Vietnam Check-Up Analysis

Trevor Moreci (tmoreci)

5/5/2023

Introduction

(1) Medical expenses for serious diseases can be very expensive, especially in countries without socialized medical systems, such as Vietnam. One way to detect issues early and prevent serious problems is to have patients get regular check-ups. There are many obstacles to getting regular check-ups, and Vietnam in particular has seen a relatively low number of patients getting regular check-ups. This analysis is prompted by the Vietnamese Ministry of Health's desire to increase the number of people who get annual health exams and discover the root causes of why people aren't getting them currently.

In order to conduct this analysis, we focused on three main questions:

1. How do people rate the value and quality of medical service, and the quality of information they receive in check-ups?
2. What factors appear to make a person less likely to get a check-up every twelve months?
3. Is the quality of information patients received during check-ups an important predictor of whether patients get check-ups, and does this differ between people with and without health insurance?

To answer these questions, we analysed a dataset of survey responses from Vietnam where respondents answered an array of questions regarding their personal information and how they feel about regular health check-ups.

(2) After conducting the analysis, we found that quality of information patients receive is an important predictor of whether or not patients get check-ups, specifically how attractive they thought the information was. Other important factors include whether or not the respondent had health insurance, the job description of the respondent, and whether or not

the respondent thought getting a check-up is a waste of time or money. In general, we found that people rated the value and quality of medical service as pretty average and didn't often give high marks.

Exploratory Data Analysis

(1) Before creating any models to help answer the questions in this analysis, we will first do some exploratory data analysis on the variables in our dataset. We will start by looking at our continuous variables. These variables are the age, height, weight, and BMI of the survey participants. As we can see from the histograms of these variables in Figure 1 and Figure 2, they all have varying degrees of right skews. Height is the closest to being normally distributed, while Age has the largest right skew. None of these skews are overly egregious though, and we should be able to work with the variables without transformations. Next, we have our variables where responses are a ranking from 1 to 5 by the participants. These variables include quality of tangibles or medical equipment at the check-up, the perceived empathy of the medical staff, the sufficiency of the information provided, the attractiveness of the information provided, the impressiveness of the information provided, and the popularity of the information provided. While these variables are scale based and could be categorical, we are treating them as continuous since many answers weren't whole numbers. All of these variables had a mode of three besides empathy, which had a mode of 5.

Next, we will look at Figure 3 and Figure 4, which show the distributions of the categorical variables in the dataset. There are a couple of things that stand out from these barplots. The first is that the majority of people that participated in this study were women. Furthermore, a large majority of people that participated had health insurance and were either a student or had a stable job. Next, looking at the value of information variables, we can see that the majority of participants thought check-ups were a waste of time, but on the other hand the majority of participants did not think they were a waste of money. Furthermore, most participants reported having some faith in the quality of medical treatment, but a slight majority of responses found check-ups to be not important. Finally, the most common response for how often check-ups should be done was once every six months.

(2) Next we will look at the distribution of our response variable HadExam in Figure 5. This is a binary variable that denotes whether or not a person has had a checkup in the last 12 months. As we can see from the bar chart, about 51% of the people surveyed have had a check-up in the last twelve months.

