2020

Statistical Analysis of Demographic Effects on Insurance Coverage of Perinatal and Neonatal Morbidity

Madeline Durbin
madeline.durbin@otterbein.edu

Statistical Analysis of Demographic Effects on Insurance Coverage of Perinatal and Neonatal Morbidity

By: Maddie Durbin
Department of Mathematical and Actuarial Sciences
Otterbein University
Westerville, OH 43081

April 7, 2020

Submitted in partial fulfillment of the requirements
For graduation with Honors

Zhen Huang, Ph.D. ________________________________
Project Advisor Advisor's signature

Pei Pei, Ph.D. ________________________________
Second Reader Second Reader's signature

Louise Captein, MFA ________________________________
Honors Representative Honors Representative's signature

Acknowledgements

Thank you to everyone who contributed to the completion of this thesis, including Dr. Zhen Huang, Ph.D., Dr. Pei Pei, Ph.D., Louise Captein, MFA, Jon Eshelman, FCAS, MAAA, Matt Michaels, ACAS, and Cristina Alvarez.

A special thank you to my awesome parents and fiancé for all their love and support. You rock!

Abstract

In the United States of America, Ohio has one of the worst neonatal and perinatal death rates. Within Ohio, Montgomery County has an above average neonatal and perinatal death rate. This statistic can be lowered if more women in Montgomery County have health insurance. They would be more likely to seek out prenatal health care, since they would no longer have to pay as much money out-of-pocket. This would allow medical professionals to be able to diagnose and treat any potential issues in the mother or child earlier. Having health insurance would also prevent mothers-to-be from seeking out other potentially dangerous options to avoid paying exorbitant amounts of money to deliver their baby, such as at-home births. This project seeks to identify whether women who fall into certain demographics have different likelihoods of having health insurance.

Data was collected from Miami Valley Hospital, located in Montgomery County. The data was then run through several different models, until one was chosen that was the most accurate and adequate. The selected model showed that women with various demographics do have different likelihoods of having health insurance. This would allow an insurance company to be able to design a product specifically for the demographics who are the most likely to be uninsured, thus increasing the number of women who have insurance, and lowering the neonatal and perinatal death rate.

Table of Contents

Page

List of Tables

Introduction

Neonatal and perinatal deaths in the United States are a serious problem. In 2017, there were 5.79 deaths per one thousand births. This means that the percent chance of death is 0.579%. For comparison, the seasonal flu has a mortality rate of only 0.1% (npr.org). In countries that are similarly wealthy as the United States in terms of GDP per capita, the average neonatal and perinatal mortality rate is only 3.4 deaths per one thousand births. Ohio in particular has one of the worst rates in the United States, at 7.2 deaths per one thousand births in 2017. This makes Ohio the eighth worst state in terms of the neonatal and perinatal death rate (healthsystemtracker.org). In the state of Ohio, Montgomery County experiences 7.5 deaths per one thousand births on average during 2012-2016 (phdmc.org). While this is a complicated issue that no single change could solve completely, if the number of people who had health insurance was greater, it would certainly help.

A study done by the National Center of Biotechnology Information found that "uninsured women receive fewer prenatal care services than their insured counterparts and report greater difficulty in obtaining the care that they believe they need" and "uninsured newborns are more likely to have adverse outcomes, including low birth weight and death, than are insured newborns"(ncbi.nlm.nih.gov). This supports the statement that if more women had health insurance, more women would seek prenatal healthcare, allowing medical experts to identify and treat potential issues with the mother or baby earlier.

The top ten causes of infant deaths[1] as noted by the CDC, can all, except for possibly SIDS, be minimized or prevented through care of medical experts. In addition to this, the average cost of having a vaginal delivery is $5,000 - $11,000 and the average cost of having a Cesarean birth is $7,500 - $14,500 (smartasset.com). If a new parent-to-be does not have health insurance, they will wind up paying out-of-pocket for this expense. This can lead to potentially dangerous at-home births and other perilous attempts to avoid paying so much money. A study by the American Journal of Obstetrics and Gynecology found that “planned home births were associated with significantly elevated neonatal mortality rates”. In fact, according to the American College of Obstetricians and Gynecologists, babies die in home births at roughly twice the rate as they do in hospital births. In addition, one complication, neonatal seizure, is three times more likely to occur at home (npr.org). However, “an average uncomplicated vaginal birth costs about 60% less in a home than in a hospital”, making it an appealing choice for an uninsured mother (americanpregnancy.org).

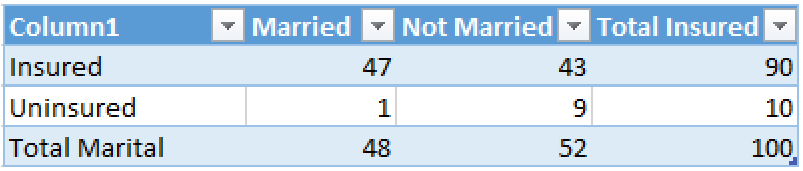
For all the reasons listed above, this project attempts to identify the demographic categories that uninsured women in Montgomery County fall into most, so that health insurance companies could market a product specifically to those demographics to help increase the number of women who have health insurance and therefore lower the neonatal and perinatal death rate.

Sample Data Variable Descriptions

This thesis attempts to identify the demographic categories that uninsured women in Montgomery County fall into most often by using data from Miami Valley Hospital, a Level 3 Trauma and a Level 3 NICU center that provides prenatal care to people from all over the county. The data consists of a sample of one hundred women from the survey that Miami Valley Hospital gives to new moms after they deliver their baby and those coming in to receive prenatal care. It asks the following: age, race, marital status, employment status, and whether they have health insurance. The dependent variable is insurance status, coded as 1 if they do not have health insurance and 0 if they do. There are four independent variables, three categorical (marital status, employment status, and race) and one continuous (age). Marital status has two levels, married and not married, coded M and NM respectively. Employment status has three levels, employed, self-employed, and unemployed, coded E, S, and U respectively. Race has five levels, Caucasian, African American, Asian, Hispanic, and Pacific Islander, coded C, AA, A, CHIS, and PI respectively. Tables 1-7 describe the dataset further.

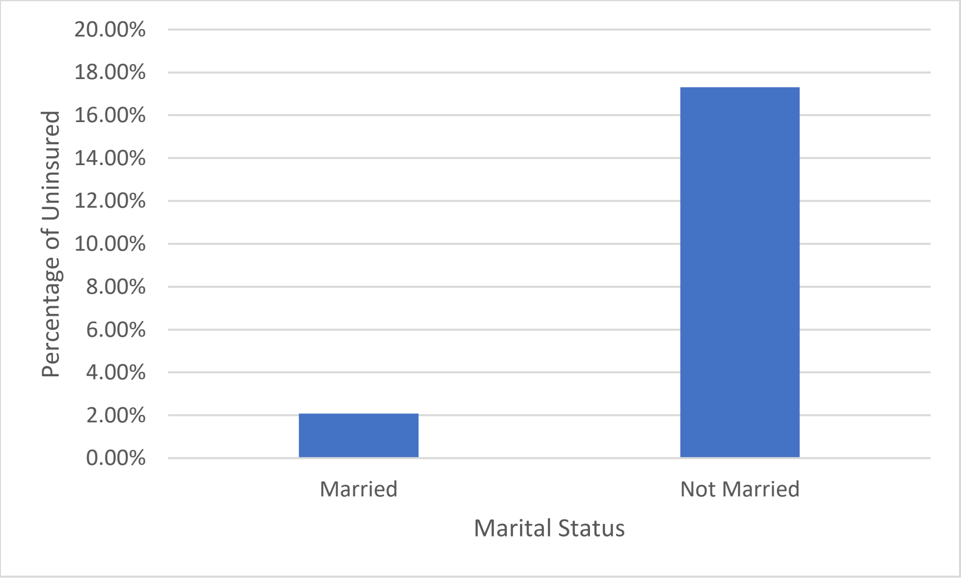
Descriptive Statistical Analysis

Table 1 – Marital Status



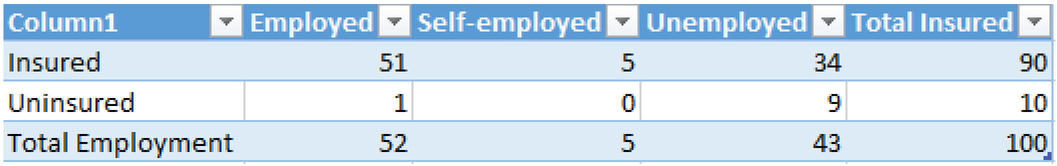
Column1	Married	Not Married	Total Insured
Insured	47	43	90
Uninsured	1	9	10
Total Marital	48	52	100

