 percent of the individuals did not graduate from high school.

Label:	Count	Total	Proportion
Not HS Grad	268	3000	0.09
HS Grad	971	3000	0.32
Some College	650	3000	0.22
College Grad	685	3000	0.23
Advanced Degree	426	3000	0.14

Figure 2: Education Level

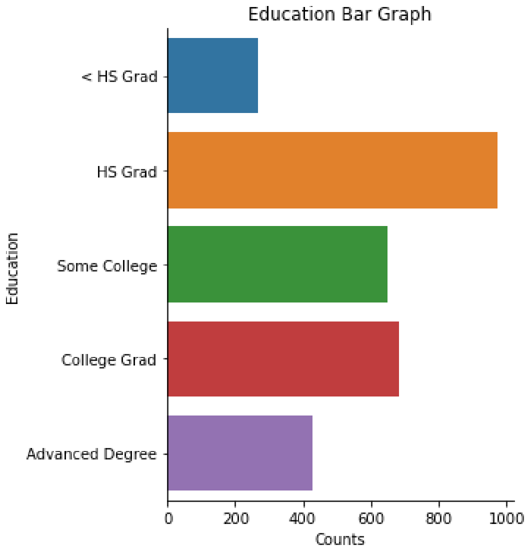

Figure 3: Education Bar Graph

In Table 3, the majority of the individuals are in the target variable have health insurance at 69 percent. The remaining 31 percent do not have health insurance.

Label:	Count	Total	Proportion
Insurance	2083	3000	0.69
No Insurance	917	3000	0.31

Table 3: Health Insurance Summary

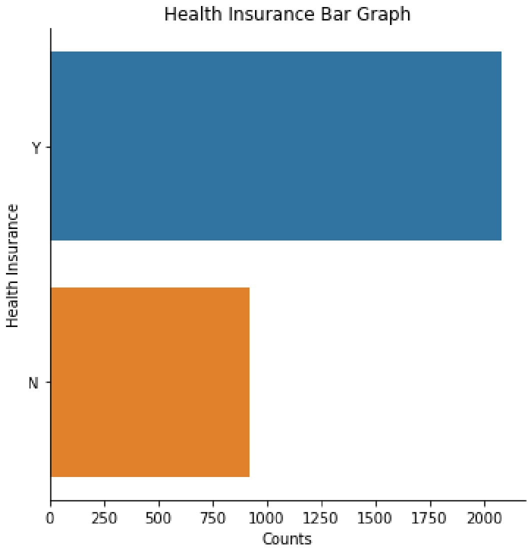

Figure 4: Health Insurance Bar Graph

3.3 Independent vs Dependent Bivariate Analysis

For our bivariate analysis we are going to compare each variable with each other to see if there is a relation between the variables. First, we compared our quantitative variables with the target variable, insurance status, using box plots.

In Figure 5, the average wage of the individuals that have health insurance is greater than the wage of individuals that do not have health insurance. From the box plot it seems the more money someone makes, the more likely they are going to have health insurance. This could be true because jobs that pay more offer better insurance benefits packages.

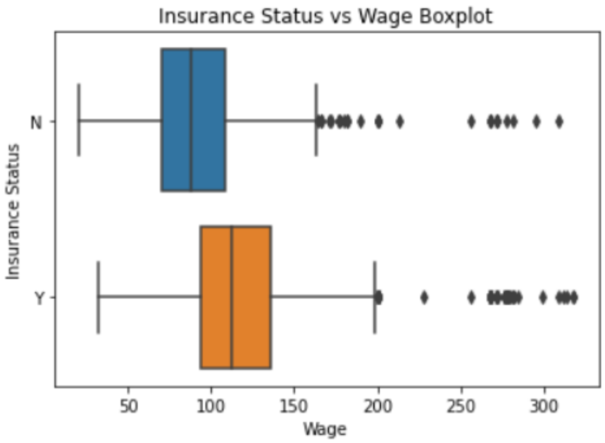

Figure 5: Wage vs Health Insurance Box plot

To compare education with health insurance status, we used a two way table to understand the frequency of the event occurring. We will also do a chi square test of independence [7] between our independent variable and our dependent variable to see if there is a relationship between the two variable. Our null hypothesis will be that there is no association between the two variables and our alternate hypothesis is that there is an association. Our significant value with be $\alpha = 0.05$.

$$H_0 : \text{The Education and Wage are Independent of Health Insurance}$$
$$H_a : \text{The Education and Wage are not Independent of Health Insurance}$$
$$\alpha < .05$$

In Table 4, we see that the majority of high school graduates have health insurance. Most individuals who did not graduate high school do not have health insurance. The chi-square statistic is 141.6346. The p-value is < 0.00001. The result is significant since $p < 0.05$. This means that there is enough evidence to conclude that there is an association between education level and health insurance status.

