36-402 DA Exam 2

Richard Kang (rjkang)

4/24/2020

Introduction

Public health researchers in Vietnam are interested in possible obstacles that prevent widespread use of regular check-ups. (1)This project aims to answer 3 questions with data from 2,068 individuals throughout Vietnam. The first question asks how people rate the value and quality of medical service. The next question wants to know what factors, if any, affect a person in getting annual check-ups. Lastly, the Ministry of Health is interested in knowing whether a marketing campaign on the quality of information gained through check-ups will motivate more people to receive check-ups, controlling for whether a person has health insurance or not. (2) This analysis concluded that respondents that rated the quality of information given in check-ups with high scores were about 1.389 times more likely than respondents that rated with low scores in knowing whether patients get check-ups, with a 95% confidence interval of 0.1889 and 10.215. Also, bar plots of ratings of the value and quality of medical service show that many people do not know the importance of getting check-ups, and believe that it is a waste of time. This will be discussed in more detail below.

Exploratory Data Analysis

The data collected by the Ministry of Health consists of 21 variables. From these variables, 9 key explanatory variables were chosen for data analysis. (1)Explanatory variables regarding basic personal information were Jobbstt(Job Status) and HealthIns (Health Insurance), which were categorical variables. In addition to these variables, there were 2 categorical variables that described how that person felt about medical check-ups. Wsttime (Waste time) asks whether the respondent feels check-ups are a waste of time. NotImp (Not Important) shows whether a person believes check-ups are not important nor urgent. A continuous variable, SuitFreq, is also a categorical variable that shows how often the respondent believes check-ups should be done. The last four continuous variables of interest are about how the

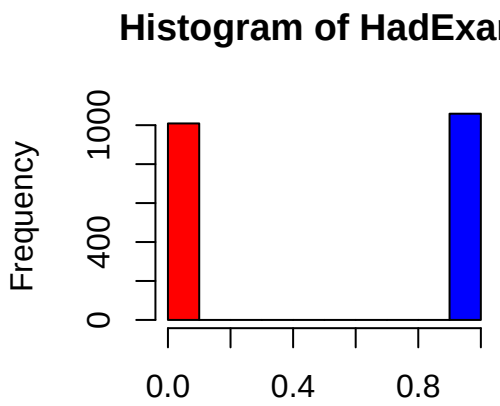

Figure 1: Histogram of the Response Variable

respondents feel about information given in check-ups. These variables will be useful in answering the Health Ministry's question on whether quality of information is an important predictor of whether patients get check-ups. Each of these variables are rated on a scale of 1 to 5. SuffInfo rates the sufficiency of information received in check-ups. AttractInfo rates the attractiveness of information in check-ups. ImpressInfo rates the impressiveness of information in check-ups. Lastly, PopularInfo rates the popularity of information received in check-ups.

Figure 2 below shows the distribution of the 6 categorical variables with non-numerical values. From a glance, we can see what each of the proportions were like. For example, there were more repondents with health insurance, and many respondents were either students or had stable jobs. (2)HadExam is our response variable, which is 1 if the respondent had an exam within the past 12 months, and 0 if not. The histogram of HadExam in Figure 1 shows that the distribution of yes's and no's is even. Within the past 12 months, about half of the respondents had taken an exam, and half had not. (3)Figure 3 is the histogram of the 4 continuous variables with numerical values regarding respondent's feelings about the quality of information in check-ups. The histograms seem to show a generally right-skewed distribution of the responses, meaning that respondents generally did not give very high scores to how they felt about check-ups. (4) Barplots that show a large difference in proportions may help understand people's motivations to get regular check-ups. The job and education status of respondents may affect how they think about the importance of check-ups.