Table 2 – Histogram of Uninsured Percentage in Marital Status



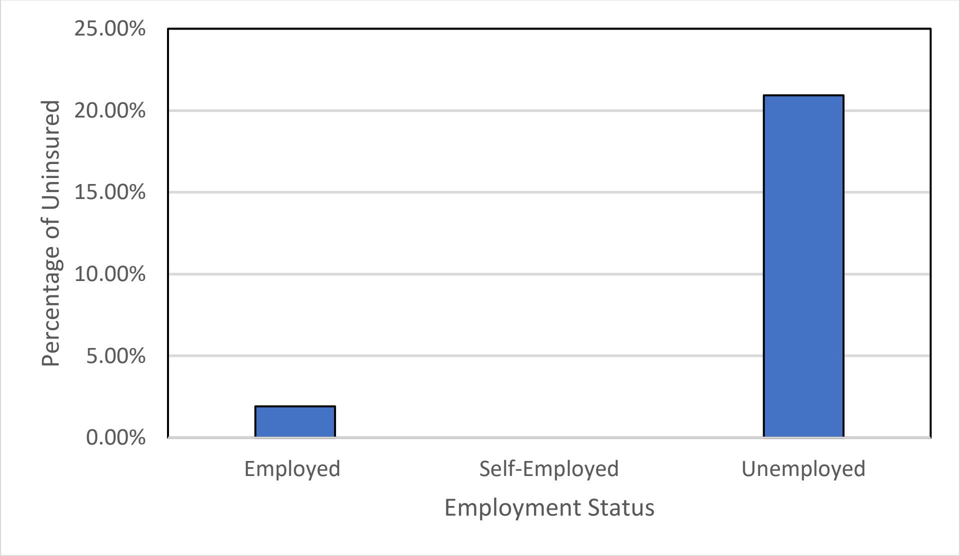
These tables show that while the insured category is nearly evenly distributed between married and not married women, there are far more not married women than married women in the uninsured category.

Table 3 – Employment Status



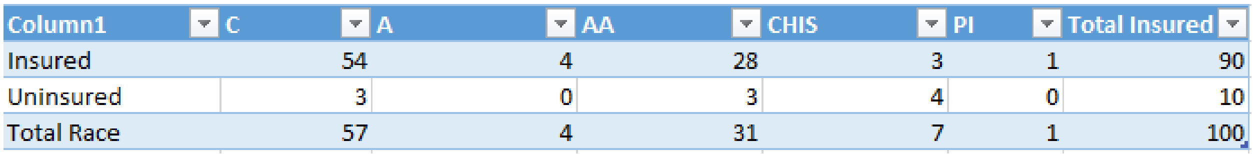
Column1	Employed	Self-employed	Unemployed	Total Insured
Insured	51	5	34	90
Uninsured	1	0	9	10
Total Employment	52	5	43	100

Table 4 – Histogram of Uninsured Percentage in Employment Status



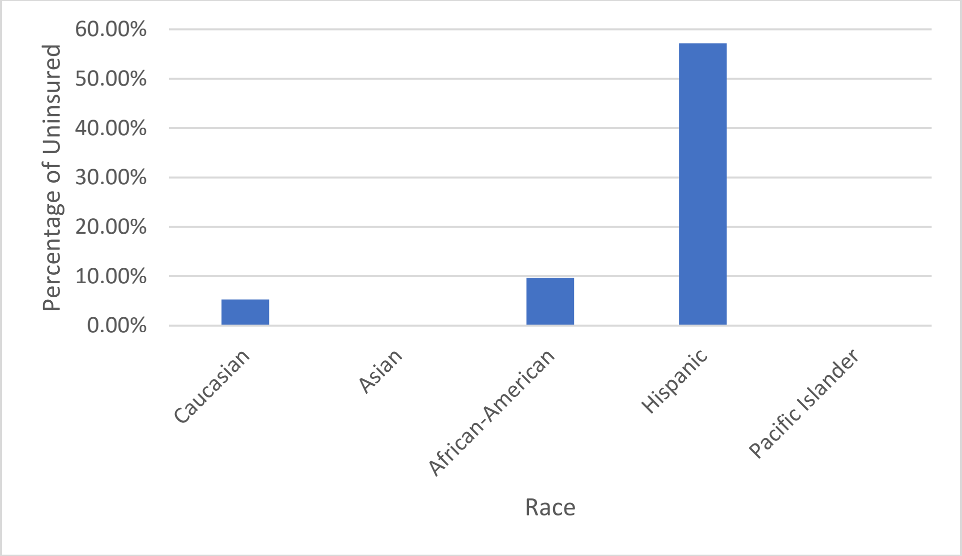
Notice that unemployment shows a similar skewness as marital status. The insured group has a substantial amount of both employed and unemployed women, but the uninsured group has quite a few more unemployed women than employed women.

Table 5 – Race



Column1	C	A	AA	CHIS	PI	Total Insured
Insured	54	4	28	3	1	90
Uninsured	3	0	3	4	0	10
Total Race	57	4	31	7	1	100

Table 6 – Histogram of Uninsured Percentage in Race



Notice that while there are only about half as many African-American women as Caucasian women, yet there are the same number of uninsured women of both groups, resulting in a higher percentage of uninsured African-American women. There is also a large percentage of Hispanic women who are uninsured, while no Asian or Pacific Islanders are uninsured.

Table 7 – Histogram of Ages

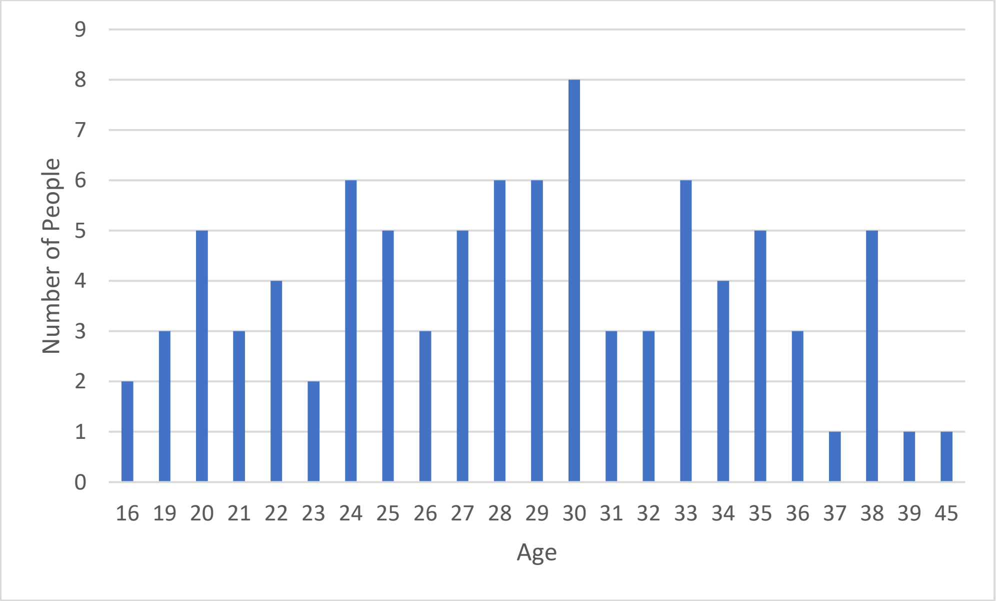
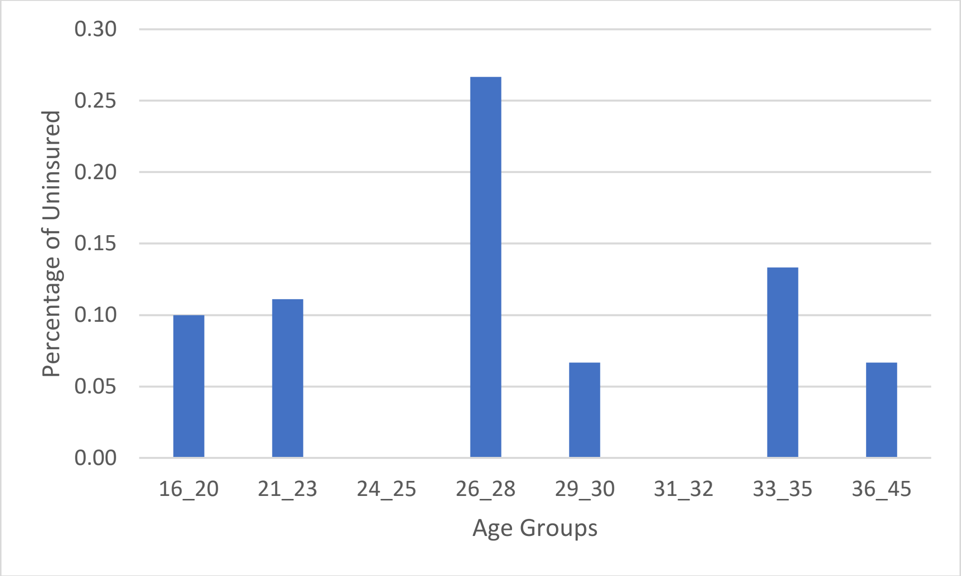


Table 8 – Histogram of Uninsured Percentage in Age



The age variable in this dataset ranges from a minimum of sixteen to a maximum of forty-five.   Age 30 has the most observations with eight, and ages 17, 18, 40, 41, 42, 43, and 44 have the least with zero. It is clear from Table 8 that the 26-28 age group has by far the greatest amount of uninsured women, while age groups 24-25 and 31-32 have none.