	No Insurance	Insurance	Total
Not HS Grad	144 (0.54)	124 (0.46)	268
HS Grad	359 (0.37)	612 (0.63)	971
Some College	183 (0.28)	467 (0.72)	650
College Grad	156 (0.23)	529 (0.77)	685
Advance Degree	75 (0.18)	351 (0.82)	426
Total	917 (0.31)	2083 (0.69)	3000

Table 4: Health Insurance vs Education Level

3.4 Independent Categorical vs Independent Quantitative Analysis

Now we are going to compare the independent categorical variables to independent quantitative variable using a box plot. We will only be focusing on education and wage because it will be used later in our testing. All other analyses will be found in the appendix.

In Figure 6, we created multiple box plots that shows the age of the individuals based on the level of education that ones achieves and the same for wage. From the box plots in Figure 6, we see that higher educations earn more in wage and also the interquartile range of each box plots are increasing in size. This means that the range of income between the 25th percentile and the 75th percentile is increasing.

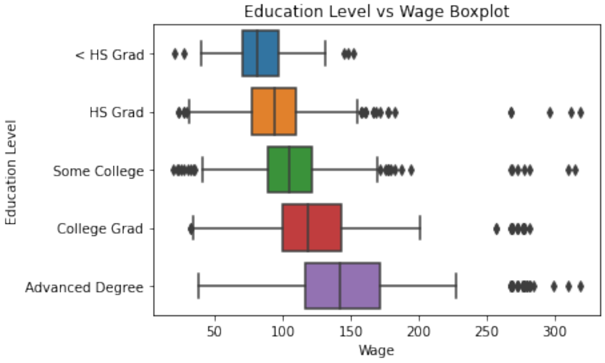

Figure 6: Education Level vs Wage Box Plot

Now that we understand our data set more, lets get started on learning more about classification and the method we will be using.

4 Logistic Regression for Classification

In this section, we are going to talk about what classification is, how it is used, and one method we want to explore. The method we are going to talk about is logistic regression and how it works. We will dive into the implementation of logistic regression, the equations, and provide some examples. Before talking about logistic regression, we first describe classification.

4.1 What is Classification?

Classification is the process of analyzing data with a data set and predicting a categorical output of the data set. The data set contains a list of variables,$(x_1, x_2, x_3, ..., x_n)$ and one output(y_1)[4, Page 128]. The x variables are known as the independent variables or features. The collection of all the features are grouped together as X. The Y variable will be the dependent variable or class. The X variables are used to understand the pattern of the data set and pattern of Y in relation to X.

Sometimes when working with classification, the data set is broken up into two sets, one for testing and one for training. A training data set is used to help learn about the data set and patterns. The training data set contains both X and Y variables. A testing data set is used to practice the method that was learned from the training data set. The testing data set only has the X variables and is trying to predict the Y variables by using the method that was learned from the training data set.When building the classifier, the classifier should not only work on the training data set, but it should also work for other data sets with the same features and class.

4.2 Logistic Regression Background

In simple linear regression, we know that the model form is $Y = \beta_0 + \beta_1 X$ where β_0 is the y-intercept and β_1 is the slope, Y is a dependent variable that can take on the values from $(-\infty, \infty)$ and X is a independent quantitative variable with values in $(-\infty, \infty)$. The parameters β_0 and β_1 are found using the least squares approach. While in logistic regression, the dependent variable, Y, is a binary with the value $\{0, 1\}$[4, Page 130], where $\{0, 1\}$ are not, numbers but are categorical classes and the value of X is in $(-\infty, \infty)$.

The problem with linear regression is that it does not capture most of the results between $Y = 1$ and $Y = 0$ well. Since logistic regression has an S - Shaped curve, and this captures our outputs are between $Y = \{0, 1\}$. This will be a a better way to model our data. To find out logistic regression equation, we start with Y in $\{0, 1\}$ and we want to create a new variable, so that our X can be in $(-\infty, \infty)$, such that the results looks like a linear regression. To get this we have to manipulate our linear regression equation in such a way to get this result.

4.3 Building the Logistic Regression Equation

First we start with conditional probability of $\text{Prob}(Y = 1|X)$[4, Page 131]. This means that when given a variable X, can we find what the probability of $Y = 1$. This is a good start because probabilities are only between 0 and 1. Using this conditional probability of $\text{Prob}(Y = 1|X)$, we can apply it to the odds ratio.

As defined in Definition 4.1, the odd ratio is the probability of an event occurring divided by the probability of the event not occurring. It is the odds of an event occurring to the event not occurring.

Definition 4.1 (Odds Ratio)**.**

$$\frac{\text{Prob}(Y = 1|X)}{\text{Prob}(Y = 0|X)} = \frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)} = r \tag{1}$$

In the odds ratio, we see that the numerator can only be in $(0, 1)$ and the denominator can only be in $(0, 1)$. We want to find a solution for $\text{Prob}(Y = 1|X)$ such that it is between $(0, 1)$ and in terms of r.

