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 genders, races, ages, . . [all variables]

There is an association between gender and flu, 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.

No association between income level and flu (contracting the new strand

We were able to develop a model to predict the probability of contracting the new flu strand based on gender and income.

The following report details our analysis.

## Examining Flu Association with Gender and Income

[Text]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table of Flu Cases by Gender** | | | | |
| **Gender** | | **Flu** | | **Total** |
| **No** | **Yes** |
| **Female** | **Frequency** | 148.00 | 45.00 | 193.00 |
| **Percent** | 42.65 | 12.97 | 55.62 |
| **Row Pct** | 76.68 | 23.32 |  |
| **Col Pct** | 63.25 | 39.82 |  |
| **Male** | **Frequency** | 86.00 | 68.00 | 154.00 |
| **Percent** | 24.78 | 19.60 | 44.38 |
| **Row Pct** | 55.84 | 44.16 |  |
| **Col Pct** | 36.75 | 60.18 |  |
| **Total** | **Frequency** | 234.00 | 113.00 | 347.00 |
| **Percent** | 67.44 | 32.56 | 100.00 |

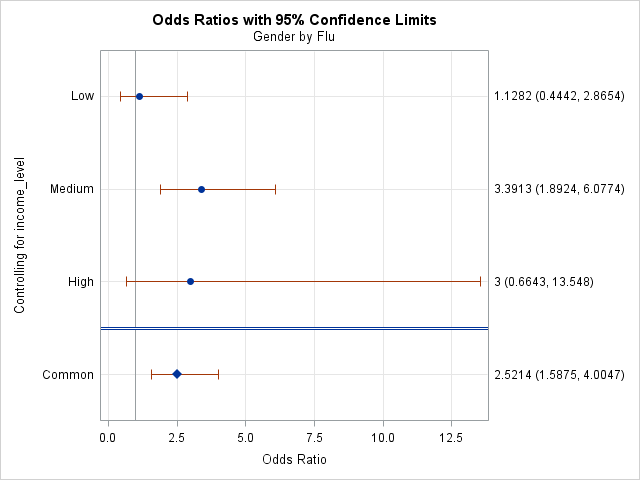
The odds ratio is 0.38 with a 95% confidence interval of 0.24 – 0.61. Since the confidence interval does not include 1, the odds ratio indicates an association exists between the variables Flu and Gender. The odds ratio of 0.38 means that female have 0.38 times the odds of catching the flu as males.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table of Flu Cases by Income Level** | | | | |
| **Income Level** | | **Flu** | | **Total** |
| **No** | **Yes** |
| **High** | **Frequency** | 34.00 | 9.00 | 43.00 |
| **Percent** | 9.80 | 2.59 | 12.39 |
| **Row Pct** | 79.07 | 20.93 |  |
| **Col Pct** | 14.53 | 7.96 |  |
| **Medium** | **Frequency** | 130.00 | 80.00 | 210.00 |
| **Percent** | 37.46 | 23.05 | 60.52 |
| **Row Pct** | 61.90 | 38.10 |  |
| **Col Pct** | 55.56 | 70.80 |  |
| **Low** | **Frequency** | 70.00 | 24.00 | 94.00 |
| **Percent** | 20.17 | 6.92 | 27.09 |
| **Row Pct** | 74.47 | 25.53 |  |
| **Col Pct** | 29.91 | 21.24 |  |
| **Total** | **Frequency** | 234.00 | 113.00 | 347.00 |
| **Percent** | 67.44 | 32.56 | 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 has 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 table above, which shows that the Medium income level has the highest row percentage of flu cases.

The Spearman Correlation coefficient of 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 adjusted confidence interval of 0.2497 – 0.6299 contains the original odds ratio of .3845.

[Text]

## Testing for Interactions Between Gender and Income

[Text]

| **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 |

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

When running the stratified analysis to test for interactions, a SAS 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 Model Using Logistic Regression

[Text]

*Note: Note sure what alpha we want to use in this section. I don’t think it changes the result, be we are asked to state it*

| **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 to the Likelihood Ratio test, the model containing all variables is overall statistically significant.

| **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 |

According the Analysis of Effects table, Gender is the only statistically significant predictor variable.

~~Variables Gender and Income are selected for the final model because Gender is statistically significant and there is evidence of confounding between Gender and Income.~~

We ran model selection . . .

The final model does not include any continuous variables, so there is no need to test the linearity assumption . . .

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

*\*\*Note: When other variables are dropped from the model, the p-value for income is now less than 0.05, but based on Peter’s comment I don’t know if that’s the right alpha.*

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| Likelihood Ratio | 23.7876 | 3 | <.0001 |
| Score | 23.2552 | 3 | <.0001 |
| Wald | 22.1157 | 3 | <.0001 |

According to the Likelihood Ratio test, the final model containing Gender and Income as predictor variables is overall statistically significant.

## Summary of Findings

[Text]