Inferential Statistical Analysis – Logistic Regression Model

An appropriate model for this dataset is binary logistic regression because it can address a binary dependent variable and is a common model amongst actuaries in the workplace.   A frequency table and two scatterplots were constructed between insurance status y and age x_1 as demonstrated in *Applied Logistic Regression*, pages 2-5, to identify if these two variables show any sort of recognizable trend. The variable age was sorted into eight approximately even categories in order to allow any trend in the generated scatterplot to be easier to see. Tables 9, 10, and 11 show that there does not seem to be any easily identifiable trend between insurance and age.

Table 9 – Scatterplot of Insurance Status and Age

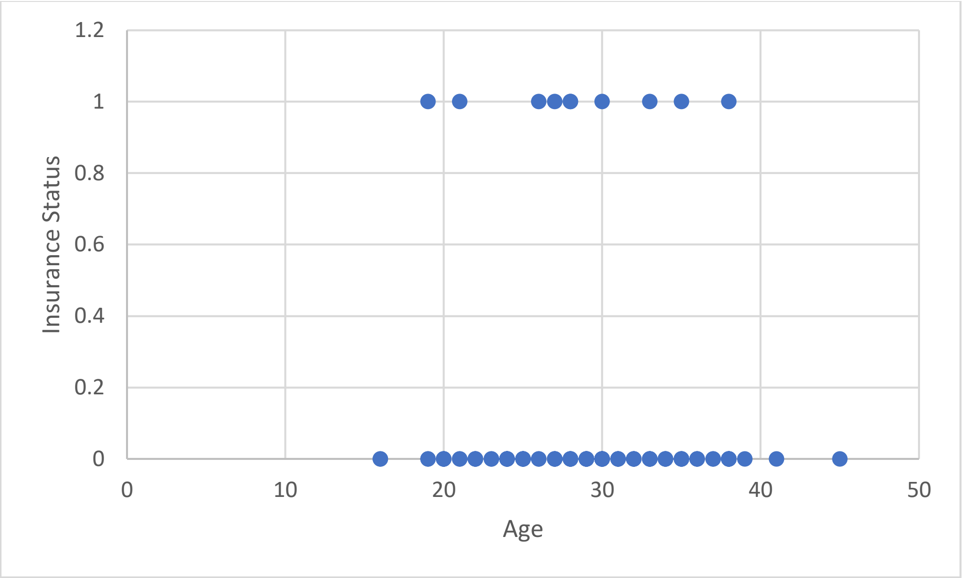
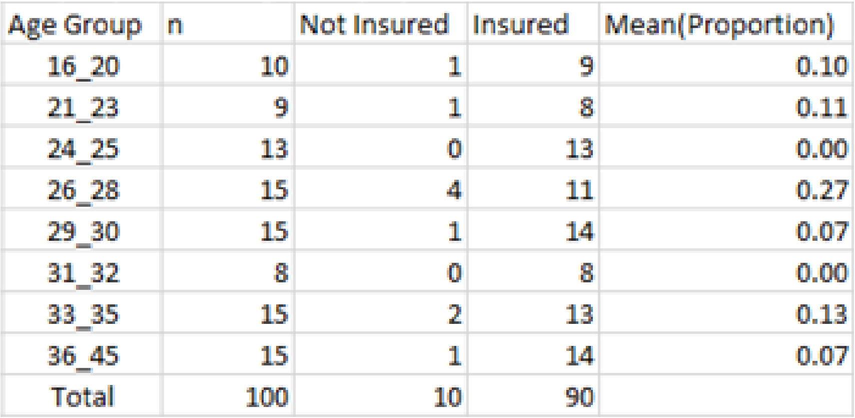
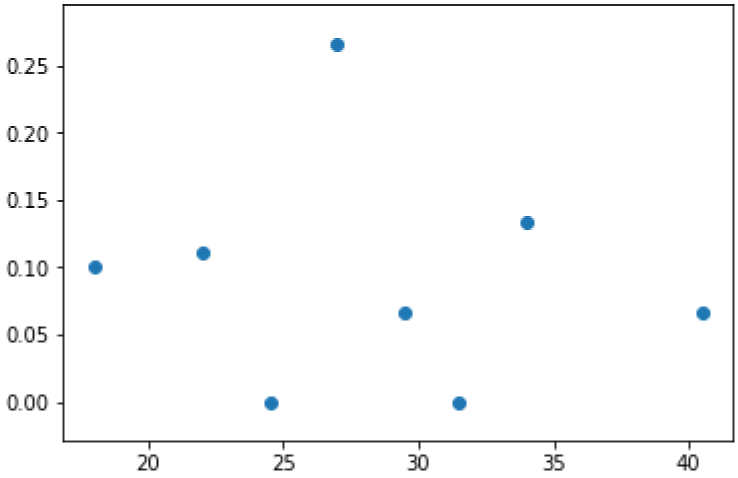


Table 10 – Frequency Table of Age Group by Insurance Status



Age Group	n	Not Insured	Insured	Mean(Proportion)
16_20	10	1	9	0.10
21_23	9	1	8	0.11
24_25	13	0	13	0.00
26_28	15	4	11	0.27
29_30	15	1	14	0.07
31_32	8	0	8	0.00
33_35	15	2	13	0.13
36_45	15	1	14	0.07
Total	100	10	90	

Table 11 – Plot of the percentage of subjects without insurance in each age group



The continuous variable age (by year) is defined as x_1. However, since this dataset also includes three categorical variables (marital status, employment status, and race), they cannot just be assigned a number to be properly defined. Instead, they have to be coded as dummy variables. This is because most machine learning models (such as logistic regression) cannot directly handle categorical variables, requiring the variable to be transformed in some way, such as creating dummy variables. To accomplish this transformation from categorical to dummy, base levels are selected and each remaining level in the variable is defined as either zero or one. A value of one for any given category signifies that the observation falls into that category and zero signifies that it falls into one of the other levels. There is no singular way to select the base level, but in this case, the normative or largest level was chosen as the base. For marital status, not married was chosen as the base level, so that

$$x_2 = \left\{ \begin{matrix} 1\ if\ subject\ is\ married \\ 0\ if\ the\ subject\ is\ not\ married \end{matrix} \right\}$$

Employed, defined as x_4, was chosen as the base level for employment status, so that

$$x_5 = \begin{Bmatrix} 1\ if\ the\ subject\ is\ self - employed \\ 0\ if\ not \end{Bmatrix}$$

$$x_6 = \begin{Bmatrix} 1\ if\ the\ subject\ is\ unemployed \\ 0\ if\ not \end{Bmatrix}$$

Note that if x_5 and x_6 are both zero, then the subject must be employed. Lastly, Caucasian, defined as x_9, was chosen as the base level for race, so

$$x_7 = \begin{Bmatrix} 1\ if\ the\ subject\ is\ Asian \\ 0\ if\ not \end{Bmatrix}$$

$$x_8 = \begin{Bmatrix} 1\ if\ the\ subject\ is\ African - American \\ 0\ if\ not \end{Bmatrix}$$