Lemma 4.1. *Let r be an arbitrary positive number. Then there exists a number u, strictly between 0 and 1, such that*

$$r = \frac{u}{1 - u}.$$

Proof. Let $u = \frac{r}{1+r}$. First we show u is less than 1 and greater than 0. Since r is positive, $r + 1 > 0$ so $u > 0$ and which is what we wanted. We want to show that

$$r = \frac{u}{1-u}.$$

Computing

$$\begin{aligned}
\frac{u}{1-u} &= \frac{\frac{r}{1+r}}{1-\frac{r}{1+r}} && \text{by substituting in } u \\
&= \frac{\frac{r}{1+r}}{1-\frac{r}{1+r}} \cdot \frac{(1+r)}{(1+r)} && \text{by multiplying by 1} \\
&= \frac{\left(\frac{r\cdot(1+r)}{1+r}\right)}{\left(1-\frac{r}{1+r}\right)\cdot(1+r)} \\
&= \frac{r}{(1+r)-r} \\
&= \frac{r}{1} \\
&= r,
\end{aligned}$$

Which is what we want to show. With the Lemma 4.3, we now know that the odds ratio is in $(0, \infty)$. □

Now we can take this information and apply it to the log odds ratio as shown in Definition 4.1. In this definition, Z is a function of Y.

Definition 4.2 (Log Odds Ratio)**.**

$$Z = \log\left(\frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)}\right) \tag{2}$$

The benefit of the Log Odds Ratio 4.2 function is that it is an S shape curve and the Prob$(Y = 1|X)$ is between $(0, 1)$. The Log Odds Ratio will capture most of our outputs since the range is $(0, 1)$. Also the Log Odds Ratio, or Z, can take on any value in $(-\infty, \infty)$ which is similar to the linear regression equation.

To find out our logistic regression equation, we are also going to substitute Z into our linear regression formula so that $Z = \beta_0 + \beta_1 X$. Now we can set our Log Odds Ratio equal to our linear regression formula so that we can get a model of the Log Odds Ratio in relation to the linear regression function

We start with both our log odds ratio function and linear regression function. We are going to use these two function to create our logistic regression equation by solving for our condition probability

$$\begin{aligned}
Z &= \log\left(\frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)}\right) \\
Z &= \beta_0 + \beta_1 X
\end{aligned}$$

First we set our two equations equal to each other, and begin solving for $\text{Prob}(Y = 1|X)$

$$\log \left(\frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)} \right) = \beta_0 + \beta_1 X.$$

To remove the natural logarithm, we exponentiate both sides to base e

$$\frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)} = e^{\beta_0 + \beta_1 X}.$$

Now by doing some algebra we manipulate the fraction on the left side of the equation to solve for our conditional probability. For simplicity, we are going to sub in $u = \text{Prob}(Y = 1|X)$

$$\begin{aligned}
\frac{u}{1-u} &= e^{\beta_0+\beta_1 X} \\
u &= (1-u)e^{\beta_0+\beta_1 X} \\
u &= e^{\beta_0+\beta_1 X} - u \cdot e^{\beta_0+\beta_1 X} \\
u + u \cdot e^{\beta_0+\beta_1 X} &= e^{\beta_0+\beta_1 X} \\
u(1 + e^{\beta_0+\beta_1 X}) &= e^{\beta_0+\beta_1 X} \\
u &= \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \\
\text{Prob}(Y = 1|X) &= \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}
\end{aligned}$$

Now this is the formula for our logistic regression function. In the next section, we are going to find out how to get the parameters for our logistic regression equation.

5 Parameter Estimation Likelihood

In linear regression, we would use a least squared method to find our parameters for the linear model but in logistic regression the method we use is maximum likelihood. First we are going to explain what least squared is and how it works in linear regression. Then we will compare least squared to maximum likelihood in linear regressions. Lastly we are going to explain how maximum likelihoods works for binary outcomes and justify our use of this method over least squared for logistic regression.

5.1 Ordinary Least Squared in Linear Regression

In least squared estimation in linear regression, the model is usually called the line of best fit as shown as $\hat{Y}$ in equation (3). $\hat{Y}$ is the predicted value based on some X. We are going to let (X_i, Y_i) be points from the data where $i \in \{1, 2, 3, .., n\}$ and (X, Y) are quantitative continuous variable. Y_i is the actual points where it accounts for the error, ϵ_i.

$$\hat{Y_i} = b_0 + b_1 X_i \tag{3}$$
$$Y_i = b_0 + b_1 X_i + \epsilon_i \tag{4}$$

We denote ϵ_i for $i \in \{1, 2, 3, ..., n\}$ where the error is the distance between the predicted value from the actual value in the data.

$$\begin{aligned}
Y_i &= b_0 + b_1 X_i + \epsilon_i \\
Y_i &= \hat{Y_i} + \epsilon_i \\
\epsilon_i &= Y_i - \hat{Y_i} \\
\epsilon_i &= Y_i - (b_0 - b_1 X_i) \qquad \text{by substituting in } \hat{Y} \text{ in equation (3)}
\end{aligned}$$

Now that we have an equation for the error between the actual data from our model, we want to find a way to calculate our parameters in our model, which is b_0, b_1. The first thing is we are going to introduce the Sum of Squared Errors.

