Logistic Regression Homework Assignment #1

*Homework Team 7: Diana Shaw, Bright Small, Andrew Williams, Logan Corbett, Rob Causey*

**Executive summary**

This new strand of the flu appears to be a universal problem affecting individuals of all gender, race, age, income level, regardless of whether they have regularly visited the doctor, live near a hospital, or recently had another strain of flu. Within this set of predictors, the strongest association is between flu and gender, with males having greater odds of getting the flu than females. The association between gender and contracting the new flu strand does not have any confounding or interactions with income level.

To predict the probability of contracting the new flu strand, the team developed a model based on gender and income. Although there is no association between income level and contracting the new flu strand when the two variables are looked at in isolation, income level is included in the model because it was found to be a statistically significant predictor when introduced into the model in the presence of gender. The final model produces 54.3% concordant pairs and only 24% discordant pairs.

|  |  |
| --- | --- |
| **Accuracy of Predictive Model Based on Gender and Income** | |
| **Percent Concordant** | 54.3 |
| **Percent Discordant** | 24 |
| **Percent Tied** | 21.7 |

The following report details our procedures and analysis.

## Examining Flu Association with Gender and Income

The contingency table below shows that males have a 44.16% chance of contracting the flu strand compared to only a 23.32% chance for females.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table of Flu Cases by Gender** | | | | |
|  | | **Flu** | | **Total** |
| **Yes** | **No** |
| **Female** | **Frequency** | 45.00 | 148.00 | 193.00 |
| **Row Pct** | 23.32 | 76.68 |  |
| **Male** | **Frequency** | 68.00 | 86.00 | 154.00 |
| **Row Pct** | 44.16 | 55.84 |  |
| **Total** | **Frequency** | 113.00 | 234.00 | 347.00 |
| **Percent** | 32.56 | 67.44 | 100.00 |

|  |  |  |
| --- | --- | --- |
| **Odds Ratio (Females/Males)** | | |
| **Value** | **95% Confidence Limits** | |
| 0.3845 | 0.2425 | 0.6097 |

The odds ratio between females and males is 0.38, meaning that females have 0.38 times the odds of catching the flu as males. The 95% confidence interval for the odds ratio is 0.24 – 0.61. Since the confidence interval does not include 1, the odds ratio indicates an association exists between the variables Flu and Gender.

Income level was also examined for potential association with flu cases, and the contingency table for flu and income level is shown in the following table.

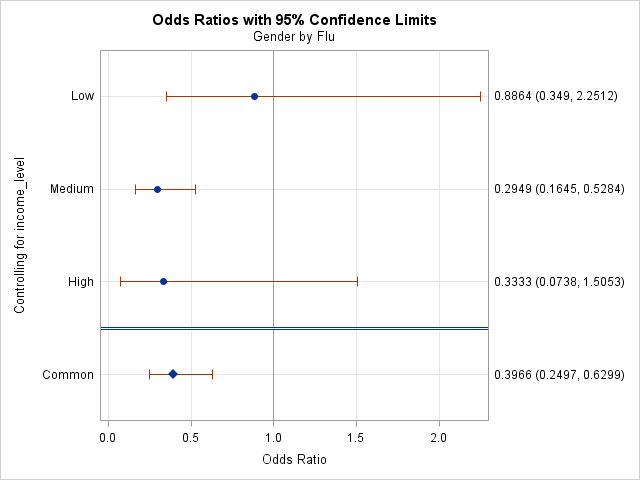
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table of Flu Cases by Income Level** | | | | |
| **Income Level** | | **Flu** | | **Total** |
| **Yes** | **No** |
| **High** | **Frequency** | 9.00 | 34.00 | 43.00 |
| **Row Pct** | 20.93 | 79.07 |  |
| **Medium** | **Frequency** | 80.00 | 130.00 | 210.00 |
| **Row Pct** | 38.10 | 61.90 |  |
| **Low** | **Frequency** | 24.00 | 70.00 | 94.00 |
| **Row Pct** | 25.53 | 74.47 |  |
| **Total** | **Frequency** | 113.00 | 234.00 | 347.00 |
| **Percent** | 32.56 | 67.44 | 100.00 |

Since income level is an ordinal variable and flu is a binary variable, the Mantel-Haenszel statistic is used to test for association. The Mantel-Haenszel statistic results in a p-value of 0.76, meaning that there is not sufficient evidence to indicate a relationship between flu cases and income level. This can also be seen in the contingency table, which shows that the Medium income level has the highest row percentage of flu cases.