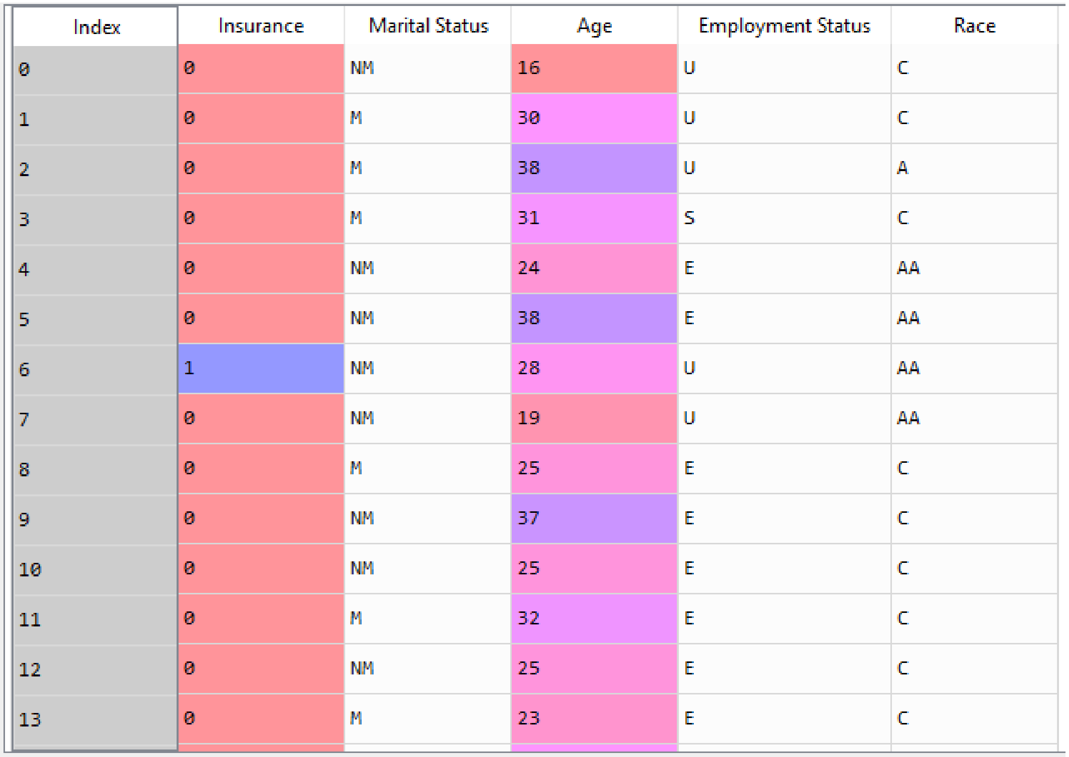
$$x_{10} = \begin{Bmatrix} 1\ if\ the\ subject\ is\ Hispanic \\ 0\ if\ not \end{Bmatrix}$$

$$x_{11} = \begin{Bmatrix} 1\ if\ the\ subject\ is\ a\ Pacific\ Islander \\ 0\ if\ not \end{Bmatrix}$$

Note that once again, if the subject is Caucasian, then x_7, x_8, x_{10}, and x_{11} must be zero. Since $x_{3,}$ x_4, and x_9 represent the chosen base levels, their coefficients are accounted for in β_0. There are several ways to incorporate the dummy variables into the data, one of which is to use one-hot encoding. This strategy "convert(s) each category value into a new column and

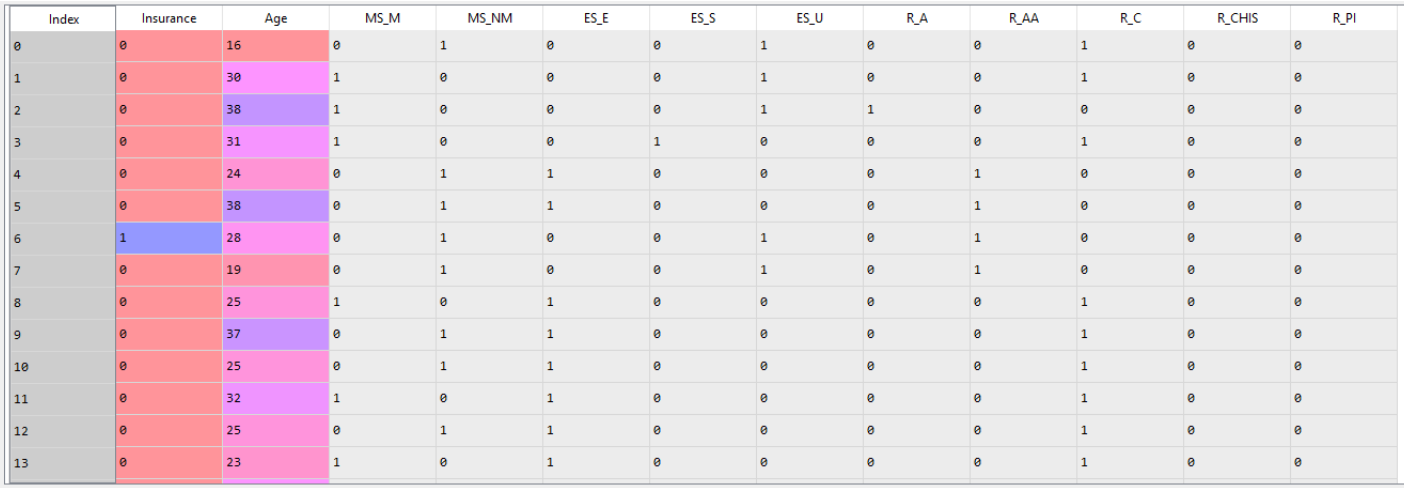
assign(s) a 1 or 0 (True/False) value to the column. This has the benefit of not weighting a value improperly" (Pathak). In this case, the pandas function get_dummies was used. Pandas is a tool that allows users to perform data analysis and manipulation quickly and easily. Tables 12 and 13 show what the data looked like in Spyder (an application that supports the Python coding language) before and after applying the one-hot encoding to create the dummy variables.

Table 12 – Part of the dataset before the one-hot encoding was applied



Index	Insurance	Marital Status	Age	Employment Status	Race
0	0	NM	16	U	C
1	0	M	30	U	C
2	0	M	38	U	A
3	0	M	31	S	C
4	0	NM	24	E	AA
5	0	NM	38	E	AA
6	1	NM	28	U	AA
7	0	NM	19	U	AA
8	0	M	25	E	C
9	0	NM	37	E	C
10	0	NM	25	E	C
11	0	M	32	E	C
12	0	NM	25	E	C
13	0	M	23	E	C

Table 13 – Part of the dataset after the one-hot encoding was applied



Index	Insurance	Age	MS_M	MS_NM	ES_E	ES_S	ES_U	R_A	R_AA	R_C	R_CHIS	R_PI
0	0	16	0	1	0	0	1	0	0	1	0	0
1	0	30	1	0	0	0	1	0	0	1	0	0
2	0	38	1	0	0	0	1	1	0	0	0	0
3	0	31	1	0	0	1	0	0	0	1	0	0
4	0	24	0	1	1	0	0	0	1	0	0	0
5	0	38	0	1	1	0	0	0	1	0	0	0
6	1	28	0	1	0	0	1	0	1	0	0	0
7	0	19	0	1	0	0	1	0	1	0	0	0
8	0	25	1	0	1	0	0	0	0	1	0	0
9	0	37	0	1	1	0	0	0	0	1	0	0
10	0	25	0	1	1	0	0	0	0	1	0	0
11	0	32	1	0	1	0	0	0	0	1	0	0
12	0	25	0	1	1	0	0	0	0	1	0	0
13	0	23	1	0	1	0	0	0	0	1	0	0

Notice that each level of each categorical variable is given its own column, with zeroes and ones to denote whether the subject had that characteristic or not. Thus, the first line of the data reads as a 16 year-old Caucasian woman who is insured, unmarried, and unemployed.

A logistic regression model estimates $E[y]$, which is defined as “the estimated probability that y = 1”, or in this case, the probability that someone is uninsured (Mendenhall 495). This probability will always lie between (or equal to) 0 or 1, with 1 being a hundred percent probability that someone does not have insurance and 0 being a zero percent chance of being uninsured, otherwise worded as a hundred percent probability that they are insured. A value of 0.5 would mean that they are equally likely to have insurance as to not have insurance.

In general, the binary logistic regression model for $E[y]$ can be written as:

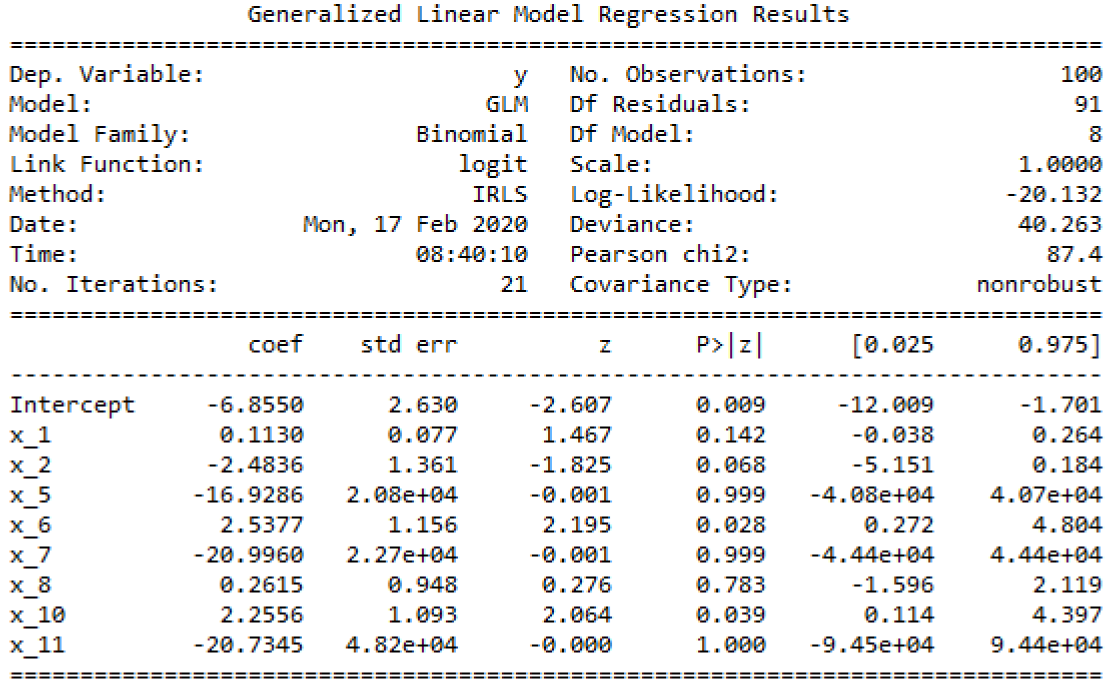
$$E[y] = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