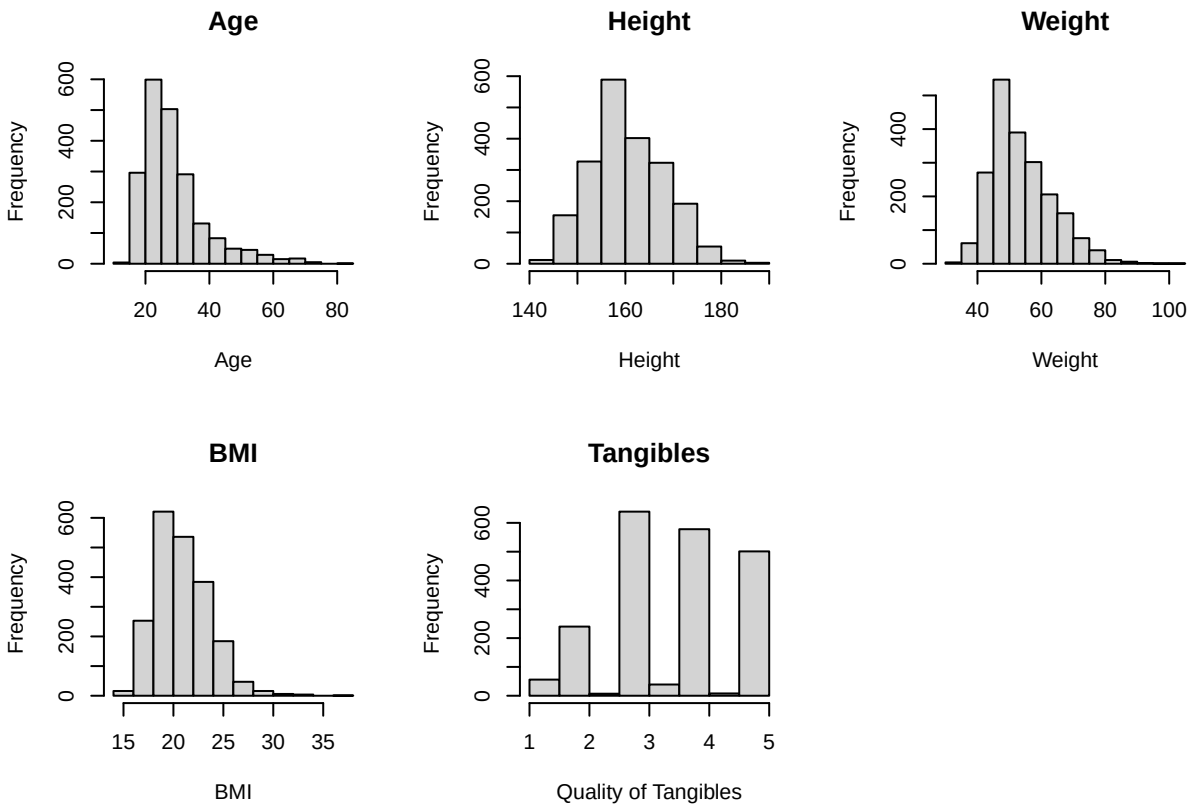

Figure 1: Histograms of Continuous Variables

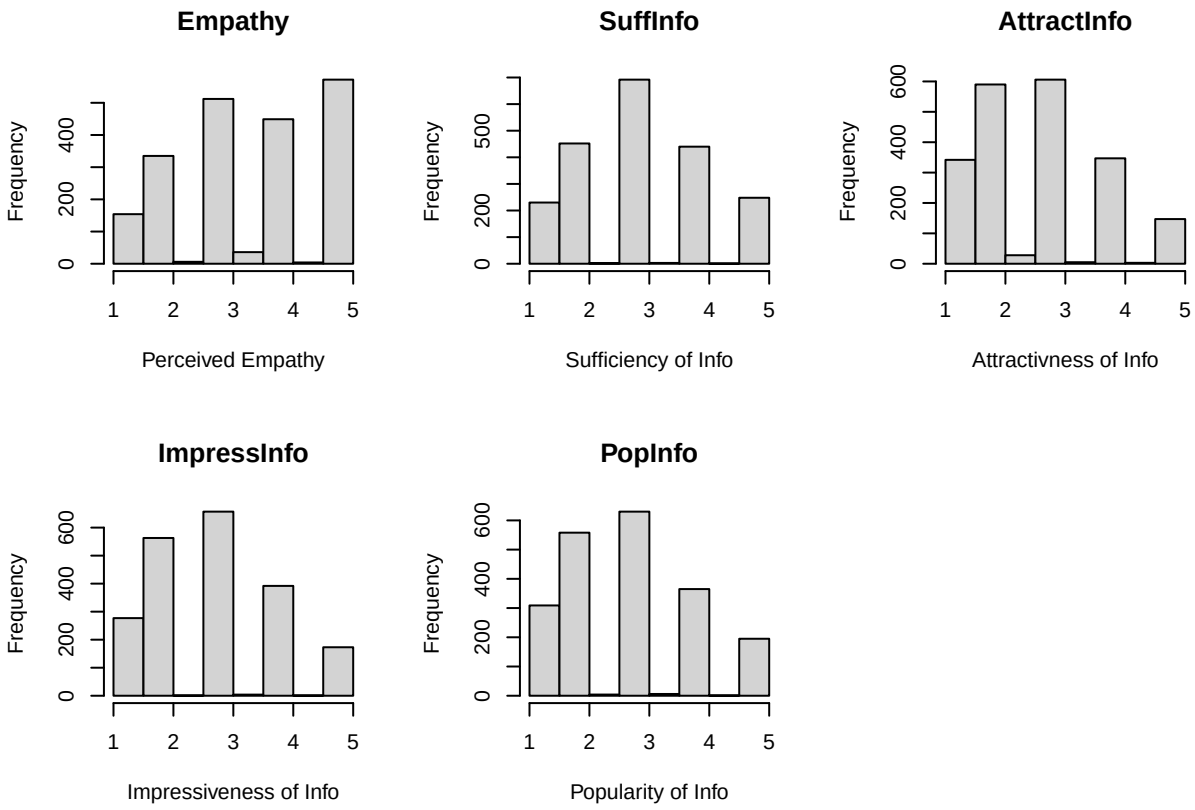