Definition 5.1 (Sum of Squared Errors(SSE))**.**

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} (\epsilon_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y_i})^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2
\end{aligned}$$

The Sum of Squared Errors is the sum of all the errors between the model and the actual data points squared. Now we are going to take this equation and minimize this so that we have the smallest error possible. We are going to this by taking the partial derivative with respect to b_0 and b_1. Then we find the critical points of each respected equation that minimizes the equation by setting it equal to 0 and solving. We mention this because this is what we want to do to our maximum likelihood equation.

5.2 Maximum likelihood

Now that we know what least squared estimation is, let's talk about what maximum likelihood is. In maximum likelihood we want to find the values of b_0, b_1 that best fits the model with the data we have[10]. We want to take a different approach to this and the first thing we are going to do is assume that ϵ_i is independent, normally distributed, has a mean of 0 and standard deviation of σ^2. This will help us take the maximum likelihood equation and turn it into something that we know from least squared estimation.

$$
\begin{aligned}
\text{Prob}(Y_i|X_i, b_0, b_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[\frac{-(Y_i - (b_0 - b_1X_i))^2}{2\sigma^2}\right] \\
&= \text{Likelihood}(b_0, b_1) \\
&= L(b_0, b_1)
\end{aligned}
$$

Now that we have an equation for our parameters $L(b_0, b_1)$, we want to maximize these because we want the parameters that give us the highest likelihoods.

$$L(b_0, b_1) \quad = \textstyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[\frac{-(Y_i-(b_0-b_1X_i))^2}{2\sigma^2}\right] \tag{5}$$

Next, we want to introduce a theorem about maximization of logs to transform our$L(b_0, b_1)$ in equation 5

Theorem 5.1. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be differentiable. Suppose the global max exist at $\hat{w}$. Then $\hat{w}$ maximizes g if and only if $\hat{w}$ maximizes $\log g$.

Proof. Suppose that $\hat{w}$ maximizes g. Then $g(\hat{w}) \geq g(w) \forall w \in \mathbb{R}$. Since log is an increasing function this means, $\log(g(\hat{w})) \geq \log(g(w)) \forall w \in \mathbb{R}$.

Now Suppose that $\hat{w}$ maximizes $\log(g)$. Then $\log(g(\hat{w})) \geq \log(g(w))\ \forall w \in \mathbb{R}$. Then since e^x is an increasing function, then $e^{\log(g(\hat{w}))} \geq e^{\log(g(w))}\ \forall w \in \mathbb{R}$. This would then simplify to $g(\hat{w}) \geq g(w)$. Thus $g(\hat{w})$ maximizes g. □

Now that we understand how the theorem works, we can use this theorem to take the log of our likelihood equation and get it to a form that we recognize from our least squared method.

$$
\begin{aligned}
\log L(b_0, b_1) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[\frac{-(Y_i - (b_0 - b_1X_i))^2}{2\sigma^2}\right]\right) \\
&= \sum_{i=1}^{n} \log\frac{1}{\sqrt{2\pi\sigma}} + \log\left(\exp\left[\frac{-(Y_i - (b_0 - b_1X_i))^2}{2\sigma^2}\right]\right) \\
&= \sum_{i=1}^{n} \log\frac{1}{\sqrt{2\pi\sigma}} + \left[\frac{-(Y_i - (b_0 - b_1X_i))^2}{2\sigma^2}\right]
\end{aligned}
$$

In our equation, $\frac{1}{\sqrt{2\pi\sigma}}$ is just another constant. So when we take two derivatives all constants are removed and are left with,

$$\log L(b_0, b_1)' = \sum_{i=1}^{n} \left[\frac{-(Y_i - (b_0 - b_1X_i))^2}{2\sigma^2}\right]' \tag{6}$$

We have the same system of equation from linear approximation method. Thus MLE and OLS for $Y_i = b_0 + b_1X_i + \epsilon_i$ is equivalent assuming ϵ_i is independent and normally distributed.