(Mendenhall p. 495).   Here, the βs represent the coefficients of the equation while each x represents a quantitative or qualitative independent variable. For this project, the equation would look like this:

$$E[y] = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{10} x_{10} + \beta_{11} x_{11})}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{10} x_{10} + \beta_{11} x_{11})}$$

where x_1 represents age, x_2 represents married, x_5 represents self-employed, x_6 represents unemployed, x_7 represents Asian, x_8 represents African-American, x_{10} represents Hispanic, and x_{11} represents Pacific Islander.   The output for this model is shown below in Table 14.

Table 14 – Output from the first version of the model



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       91
Model Family:                Binomial   Df Model:                            8
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -20.132
Date:                Mon, 17 Feb 2020   Deviance:                       40.263
Time:                        08:40:10   Pearson chi2:                     87.4
No. Iterations:                    21   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -6.8550      2.630     -2.607      0.009     -12.009      -1.701
x_1            0.1130      0.077      1.467      0.142      -0.038       0.264
x_2           -2.4836      1.361     -1.825      0.068      -5.151       0.184
x_5          -16.9286   2.08e+04     -0.001      0.999   -4.08e+04    4.07e+04
x_6            2.5377      1.156      2.195      0.028       0.272       4.804
x_7          -20.9960   2.27e+04     -0.001      0.999   -4.44e+04    4.44e+04
x_8            0.2615      0.948      0.276      0.783      -1.596       2.119
x_10           2.2556      1.093      2.064      0.039       0.114       4.397
x_11         -20.7345   4.82e+04     -0.000      1.000   -9.45e+04    9.44e+04
==============================================================================
```

The output was calculated using the statsmodels glm function with a binomial family. This function calculates several different statistics, including the log-likelihood, deviance, Pearson chi-square value, coefficients, standard error, p-values, and confidence intervals. Statsmodels glm function uses the maximum likelihood estimation method to estimate the β parameters. This method has several desirable characteristics, including “that the data need not be replicated to apply maximum likelihood estimation” (Mendenhall 496). Instead of the standard *F* and *t* distributions of least squares regression, “the test statistics for testing individual parameters and overall model adequacy have approximate chi-square (X^2) distributions. The X^2 distribution is similar to the *F* distribution in that it depends on degrees of freedom and is nonnegative” (Mendenhall 496-497). The Pearson chi-square statistic can be used to determine model adequacy using the chi-square test. The null and alternative hypotheses for the chi-square test for this model are:

$$H_0: \beta_1 = \beta_2 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_{10} = \beta_{11} = 0$$

$$H_a: at\ least\ one\ of\ \beta_i \neq 0, \qquad i = 1, 2, 5, 6, 7, 8, 10, 11$$

For this test, the null hypothesis is rejected if the p-value is < α or if the test statistic is > the X^2 value. The test statistic is given by the model output as 87.4. If a significance level of α = 0.05 is used, the X^2 value with $k = 8$ degrees of freedom is equal to 15.5073. Since 87.4 > 15.5073, the null hypothesis is rejected and the model is adequate for predicting insurance status y.

The log-likelihood and deviance values provide a way to measure how good the model is at fitting the data. The model fits the data better as both statistics move closer to zero, however, they have no real inferential value unless different versions of the same

model are compared. The values for each $\hat{\beta}_i$ are given, as well as the standard error, and these values are used to compute the test statistic z that can be used in a hypothesis test to determine whether an individual term is significant. The null and alternative hypotheses for this test are

$$H_0:\ \beta_k = 0$$

$$H_a:\ \beta_k \neq 0, \qquad k = 1, 2, 5, 6, 7, 8, 10, or\ 11$$

Z is calculated to find the p-value. If the p-value is less than a chosen alpha of 0.05, the variable is significant to the predicted value of the model. These p-values are shown in the output, in the column labeled $P > |z|$. Notice that there are two p-values less than 0.05, x_6 and x_{10}, which represent unemployment and Hispanic. For these variables, the null hypothesis can be rejected and thus they are statistically significant to the prediction of how likely a woman is to have health insurance. X_2, representing married women, is close enough to 0.05 that it could be considered significant as well.

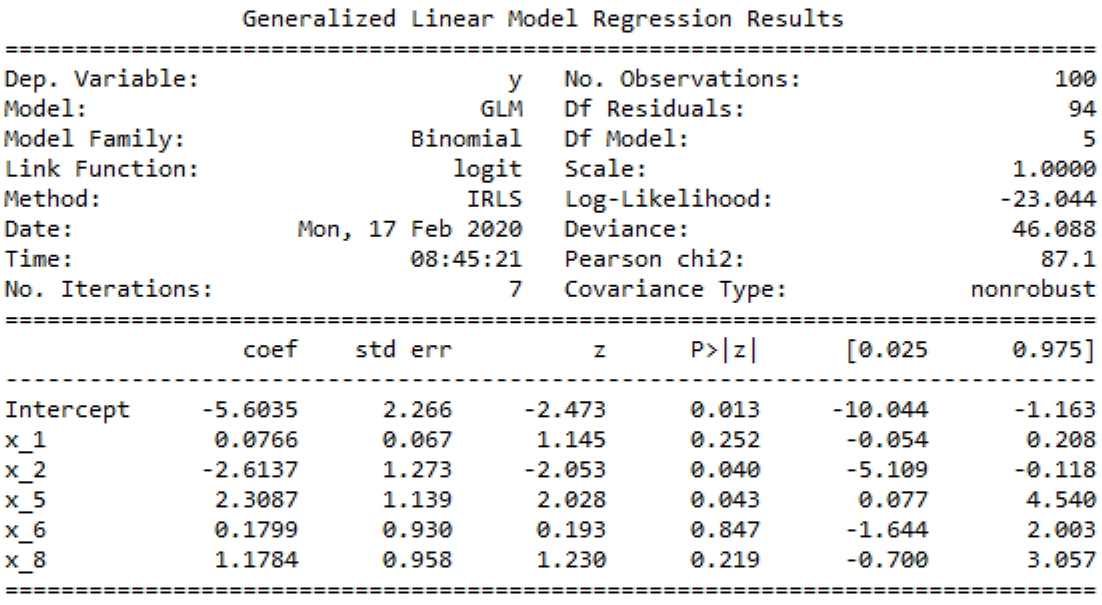
To interpret the beta values, analysts typically "compute $\left(e^{\widehat{\beta}_i}\right) - 1$, which is an estimate of the percentage increase or decrease in the odds" (Mendenhall p. 500). For example, $\hat{\beta}_1$ was computed to be 0.113, so $e^{\widehat{\beta}_1}$ = 1.1196, and $\left(e^{\widehat{\beta}_1}\right) - 1$ = 0.1196. This means that for each additional year older the woman is (x_1), we estimate the odds of being uninsured to increase by 11.96%, holding $x_2, x_5, x_6, x_7, x_8, x_{10}$, and x_{11} fixed. A similar analysis can be applied to the predictive variables that were found to be statistically significant or close to it: x_2, x_6 , and x_{10}. $\hat{\beta}_2 = -2.4836$, so $e^{\widehat{\beta}_2}$ = 0.0834, and $\left(e^{\widehat{\beta}_2}\right) - 1$ = -0.9166. This means that if a woman is married, we estimate the odds of being uninsured to decrease by 91.66%, holding $x_1, x_5, x_6, x_7, x_8, x_{10}$, and x_{11} fixed. Next, $\hat{\beta}_6 = 2.5377$, so $e^{\widehat{\beta}_6}$ = 12.6505, and $\left(e^{\widehat{\beta}_6}\right) - 1$ = 11.6505. This means that if a woman is unemployed, we estimate the odds of being uninsured to increase by 1165%, holding $x_1, x_2, x_5, x_7, x_8, x_{10}$, and x_{11} fixed. Lastly, $\hat{\beta}_{10} = 2.2556$, so $e^{\widehat{\beta}_{10}}$ = 9.541, and $\left(e^{\widehat{\beta}_{10}}\right) - 1$ = 8.541. This means that if a woman is Hispanic, we estimate the odds of being uninsured to increase by 854.1%, holding $x_1, x_2, x_5, x_6, x_7, x_8$, and x_{11} fixed. Notice that all of these statements coincide with what intuition would assume.

However, the detail to recognize about this model is that x_5, x_7, and x_{11}, representing self-employment, Asian, and Pacific Islander respectively, have very large beta values and standard errors. This is likely because the dataset has very little data in each of these categories and each observation happens to also be insured, hence the large coefficients. This would cause the model to be very inaccurate in predicting whether or not someone with these characteristics has insurance. To fix this, the strategy of grouping was implemented. The self-employed level was combined with employment and Hispanic, Asian, and Pacific Islander were combined to form their own level. The latter three categories were combined because they are the smallest and just combining Asian and Pacific Islander would not have solved the problem. With this strategy,

$$x_8 = \begin{Bmatrix} 1\ if\ the\ subject\ is\ Hispanic, Asian, or\ Pacific\ Islander \\ 0\ if\ not \end{Bmatrix}$$

and self-employment is now represented in β_0. Table 15 shows the output of the model with the new groupings.

Table 15 – Output from the second version of the model (grouping)



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       94
Model Family:                Binomial   Df Model:                            5
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -23.044
Date:                Mon, 17 Feb 2020   Deviance:                       46.088
Time:                        08:45:21   Pearson chi2:                     87.1
No. Iterations:                     7   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.6035      2.266     -2.473      0.013     -10.044      -1.163
x_1            0.0766      0.067      1.145      0.252      -0.054       0.208
x_2           -2.6137      1.273     -2.053      0.040      -5.109      -0.118
x_5            2.3087      1.139      2.028      0.043       0.077       4.540
x_6            0.1799      0.930      0.193      0.847      -1.644       2.003
x_8            1.1784      0.958      1.230      0.219      -0.700       3.057
==============================================================================
```