The Spearman Correlation coefficient between flu and income level is 0.0263 (indicating very weak association) confirms that there is no association.

## Testing for Confounding Between Gender and Income

When controlling for Income, there is no evidence of confounding because the confidence interval for the adjusted odds ratio is 0.2497 – 0.6299 and contains the original odds ratio of .3845. The confidence interval for the adjusted odds ratio is shown in the following chart.



## Testing for Interactions Between Gender and Income

Using results from calculation of the Breslow-Day-Tarone statistic and the Zelen’s exact test, there is not significant evidence of an interaction between income and gender.

| **Tests for Homogeneity of Odds Ratios** | |
| --- | --- |
| **Breslow-Day-Tarone Chi-Square** | 3.9681 |
| DF | 2 |
| Pr > ChiSq | 0.1375 |
|  |  |
| **Zelen's Exact Test (P)** | 0.0099 |
| Exact Pr <= P | 0.1380 |

When running the stratified analysis to test for interactions in SAS, a warning was observed that said that 25% of the cells have expected counts less than 5. This means that Zelen’s exact test is the better test for this data set.

## Predictive Modeling Using Logistic Regression

Using logistic regression techniques, a predictive model can be developed to predict the probability of an individual contracting the new strain of flu.

When running the Likelihood Ratio test on an initial model containing all variables, the results show that the model is overall statistically significant. Since our total sample size is 347, the alpha level for testing significance should be set no higher than 0.032, and the p-value for the global Likelihood Ratio test falls well below that (see table below).

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| Likelihood Ratio | 32.8588 | 10 | 0.0003 |
| Score | 31.4880 | 10 | 0.0005 |
| Wald | 28.0824 | 10 | 0.0018 |

According the Analysis of Effects table below, Gender is the only statistically significant predictor variable when all variables are present in the model.

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| Gender | 1 | 14.4226 | 0.0001 |
| Age | 1 | 0.9601 | 0.3272 |
| Distance | 1 | 0.0155 | 0.9011 |
| Income | 2 | 5.8659 | 0.0532 |
| Previous | 1 | 0.2345 | 0.6282 |
| Race | 3 | 6.0396 | 0.1097 |
| Visits | 1 | 0.2966 | 0.5860 |

Using the stepwise model selection technique with an entry criteria of p < 0.3 and a stay criteria of   
p < 0.1 results in a model with only gender and income as predictor variables. When these two variables are in the model together, gender has a p-value well below the cutoff of 0.032 and income has a p-value just above the cutoff.

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| Gender | 1 | 15.4844 | <.0001 |
| Income | 2 | 6.5806 | 0.0372 |

As can be seen in the tables below, keeping income in the model increases the predictive accuracy of the model by more than 16% with respect to concordant pairs, so a judgment call is made to keep income in the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy of Model Based on Gender and Income** | |  | **Accuracy of Model Based on Gender** | |
| **Percent Concordant** | 54.3 |  | **Percent Concordant** | 38.1 |
| **Percent Discordant** | 24 |  | **Percent Discordant** | 14.6 |
| **Percent Tied** | 21.7 |  | **Percent Tied** | 47.3 |

The final model does not include any continuous variables, so there is no need to test the linearity assumption for logistic regression models. A Likelihood Ratio test of the final model containing Gender and Income as predictor variables is overall statistically significant (see test results below).

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| Likelihood Ratio | 23.7876 | 3 | <.0001 |