The next thing we want to do to is to recall Bernoulli distribution to connect with Maximum likelihood

5.3 Recall Bernoulli

Now we are going to recall Bernoulli distribution because it is the likeliness of 1 of 2 events occurring. We are using this so that we can connect our independent, normally distributed likelihood to our log odds. We now have, y $Binomial(1, \alpha) = y$ $Bernoulli(\alpha)$ for $\alpha \in (0, 1)$, then the $\text{Prob}(y = b) = \alpha^b(1 - \alpha)^{1-\alpha}$ where $b = 0, 1$. We want to show the properties of this with lemma (5.2)

Lemma 5.2. Let $\theta = \frac{e^w}{1+e^w}, w \in \mathbb{R}$ then,

1. $\theta \in (0, 1)$
2. $1 - \theta = (1 + e^w)^{-1}$

Proof. 1. To show that $\theta \in (0, 1)$, we will show that $\theta > 0$ and $\theta < 1$. First looking at $\theta = \frac{e^w}{1+e^w}$ it is always positive because e is positive so $e^w > 0$ and $1 + e^w > 0$. Also for $\frac{e^w}{1+e^w}$, $e^w < 1 + e^w$ so $\frac{e^w}{1+e^w} < 1$.

2.To show that $1 - \theta = (1 + e^w)^{-1}$ we can substitute in $\frac{e^w}{1+e^w}$ for θ and we get,

$$\begin{aligned} 1 - \theta &= 1 - \frac{e^w}{1 + e^w} \\ &= \frac{1 + e^w}{1 + e^w} - \frac{e^w}{1 + e^w} \\ &= \frac{1 + e^w - e^w}{1 + e^w} \\ &= \frac{1}{1 + e^w} \\ &= (1 + e^w)^{-1}. \end{aligned}$$

Now we have shown that $\theta \in (0, 1)$ and $1 - \theta = (1 + e^w)^{-1}$. □

Now we recall that in our logistic equation $Z_i = \text{Log Odds}$. We can set up the equation and simplify.

$$\begin{aligned} Z_i &= \text{Log Odds} \\ &= \log\left(\frac{\text{Prob}(Y_i = 1|X_i)}{1 - \text{Prob}(Y_i = 1|X_i)}\right) = \beta_0 + \beta_1 X_i \end{aligned}$$

Now we can substitute $\theta(X_i) = \text{Prob}(Y_i = 1|X_i)$ to simplify to get,

$$= \log\left(\frac{\theta(X_i)}{1 - \theta(X_i)}\right) = \beta_0 + \beta_1 X_i. \tag{7}$$

Now we have connected our lemma with our logistic equation model and we want to use this to help connect it to our MLE parameter estimation.

5.4 Connecting MLE to Bernoulli

Using Y_i Bernoulli $(\theta(X_i)$, we are going to connect this to the MLE equation model which is 8. We are going to simplify our variables to make things easier to transfer into our next step.

$$\text{Prob}(Y_1, .., Y_n|X_i, b_0, b_1) = \prod_{i=1}^{n}(\theta(X_i)^{Y_i})(1-\theta(X_i)^{1-Y_i} \tag{8}$$

$$L(b_0, b_1) = \prod_{i=1}^{n}(\theta(X_i)^{Y_i})(1-\theta(X_i)^{1-Y_i} \tag{9}$$

$$L(\theta) = \prod_{i=1}^{n}(\theta(X_i)^{Y_i})(1-\theta(X_i)^{1-Y_i} \tag{10}$$

The next thing we want to do is maximize $L(\theta)$, but first we want to simplify into a form we know.

$$\begin{aligned}
\log L(\theta) &= \log \prod_{i=1}^{n}(\theta(X_i)^{Y_i})(1-\theta(X_i)^{1-Y_i} \\
&= \sum_{i=1}^{n} \log(\theta(X_i)^{Y_i})(1-\theta(X_i)^{1-Y_i} \\
&= \sum_{i=1}^{n} \log(\theta(X_i)^{Y_i}) + \log(1-\theta(X_i)^{1-Y_i} \\
&= \sum_{i=1}^{n}(Y_i)\log(\theta(X_i)) + (1-Y_i)\log(1-\theta(X_i)) \\
&= \sum_{i=1}^{n}(Y_i)\log(\theta(X_i)) + \log(1-\theta(X_i)) - (Y_i)\log(1-\theta(X_i)) \\
&= \sum_{i=1}^{n}(Y_i)\log(\theta(X_i)) - (Y_i)\log(1-\theta(X_i)) + \log(1-\theta(X_i)) \\
&= \sum_{i=1}^{n}(Y_i)\log\left(\frac{\theta(X_i)}{1-\theta(X_i)}\right) + \log(1-\theta(X_i))
\end{aligned}$$

After all the simplification, we recognize this form from Equation 7. Now lets plug in Log Odds Ratio and Lemma 5.2.

$$\begin{aligned}
\log L(b_0, b_1) &= \sum_{i=1}^{n}(Y_i)\log\left(\frac{\theta(X_i)}{1-\theta(X_i)}\right) + \log(1-\theta(X_i)) \\
&= \sum_{i=1}^{n}(Y_i)(\beta_0+\beta_1 X_i) + \log(1-e^{\beta_0+\beta_1 X_i}) \qquad \text{where } w = \beta_0 + \beta_1 X_i
\end{aligned}$$

After all simplification we now have an equation to find and estimate our parameters for our logistic equation and we are going to do this by maximizing our parameters. We are doing this because we want to find the two parameter values, (β_0, β_1), that gives the highest likelihood of an event occurring. The thing is, there is no way to solve this explicitly. There are other ways such as numerical optimization, Newtons-Raphson Methods [2]. For our purposes, we are going to use python to compute our parameters.