The grouping fixed the high coefficients and standard errors, however, the log-likelihood and deviance got further from zero, and the Pearson chi-square statistic decreased, although not by a large amount. The null and alternative hypotheses for the chi-square test are

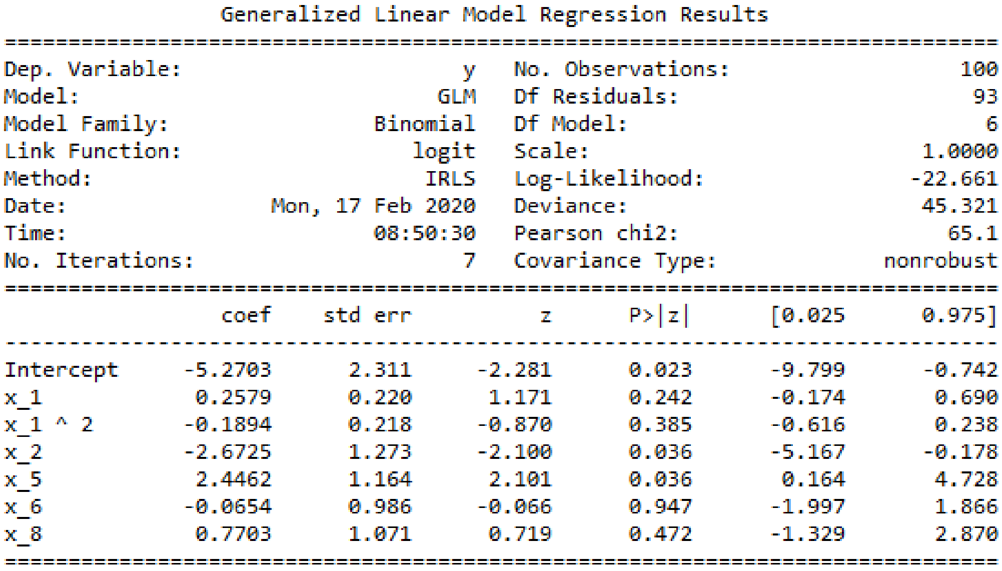
$$H_0: \beta_1 = \beta_2 = \beta_5 = \beta_6 = \beta_8 = 0$$

$$H_a: at\ least\ one\ of\ \beta_i \neq 0, \qquad i = 1, 2, 5, 6, 8$$

If a significance level of α = 0.05 is used, the X^2 value with $k = 5$ degrees of freedom is equal to 11.0705. Since 87.4 > 11.0705, the null hypothesis is rejected and the model is adequate for predicting insurance status y. Notice that in this model, the intercept, x_2, and x_5 have p-values that are less than 0.05, thus showing that these variables are statistically significant to the prediction of how likely a woman is to have health insurance.

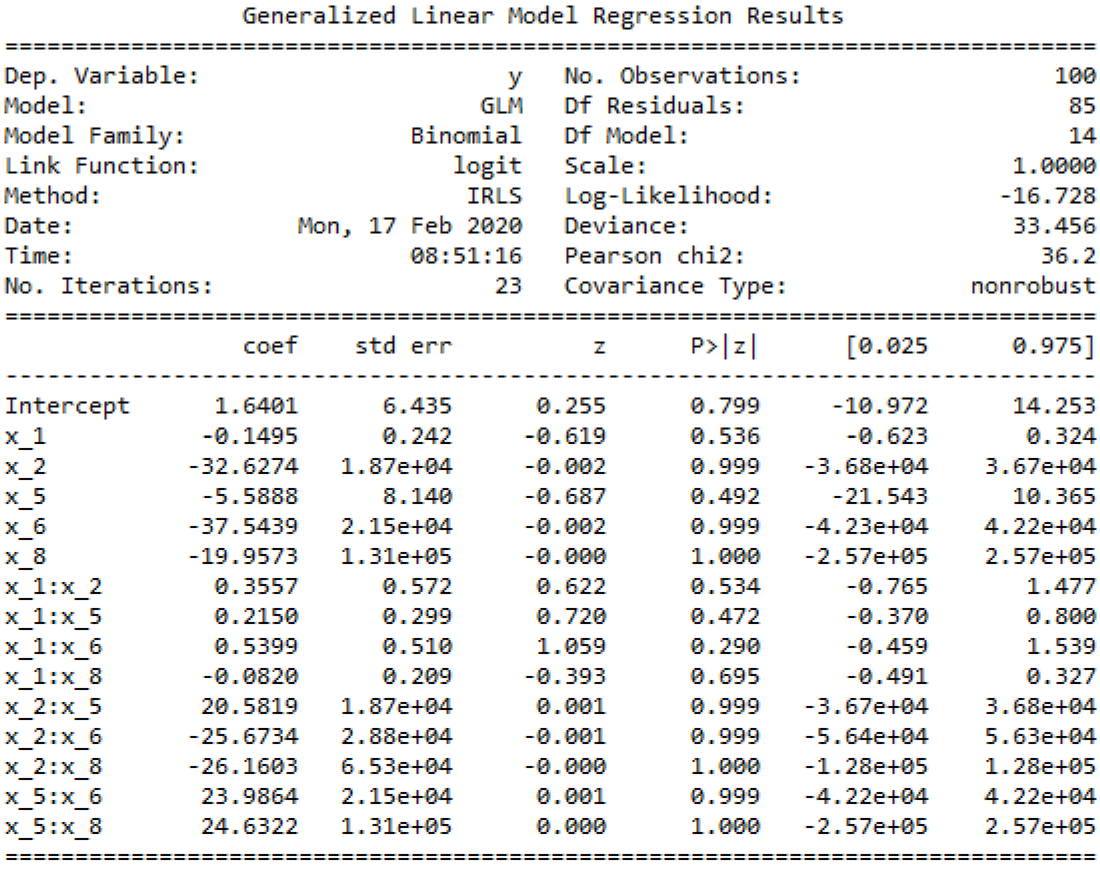
Several other versions of the grouped model were tried in order to see which one provided the best fit for the data. A model with a squared term for x_1 was implemented as well as several combinations of interaction terms. Although the scatterplot between insurance status and age did not reveal a trend, a model that included a squared term was tested to see if a quadratic model would fit the data better than a linear one. In the first two models, the assumption was made that all of the predictive variables were independent of each other in order to start with a simpler model. However, in reality it is unlikely that age, employment status, marital status, and race have absolutely no effect on another. Thus, interaction terms were added in the third model. The interaction terms allow for possible relationships any two variables may have with each other. Tables 16, 17, and 18 show each model's output.

