|  |
| --- |
|  |
| Analysis of Incidence of New Flu Strain |
| August 27, 2013 |

|  |
| --- |
| Michael Davidoff, Alex Lentz, Jim Box, Phillip Domschke, & Sarah Rawls  27-Aug-13 |

# Key Findings

We performed an analysis on the data to find some associations between the incidence of a new strain of flu and several variables. The key findings are summarized below:

* There is a statistically significant difference in the incidence of the new strain of the flu and gender. Females had 2.6 times the odds of avoiding the flu than did males.
* Income level is provided in the dataset with values of Low, Medium, and High. There is no statistically significant relationship between these levels of income and incidence of the new strain of the flu.
* The relationship between Gender and Flu is not affected by an interaction with Income Level, nor is there any confounding involving Income level.
* Of all of the predictor variables contained in the dataset, the best model for predicting incidence of the new strain of the flu is one that contains only gender, with the basic finding that Males are much more likely to contract this strain of the flu for the given sample.
* Income level was close to being statistically significant, and it may warrant further investigation. One approach is to look at how the values were classified into the three groups (most were in the medium income level). We recommend exploring more levels or looking at the data as a continuous variable.

Contents

[Key Findings 1](#_Toc365201444)

[Analytical Methods 2](#_Toc365201445)

[Incidence of Flu by Gender 2](#_Toc365201446)

[Incidence of Flu by Income Level 2](#_Toc365201447)

[Incidence of Flu by Gender, Controlling for Income Level 3](#_Toc365201448)

[Model of Probability of Having the Flu 3](#_Toc365201449)

# Analytical Methods

## Incidence of Flu by Gender

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | No Flu | Flu | Total |
| Female | 148 (77%) | 45 (23%) | 193 |
| Male | 86 (56%) | 68 (44%) | 154 |
| Overall | 154 (44%) | 193 (56%) | 347 |

The Odds Ratio was calculated as 2.6, which means that Females have 2.6 times the odds of being Flu-free than do males. The 95% confidence interval for this ratio was (1.64, 4.12), meaning that since the interval does not include 1, this is a statistically significant odds ratio. Additionally, the Mantel-Haenszel Chi-Square test for significance returned a p-value of <0.0001. This is less than the alpha level of 0.0003, which would indicate a strong relationship given our sample size.

## Incidence of Flu by Income Level

|  |  |  |  |
| --- | --- | --- | --- |
| Income | No Flu | Flu | Total |
| Low | 70 (74%) | 24 (26%) | 94 |
| Medium | 130 (62%) | 80 (38%) | 210 |
| High | 34 (79%) | 9 (21%) | 43 |
| Overall | 234 (67%) | 113 (33%) | 347 |

After recoding the income levels, we are able to use the Mantel-Haenszel Chi-Square test of significance between ordinal variables. The result of this test returned a p-value of 0.7633, which is way above the alpha level of 0.0003. It is also above the alpha level of 0.009, which would indicate a weak relationship. As a result, there is no significant relationship between a person’s income level and incidence of this flu strain.

We looked at the Spearman Correlation Coefficient, which looks at the association of the rank of the variables. The value for these variables is 0.0263, which is very close to zero, indicating there is no linear relationship between the ranks of Income level and contracting the flu.

## Incidence of Flu by Gender, Controlling for Income Level

We performed a stratified analysis of Flu by Gender, controlling for Income Level. After controlling for income, the adjusted odds ratio was 2.52, with a 95% confidence interval of (1.588, 4.004). That means that after controlling for income, females have 2.52 times the odds of avoiding the flu than do males. Since this confidence interval contains 2.60, the odds ratio without controlling for income, there does not appear to be any confounding with income level.

We looked to see if there were any interactions between the Income level and the relationship between Gender and Flu. The Tarone’s adjustment for the Breslow-Day statistic had a p-value of 0.1375, which was well about our alpha level of 0.0003. As a result, we cannot reject the null hypothesis that there is no interaction. We also looked at Zelen’s Exact test to account for the problems with small expected cell counts in some of our tests. This test returned a p-value of 0.1380, confirming that there is no interaction.

# Model of Probability of Having the Flu

We performed a logistic regression to determine the probability of contracting the new strain of the flu based on the full set of variables available to us:

* Race (African American, White, Hispanic compared to Other)
* Gender (Male compared to Female)
* Income Level (Low and Medium compared to High)
* Previous Flu (Flu in the past three years compared to flu-free)
* Age (years)
* Distance (miles to closet hospital)
* Visits (annual visits to the doctor’s office)

From our earlier analyses, there were no interactions required for inclusion in the model.

Prior to fitting the model, we checked the assumption that the logit of the probability of getting the flu is linearly related to the continuous variables. We tested this by adding variables for age\*log(age), distance\*log(distance), and visits\*log(visits), and performing a regression. None of these new variables were significant, so the assumption of linearity is confirmed and we did not require a Box-Tidwell correction.

When running the full model, the Global Hypothesis Test returned a p-value of 0.0003 for the likelihood ratio, meaning that there is enough evidence to reject the null hypothesis and conclude that the model has some predictive ability.

From the individual variables, Gender (p-value < 0.0001) and Income (0.0532) seemed to be the best effects to include. This was confirmed when we used a backwards stepwise selection technique to explore the covariates.

We ran the model a second time with just Income Level and Gender. Again, the Global Hypothesis Test was significant at alpha = 0.0003, meaning that the model had some validity. With just these two predictors in the model, Gender was still significant with a p-value of <0.0001, but Income had a p-value of 0.0372, which was above our alpha value. With this sample size, the p-value is too large to say anything other than there is weak significance. Therefore, we propose the model with just Gender as a predictor. This model shows the same p-values for both Gender and the overall hypothesis test as previously discussed, and as such, both are significant at the 0.0003 alpha level.