6 Applying Logistic Regression

Now that we know the method of logistic regression classification, we will apply it on our wage data set and compare the resulting prediction from logistic regression to our true results from the data set.

6.1 Implementation of Logistic Regression

For the implication of logistic regression our Y axis will be our target value, health insurance status, with the outcome of yes and no as 1 and 0 respectively. Our X axis will be our quantitative variables wage. Age will not be used for our logistic regression because it is not represented well in logistic regression, only quantitative X variables can be used for our classification.

As we can see from Figure 7, this logistic regression does not capture our wage data well. It is missing all of the wages of individuals who do not have health insurance. This is because the histogram is right skewed as shown in Figure 7. To resolve this issue we applied a log transformation to our wages to create a better spread of the data.

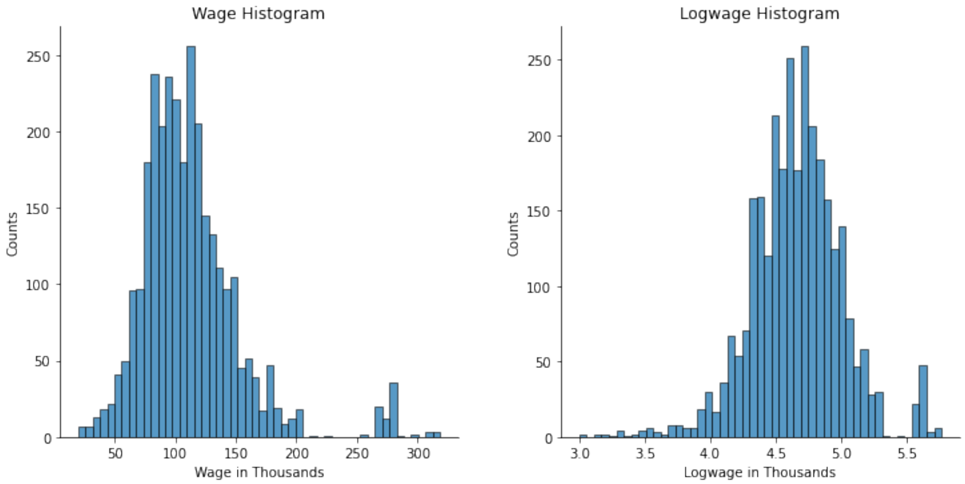

Figure 7: Wage vs Log Wage Logistic Regression

After the log transformation, our new histogram, Figure 7, the data looks more normally distributed than right skewed. Now in Figure 7, after the log transformation, we see that we have a logistic curve that will better represent the wage data and one we can use to predict out out comes.

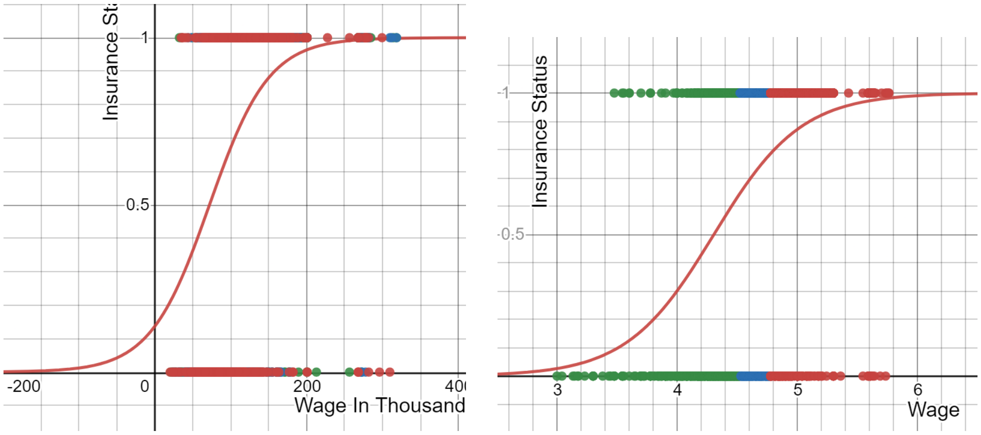

Table 5: Wage vs Log Wage Histogram

Now, we are ready to apply logistic regression to our data set. We will be using python [9] to create our logistic regression, do our classification, and produce our results.