Table 16 – Output of the third version of the model (squared term)



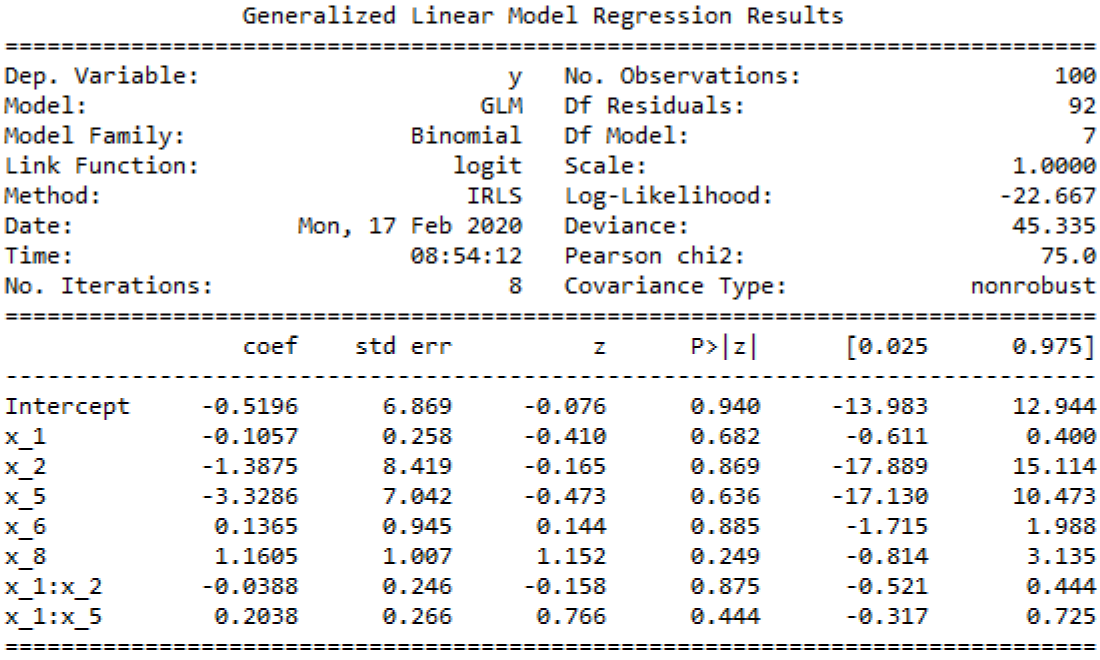
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       93
Model Family:                Binomial   Df Model:                            6
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -22.661
Date:                Mon, 17 Feb 2020   Deviance:                       45.321
Time:                        08:50:30   Pearson chi2:                     65.1
No. Iterations:                     7   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.2703      2.311     -2.281      0.023      -9.799      -0.742
x_1            0.2579      0.220      1.171      0.242      -0.174       0.690
x_1 ^ 2       -0.1894      0.218     -0.870      0.385      -0.616       0.238
x_2           -2.6725      1.273     -2.100      0.036      -5.167      -0.178
x_5            2.4462      1.164      2.101      0.036       0.164       4.728
x_6           -0.0654      0.986     -0.066      0.947      -1.997       1.866
x_8            0.7703      1.071      0.719      0.472      -1.329       2.870
==============================================================================
```

Table 17 – Output of the fourth version of the model (possible interaction terms)



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       85
Model Family:                Binomial   Df Model:                           14
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -16.728
Date:                Mon, 17 Feb 2020   Deviance:                       33.456
Time:                        08:51:16   Pearson chi2:                     36.2
No. Iterations:                    23   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.6401      6.435      0.255      0.799     -10.972      14.253
x_1           -0.1495      0.242     -0.619      0.536      -0.623       0.324
x_2          -32.6274   1.87e+04     -0.002      0.999   -3.68e+04    3.67e+04
x_5           -5.5888      8.140     -0.687      0.492     -21.543      10.365
x_6          -37.5439   2.15e+04     -0.002      0.999   -4.23e+04    4.22e+04
x_8          -19.9573   1.31e+05     -0.000      1.000   -2.57e+05    2.57e+05
x_1:x_2        0.3557      0.572      0.622      0.534      -0.765       1.477
x_1:x_5        0.2150      0.299      0.720      0.472      -0.370       0.800
x_1:x_6        0.5399      0.510      1.059      0.290      -0.459       1.539
x_1:x_8       -0.0820      0.209     -0.393      0.695      -0.491       0.327
x_2:x_5       20.5819   1.87e+04      0.001      0.999   -3.67e+04    3.68e+04
x_2:x_6      -25.6734   2.88e+04     -0.001      0.999   -5.64e+04    5.63e+04
x_2:x_8      -26.1603   6.53e+04     -0.000      1.000   -1.28e+05    1.28e+05
x_5:x_6       23.9864   2.15e+04      0.001      0.999   -4.22e+04    4.22e+04
x_5:x_8       24.6322   1.31e+05      0.000      1.000   -2.57e+05    2.57e+05
==============================================================================
```

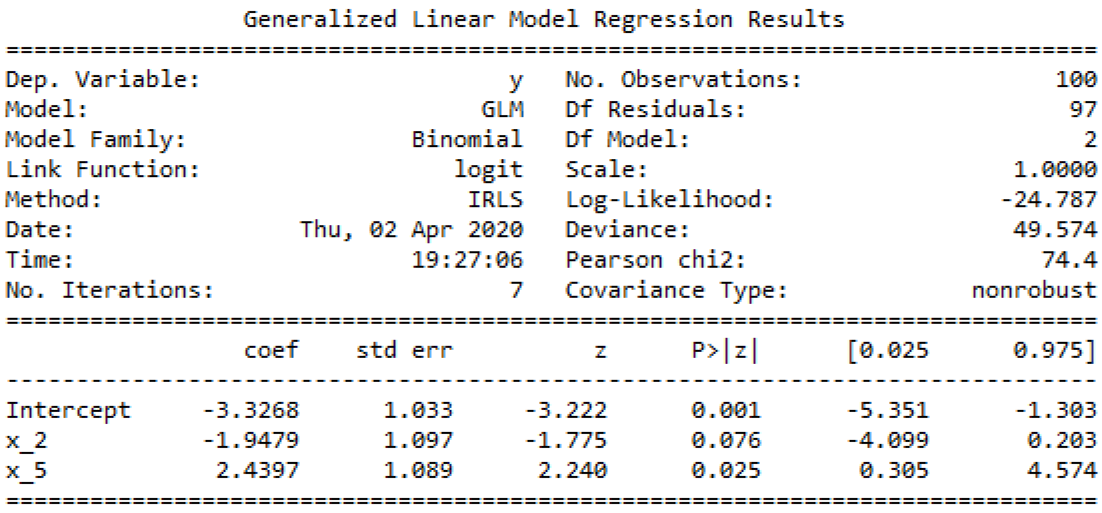
Table 18 – Output of the fifth version of the model (leaving out some interaction terms)



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       92
Model Family:                Binomial   Df Model:                            7
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -22.667
Date:                Mon, 17 Feb 2020   Deviance:                       45.335
Time:                        08:54:12   Pearson chi2:                     75.0
No. Iterations:                     8   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.5196      6.869     -0.076      0.940     -13.983      12.944
x_1           -0.1057      0.258     -0.410      0.682      -0.611       0.400
x_2           -1.3875      8.419     -0.165      0.869     -17.889      15.114
x_5           -3.3286      7.042     -0.473      0.636     -17.130      10.473
x_6            0.1365      0.945      0.144      0.885      -1.715       1.988
x_8            1.1605      1.007      1.152      0.249      -0.814       3.135
x_1:x_2       -0.0388      0.246     -0.158      0.875      -0.521       0.444
x_1:x_5        0.2038      0.266      0.766      0.444      -0.317       0.725
==============================================================================
```

Since the p-value of 0.385 is not < .05, the squared term in Table 16 was deemed to not be significant to the prediction of insurance status. Thus, the decision was made to leave it out of future versions of the model. After seeing the output in Table 17, the decision was made to leave out some interaction terms as well. This is because some interactions did not make sense (such as age and race) or because they produced unreasonably large coefficients and standard errors. Each model also proved to be adequate for predicting insurance status y using a chi-square test, since they all had Pearson chi-square statistics that were greater than their needed chi-square values of 12.5916, 23.6848, and 14.0671, respectively. However, none of their Pearson chi-square statistics were as large as the second model. Once the fifth model ran, although all the statistics were reasonable, none of the p-values for the interaction terms were less than 0.05. This means that while there may be relationships between predictor variables, they are not significant enough to cause the prediction of y to be inaccurate. A sixth model was run where the second model was used as the base, but all terms with a p-value greater than 0.05 were removed, leaving x_2 and x_5. The output for this model is shown below.

Table 19 – Output of the sixth model (statistically significant terms only)



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       97
Model Family:                Binomial   Df Model:                            2
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -24.787
Date:                Thu, 02 Apr 2020   Deviance:                       49.574
Time:                        19:27:06   Pearson chi2:                     74.4
No. Iterations:                     7   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.3268      1.033     -3.222      0.001      -5.351      -1.303
x_2           -1.9479      1.097     -1.775      0.076      -4.099       0.203
x_5            2.4397      1.089      2.240      0.025       0.305       4.574
==============================================================================
```

This model also proves to be adequate for predicting insurance status y using a chi-square test, since it had a Pearson chi-square statistic of 74.4, which is greater than 5.99147, but not as large as the second model's statistic of 87.1. Hence it was determined that the second model should be the equation used to predict $E[y]$, as it was the most accurate and adequate.