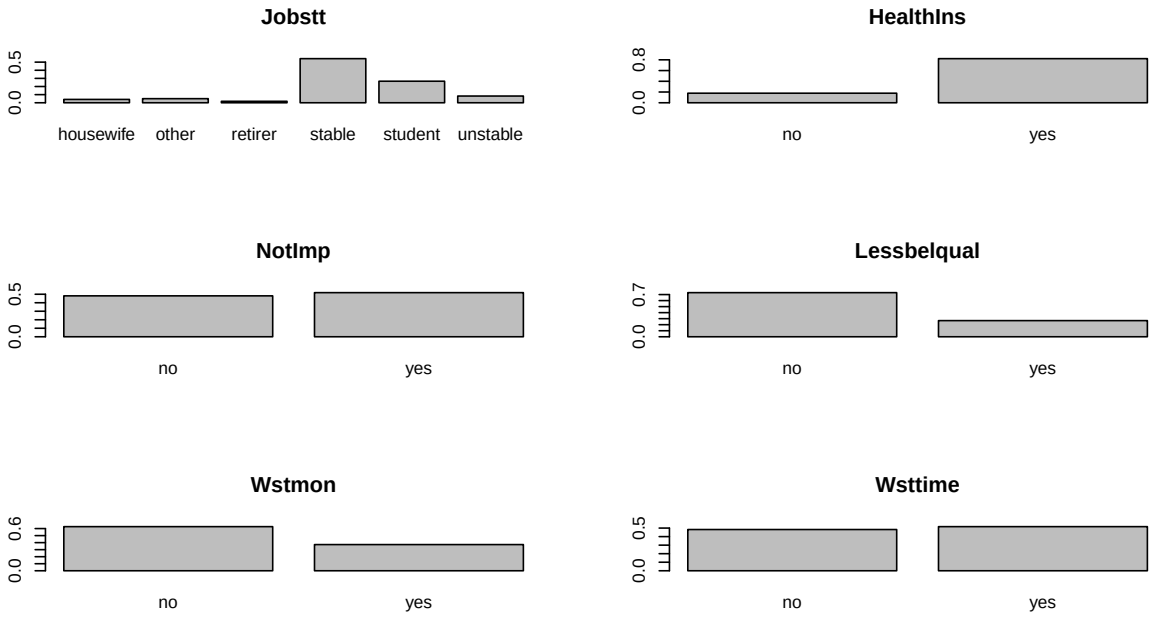

Figure 2: Bar Plots of Categorical Variables

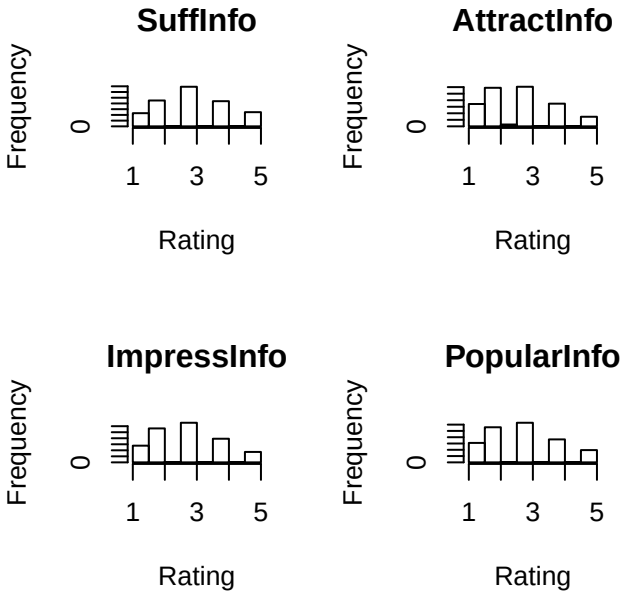

Figure 3: Histogram of Quality of Information Variables

Initial Modeling and Diagnostics

(1)The first logistic model includes all variables on demographics, as well as all variables on the value and quality of medical service. (2)The second model was fitted using a stepwise selection procedure to remove variables that do not contribute much in predicting the outcome. The result of running this procedure was a model with Jobstt, Wsttime, NotImp and SuitFreq as the explanatory variables. (3)The third model was fit by adding HealthIns and variables regarding Quality of Information. Interaction terms between HealthIns and the other variables were also included to see whether having HealthIns affected the quality of information ratings.

(4)A goodness of fit test for model3 was conducted by subtracting the log-likelihood of model3 from the log-likelihood of a saturated model. The null hypothesis for this test was model3 is correct, compared to the alternative hypothesis that the saturated model is correct. A p-value of 0.99998 tells us that there is no reason to reject the null hypothesis. (5)Figure 4 shows the calibration plot for model3. Three methods were used to get a good idea of how close the estimated probabilities of model3 were to the fractions of Y=1 cases. The kernel and spline methods suggest that model3 is indeed well calibrated. However, the volatility of the knn method makes it difficult to confirm calibration with certainty. Also, for all 3 calibration methods, we see that accuracy falls off near the extremes, when our predicted probability is close to 0 or 1. To improve performance in these areas, we could consider adding explanatory variables or use an additive model that accounts for the loss of accuracy in these areas.