6.2 Logistic Regression Wage Results

After running our data set through the package named "statsmodels" [8] in python the model we obtained is:

$$\text{Prob}(Y = 1|X) = \frac{e^{-11.873+2.758X}}{1 + e^{-11.873+2.758X}}. \tag{11}$$

This is going to be the model we will use to predict the health insurance status of our data set using wage.

The way we classified our results is if the probability is greater than 0.5 we will classify that as the person does have health insurance and if the probability is less than 0.5 we will classify that as the person does not have health insurance.

$$\text{Prob}(Y = 1|X) = \begin{cases} 1 \text{ or Yes} & \text{if Prob}(Y = 1|X) > 0.5 \\ 0 \text{ or No} & \text{if Prob}(Y = 0|X) < 0.5 \end{cases} \tag{12}$$

With our equation we can now begin to classify and predict our data. To show how we will classify for health insurance status we can take an example if the wage was 80, in thousands of dollars. We first take log of the wage. So we get $\log(80) = 4.382$ and we now take this new value and plug it into the logistic regression equation(6):

$$\text{Prob}(Y = 1|X) = \frac{e^{-11.873+2.758*4.382}}{1 + e^{-11.873+2.758*4.382}} = 0.552. \tag{13}$$

So the output from our example logistic equation(7) was 0.552 and from the way we are classifying this we see that the value is greater than 0.5, so we would classify this individual as the person does have health insurance.

To interpret our results we are going to use a two way table or could also be called a confusion matrix. A confusion matrix is used to compare our predicted or classified results to the actual

results of the data set. There are four possible results of the matrix. If the actual and the predicted results are the same we call it a positive result because we predicted it correctly. If the predicted and actual results are different we call it a false positive result because we predicted it incorrectly. We are also going to calculate accuracy of the results by dividing everything we predicted correctly by the over total amount of individuals in our data set, which is 3000.
As we can see from Table 6,from our positive results we predicted 2243 out of the 3000 individuals health insurance status correctly, where 293 are no and 1950 are yes, which is about 74.8% correct. From our false positive results we predicted a total of 757 responses incorrectly, where 133 responses should have been yes for health insurance but with our logistic regression equation it classified it as no health insurance and 624 responses should have been no but our logistic equation predicted it as yes.

		Predicted		
		No	Yes	Total
Actual	No	Positive: 293	False Positive: 624	917
	Yes	False Negative: 133	Positive: 1950	2083
	Total	426	2574	3000

Table 6: Wage Result Table

So we see that when classifying with wage as our independent variable in our logistic equation we are about 74.8% accurate to the original data set. We do not know if this is a good variable for classifying health insurance but a way to find out is we can use other variables and classify with those variables to see how accurate they are. In the next section, we are going to classify health insurance with some categorical variables and see how accurate are those variables and if they are better for our predictor.

7 Logistic Regression With Categorical Predictor

Now we are going to use independent categorical variables to predict the status of health insurance. The two variables we are going to use education. Now for classifying using categorical variables, there is one thing we have to do different to our logistic regression. We are going to need a reference variable. A reference variable is a variable within the categorical variable that is used to compare against all the other variables within the categorical variable chosen. Usually the reference variable is the one that appears the most within the data set. For our purposes, the reference variable for education will be HS Grad since they both occur with the highest frequency in the data set for their respected category.

7.1 Predicting Health Insurance Status With Education

So for education our reference variable is going to be HS grad. So knowing this and running our model in statsmodels in python we get our logistic regression equation to be:

$$\text{Prob}(Y = 1|X) = \frac{e^{0.533410-0.682942X_1+1.009888X_2+0.687723X_3+0.403433X_4}}{1 + e^{0.533410-0.682942X_1+1.009888X_2+0.687723X_3+0.403433X_4}} \tag{14}$$

First, we are going to relabel our education status variables, if the education status is anything else but our reference variable, HS Grad, we are going to label it as 1, and for any other variable it will be labeled as 0 as shown in Figure 8. Then from here we will classify the Prob$(Y = 1|X)$ the same as we did for wage.

$$X_1 = \begin{cases} 1 & \text{if Education is} < \text{HS Grad} \\ 0 & \text{if anything else} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if Education is Advanced Degree} \\ 0 & \text{if anything else} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if Education is College Grad} \\ 0 & \text{if anything else} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if Education is Some College} \\ 0 & \text{if anything else} \end{cases}$$

Figure 8: Labels for Education Status

After calculating all of the data from education we see that in our confusion matrix from Figure 7, we successfully predicted 144 individuals do not health insurance correctly and 1959 individual do have health insurance correctly. We also predicted 124 individuals that should have had health insurance but did not from our equation and 733 individuals that should not have had health insurance but did from our equation.

		Predicted		
		No	Yes	Total
Actual	No	Positive: 144	False Positive: 773	917
	Yes	False Negative: 124	Positive: 1959	2083
	Total	268	2732	3000

Table 7: Education Result Table

Overall our logistic regression for education was 70.1%. Now, we have seen how a categorical variable, education, works with logistic regression. In the next section, we are going to review the results from both variables.