Consider a scenario where the modeler wants to predict how likely a 22 year-old, unmarried, employed, Caucasian woman is to be uninsured. The equation for the chosen model is

$$E[y] = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_6 x_6 + \beta_8 x_8)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_6 x_6 + \beta_8 x_8)}$$

X_2, x_5, x_6, and x_8 are all zero, since not married, employed, and Caucasian are base levels. The equation is reduced to

$$E[y] = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

After plugging in the coefficient values shown in Table 14, the equation is

$$\widehat{E[y]} = \frac{\exp\left(-5.6035 + 0.0766(22)\right)}{1 + \exp\left(-5.6035 + 0.0766(22)\right)} = 0.0195$$

This can be interpreted as there is a 1.95% chance that a 22 year-old unmarried, employed, Caucasian woman does not have health insurance. Next, consider a scenario where the modeler would like to predict how likely it is that a 30 year-old, married, unemployed, African-American woman does not have insurance. Then the equation would be as follows:

$$E[y] = \frac{\exp(\beta_0 + \beta_1(30) + \beta_2(1) + \beta_5(1) + \beta_6(1) + \beta_8(0))}{1 + \exp(\beta_0 + \beta_1(30) + \beta_2(1) + \beta_5(1) + \beta_6(1) + \beta_8(0))}$$

With the coefficient values plugged in, the modeler has

$$\widehat{E[y]} = \frac{\exp(-5.6035 + 0.0766(30) + \ -2.6137(1) + \ 2.3087(1) + \ 0.1799(1))}{1 + \exp(-5.6035 + 0.0766(30) + \ -2.6137(1) + \ 2.3087(1) + \ 0.1799(1))}$$

$$= \ .0314$$

This is interpreted as there is a 3.14% chance that a 30 year-old, married, unemployed, African-American woman does not have health insurance. Thus, a woman with these demographic characteristics is nearly twice as likely to be uninsured than a 22 year-old unmarried, employed, Caucasian woman.

Contingency chi-square tables were created using the statsmodels chi2_contingency function to test "the association between the row and column variables in a … table" (Two-Way Tables and the Chi-Square Test). This method provides a direct comparison of each independent categorical variable to the dependent one. "The null hypothesis H_0 assumes that there is no association between the variables (in other words, one variable does not vary according to the other variable), while the alternative hypothesis H_a claims that some association does exist" (Two-Way Tables and the Chi-Square Test). Marital status, employment status, and race all produced significant results, with respective p-values of 0.028, 0.007, and 0.0007 being less than an α of 0.05. A one-way ANOVA test was performed using the statsmodels f_oneway function to compare the continuous variable, age, against insurance status. These two variables did not prove to have a significant relationship as the test produced a p-value of 0.822, which is greater than 0.05, so the null hypothesis was rejected.

Conclusion

In conclusion, after an extensive analysis of data from Miami Valley Hospital in Montgomery County, Ohio, it was determined that the demographic characteristics marital status, employment status, and race do affect how likely a woman is to not have health insurance. It was also found through other statistical tests that race had the most compelling association with insurance, although employment status and marital status also had significant associations. While this correlation is by no means assumed to be causal, it is known that they are associated, and more research could be done to find out what exactly the relationship is. It was also found that different combinations of demographic characteristics produced different likelihoods of being uninsured. Thus, an insurance company would be able to take the chosen model and test all combinations of demographics to find the ones that produce the lowest probability of being insured. This would allow them to create and market a health insurance product specifically towards those demographics with the hope of more women being insured and lowering the neonatal and perinatal death rates in Montgomery County. For future study, a 70/30 train/test data split should be implemented in order to prevent the model overfitting or underfitting. It was not incorporated into this thesis because the necessary code could not be reconciled with code used to create the models. Variable screening should also be considered, as it was not able to be executed in this project due to the technique’s complexity in Python. A larger dataset should be used as well to confirm the results from this undertaking. All of these methods will help ensure that the products developed by the insurance companies are as effective as possible.

Appendix

[1]Ten leading causes of infant deaths:

1. Congenital malformations, deformations and chromosomal abnormalities (congenital malformations)
2. Disorders related to short gestation and low birth weight, not elsewhere classified (low birth weight)
3. Newborn affected by maternal complications of pregnancy (maternal complications)
4. Sudden infant death syndrome (SIDS)
5. Accidents (unintentional injuries)
6. Newborn affected by complications of placenta, cord and membranes (cord and placental complications)
7. Bacterial sepsis of newborn
8. Diseases of the circulatory system
9. Respiratory distress of newborn
10. Neonatal hemorrhage

Bibliography

"2 x 2 Contingency Chi-Square." *Pdx.Edu*, web.pdx.edu/~newsomj/uvclass/ho_chisq.pdf.

"Deaths and Mortality." *Cdc.gov*, U.S. Department of Health and Human Services, 3 May 2017, www.cdc.gov/nchs/fastats/deaths.htm.

"Home Birth: Benefits and Tips." *Americanpregnancy.org*, American Pregnancy Association, 13 Oct. 2019, americanpregnancy.org/labor-and-birth/home-birth.

Hosmer, David W., and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2000.

Houston, Rickie. "What Is the Cost of Having a Baby in 2019?" *SmartAsset*, SmartAsset, 8 Oct. 2019, smartasset.com/financial-advisor/cost-of-having-a-baby.

Huang, Pien. "How The Novel Coronavirus And The Flu Are Alike ... And Different." *Npr.org*, 20 Mar. 2020, choice.npr.org/index.html?origin=https://www.npr.org/sections/goatsandsoda/2020/03/20/815408287/how-the-novel-coronavirus-and-the-flu-are-alike-and-different.

"Infant Mortality Data Montgomery County, Ohio 2012-2016." *Public Health Dayton and Montgomery County*, www.phdmc.org/epidemiology/health-profiles/774-2016-montgomery-county-infant-mortality/file.

Institute of Medicine (US) Committee on the Consequences of Uninsurance. "Health-Related Outcomes for Children, Pregnant Women, and Newborns - Health Insurance Is a

Family Matter - NCBI Bookshelf.” *Ncbi.Nlm.Nih.Gov*, National Center of Biotechnology Information, www.ncbi.nlm.nih.gov/books/NBK221019. Accessed 29 Mar. 2020.

Kamal, Rabah. “What Do We Know about Infant Mortality in the U.S. and Comparable Countries?” *Peterson-Kaiser Health System Tracker*, 18 Oct. 2019, www.healthsystemtracker.org/chart-collection/infant-mortality-u-s-compare-countries/#item-infant-mortality-is-higher-in-the-u-s-than-in-comparable-countries_2019.

McClurg, Lesley. “Home Birth Can Be Appealing, but How Safe Is It?” *Npr.org*, 11 Mar. 2019, choice.npr.org/index.html?origin=https://www.npr.org/sections/health-shots/2019/03/11/700829719/home-birth-can-be-appealing-but-how-safe-is-it.

Mendenhall, William, and Terry Sincich. A Second Course in Statistics: Regression Analysis. Prentice-Hall, 1996.

Pathak, Manish. “(Tutorial) Handling Categorical Data in Python.” *DataCamp Community*, 6 Jan. 2020, www.datacamp.com/community/tutorials/categorical-data.

“Random sampling for patients seeking care at OB tertiary care center in Montgomery County, Ohio 11/1/2019 – 12/1/2019.” Miami Valley Hospital.

“Two-Way Tables and the Chi-Square Test.” *Stat.Yale.Edu*, www.stat.yale.edu/Courses/1997-98/101/chisq.htm.

Wax, Joseph, et al. “Maternal and Newborn Outcomes in Planned Home Birth vs Planned Hospital Births: A Metaanalysis.” *The American Journal of Obstetrics and Gynecology*, vol. 203, no. 3, 2010, p. 243, www.ajog.org/article/S0002-9378(10)00671-X/fulltext.