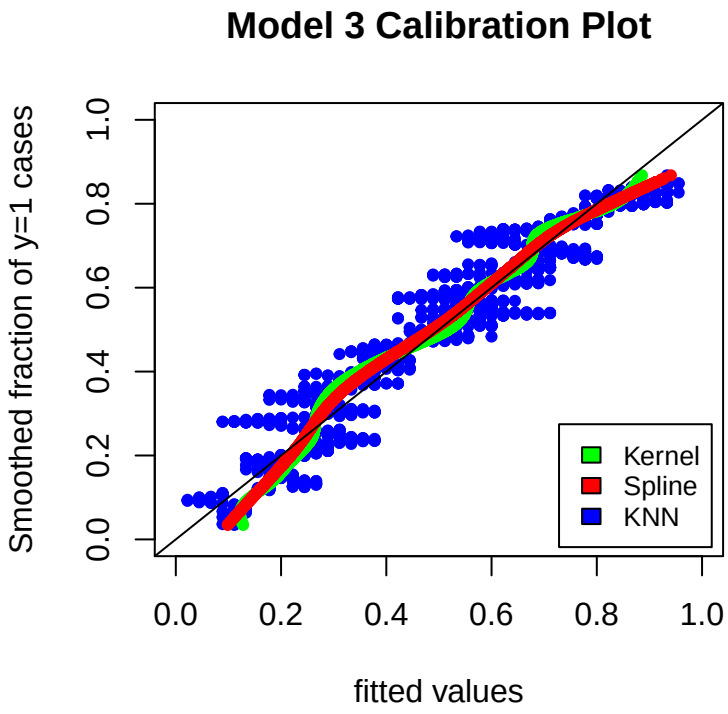

Figure 4: Calibration Plot for Model3

Model Inference and Results

Table 1: Anova Test Between Model3 and Model without HealthIns Interaction

Resid..Df	Resid..Dev	Df	Deviance	Pr..Chi.
2048	2403.575	NA	NA	NA
2052	2404.857	-4	-1.281853	0.8644469

We are now interested in interpreting the interaction terms of Model3. (1)Holding other variables constant and assigning 'yes' for HealthIns tells us that people who have HealthIns are 0.66 times more likely to have received a check-up than those who did not have HealthIns. Then, the model with the interaction terms was compared against a model without interaction terms. (2)As shown in table 1, doing an anova chi-squared test between the two models concluded in an insignificant p-value of 0.864, which tells us that interaction terms between HealthIns and the quality of information variables were not necessary, and the null-hypothesis of the model without interactions were sufficient in predicting whether a respondent had an exam in the past 12 months. This reduced model will be used from this point on.

$$logit(p_{HadExam}) = B_0 + B_1 Jobstt + B_2 Wsttime + B_3 NotImp + B_4 SuitFreq$$
$$+B_5 HealthIns + B_6 SuffInfo + B_7 AttractInfo + B_8 ImpressInfo + B_9 PopularInfo$$

To make the result interpretable for the Assistant Minister, the ratio between the odds of having a checkup for people with the most belief in the quality of information and the least belief in the quality of information was calculated. The result is shown below in table 2. (3)The ratio was 1.389, which means people that rated 5 for every quality of information question was 1.389 times more likely to have had an exam in the past year than a person that rated 1 for every quality of information question. We wanted to provide a confidence interval for this ratio. (4) The 95% confidence interval for the ratio is approximately between 0.189 and 10.214. This means that a person that rated 5 for all quality of information questions was more likely to have had a check-up, with a likelihood between 0.189 and 10.214. This is a strangely wide interval, that begs the question, is rating and HadExam even related? Whether a person had insurance was not considered because we concluded that hadInsurance did not significantly interact with the other variables.

Table 2: 95% CI For Ratio of a Subject that Gave Highest Rating vs Lowest Rating

Point.Estimate	Lower	Upper
1.38947	0.1889918	10.21542

Conclusions

(1)Figure 2 in the EDA section shows people's feelings towards the value and quality of medical service. Wsttime and NotImp had a no/yes ratio of approximately 48:52. WstMon and Lessbelqual had a better ratio, with many more people saying that they did not feel check-ups were a waste of money or that they had little faith in the quality. In these aspects, it would be recommended that the Ministry of Health design their marketing campaign to highlight the importance of check-ups, and that it is not a waste of time. Figure 3 in the EDA section shows people's feelings towards the quality of information they receive in check-ups. Although SuffInfo had an approximately normal distribution, the other three variables, AttractInfo, ImpressInfo and PopularInfo had a right-skewed distribution that shows in general, people rated the quality of information from check-ups poorly. In addition to the marketing campaign on importance, the Ministry of Health should provide an overall better quality information to patients that come to check-ups.

To assess the effect of people's feelings towards the quality of information on whether a person is likely to get a check-up, a 95% confidence interval was constructed. The result of this assessment was an interval between 0.189 and 10.215, with a point-estimate of 1.389. Although the point-estimate is greater than 1, telling us that people who rate the quality of information higher is more likely to have had a check-up, the wide confidence interval covers numbers below 1, which says that people with high ratings for quality of information are less likely to have had a check-up than those with low ratings. For this reason, it is difficult to suggest that quality of information is an important predictor for the response variable, HadExam. To answer the question about whether health insurance had interactions with the quality of information variables, a comparison of two models, one with interaction and one without, was performed, and concluded that the model with interaction did not perform better than the model without interaction. Therefore, having health insurance did not interact with quality of information variables significantly.

(2)The analysis seems to say that people are not getting check-ups because they do not

think it is important, and they believe it is a waste of time. This could be related to how people responded in the quality of information variables, which has a generally right-skewed distribution. People may have had a previously bad experience in check-ups, and receiving unimpressive, unattractive or unpopular information may have people think that future check-ups would be unnecessary. (3)There are a few aspects that could have made a better analysis. There could be confounding variables that weren't collected, such as the education level of a person, or financial situation that could have helped in predicting whether people get check-ups. Collecting such variables may help in the calibration of the model that we fit. As mentioned before, our model seemed to be calibrated fairly well, except near the extremes, and better data collection may be a possible solution.