7.2 Summary of Single Variable Predictor

Overall, wage was the most accurate of the two variables from Table 7.2 with our logistic regression equation at 74.8% accurate. This does make sense through intuition where your income does effect the odds of someone having health insurance or not. The more income someone has the more likely they are going to have health insurance and the other way around. There could be effects of race and education but we have seen through our predicting that wage is the most accurate.

Variable	Accuracy
Wage	74.8%
Education	70.1%

Table 8: Accuracy of Each Variable

So far we have been working with logistic regression with one variable. We now want to build a more complex model by combining multiple variables and create a multi variable logistic regression and predict with this.

8 Logistic Regression With Multiple Predictor

Now we want to build a more complex model that can capture more than one of our variables. In the next models we are going to combine our quantitative variable wage with our categorical models education to see if the models become more accurate.

First we are going to build a logistic regression equation by combining wage and education. When we run our model in python, we are going to keep the same reference variables from section 7.1 and we are going to add another variable wage to our model. Wage will act like a modifier from our education logistic regression model. So we are looking to see what the impact on our model would be if we have the education status and the wage will make. From our python output we find that our logistic regression equation coefficients and odds ratios are:

Variable Names	Coefficient	Odds Ratio	P-Value
Y - Intercept	-11.51	1.70	0.001
Wage	2.68	16.80	0.001
< HS Grad	-0.44	0.51	0.003
Advanced Degree	1.02	2.75	0.909
College Grad	1.14	1.99	0.295
Some College	1.14	1.49	0.254

Table 9: Multiple Predictor with Wage and Education Coefficient and Odds Ratio Results

From Table 9, we see that wage still has a huge impact on the insurance status when adjusted for education level, since the p-value is small at 0.001. The odd ratio for wage is 16.80. This means that someone who meets the wage requirement is almost 17 times more like to have health insurance than no health insurance. Also individuals who have a education level of < HS Grad are also significant when adjusted for wage at a p-value of 0.003. The odd ratio for this variable is 0.51. This means that if the education level for an individual was < HS Grad, they are half as likely to have health insurance than not having health insurance. This could also be interpreted as the individual is 2 times as likely to not have health insurance than having health insurance.

Using our multiple logistic regression model we are going to predict the insurance status of all of our individuals in our data set. We see in Table 10 we predicted 1981 with health insurance and 249 with no health insurance correctly. While we incorrectly predicted 102 individuals as no when they were actually yes and 668 as yes when they were actually no.

		Predicted		
		No	Yes	Total
Actual	No	Positive: 249	False Positive: 668	917
	Yes	False Negative: 102	Positive: 1981	2083
	Total	351	2649	3000

Table 10: Multiple Logistic Regression Result Table

Overall, the accuracy of our multiple logistic regression was 74.33% accurate. This is much better than education accuracy from Table 7.2 but not as good as wage but it is close. When modifying our education model and adding on wage to make a multiple logistic regression model that has made our education model better. This could mean that wage does make a big impact on health insurance status of an individual. With the p-value being very small and the odd ratio being relatively high at 17 times as likely, wage does make a bigger impact than education level.

9 Limitations

Throughout our testing we realized that their are many things that could have effected our results from our models. Some of the things that we considered to be a limitation is inflation of wage and the type of data set that we choose.

The first limitation we realize was that we were working with wage between different years. The problem is that there is inflation each year. Inflation is the value of things are increasing so it would decrease the purchasing power of money. So when looking at all of our wage values in our data set that was collected between different years we find out that the value of wage between years are different in purchasing power. The best way to avoid this issue is for us to pick a data set that is in the same year. This will rule out the inflation because all of the wage will be worth the same.

The next limitation that we notice was the data set we chose. We noticed that when testing our variables on our models, they were not as accurate as we thought. When looking into the history of our data set, we find that it was meant to be used to predict wages. The best way to get around this is to pick a data set that is meant for predicting binary outcomes. A few that we came up with was credit default or Titanic survival predictions.

References

[1] Michael Waskom ˙mwaskom/seaborn: v0.8.1 (september 2017). September 2017.

[2] Mircea Cirnu and Irina Badralexi. On Newton-Raphson Method. *Romanian Economic Business Review*, 5(1):91–94, May 2011.

[3] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[4] Garath James. *An introduction to statistical learning : with applications in R.* Springer texts in statistics ; 103. Springer, New York, NY, 2013.

[5] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[6] Steve Miller. Wage: Mid-atlantic wage data.

[7] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, July 1900.

[8] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[9] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA, 2009.

[10] Sanford Weisberg. *Applied linear regression.* Wiley series in probability and statistics. Wiley, Somerset, 4th ed edition, 2014.