Figure 2: Histograms of Continuous Variables

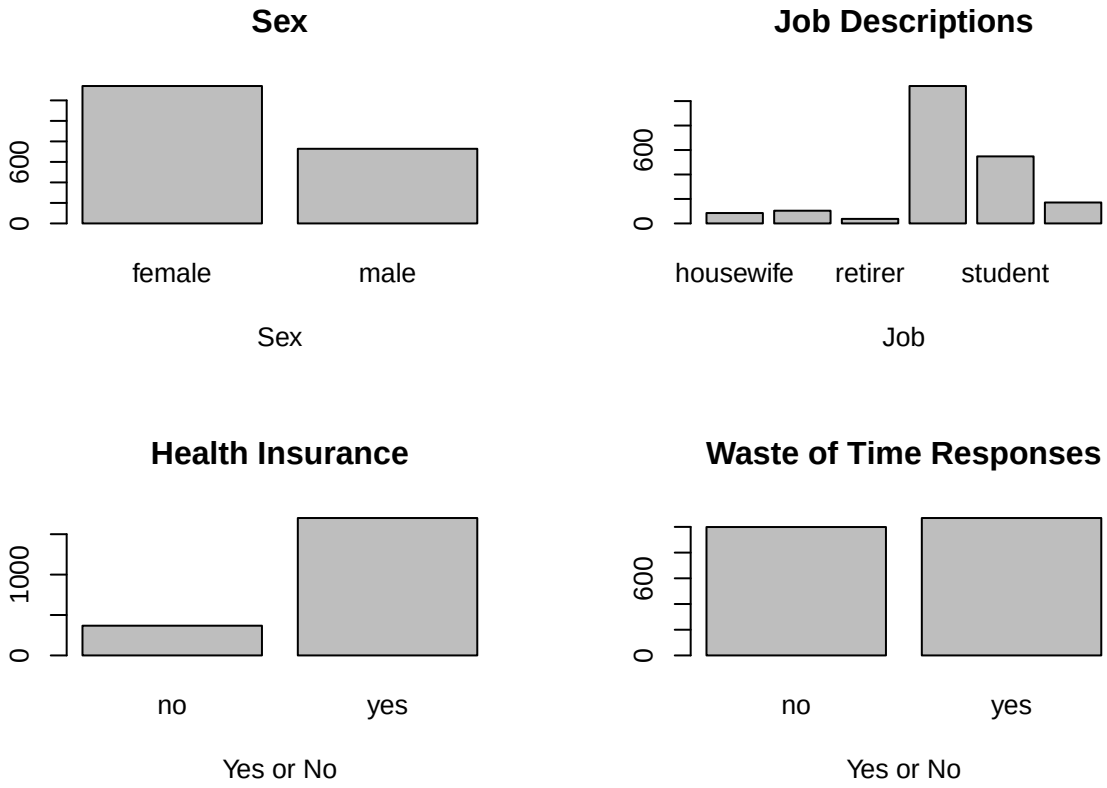

Figure 3: Barplots of Categorical Variables

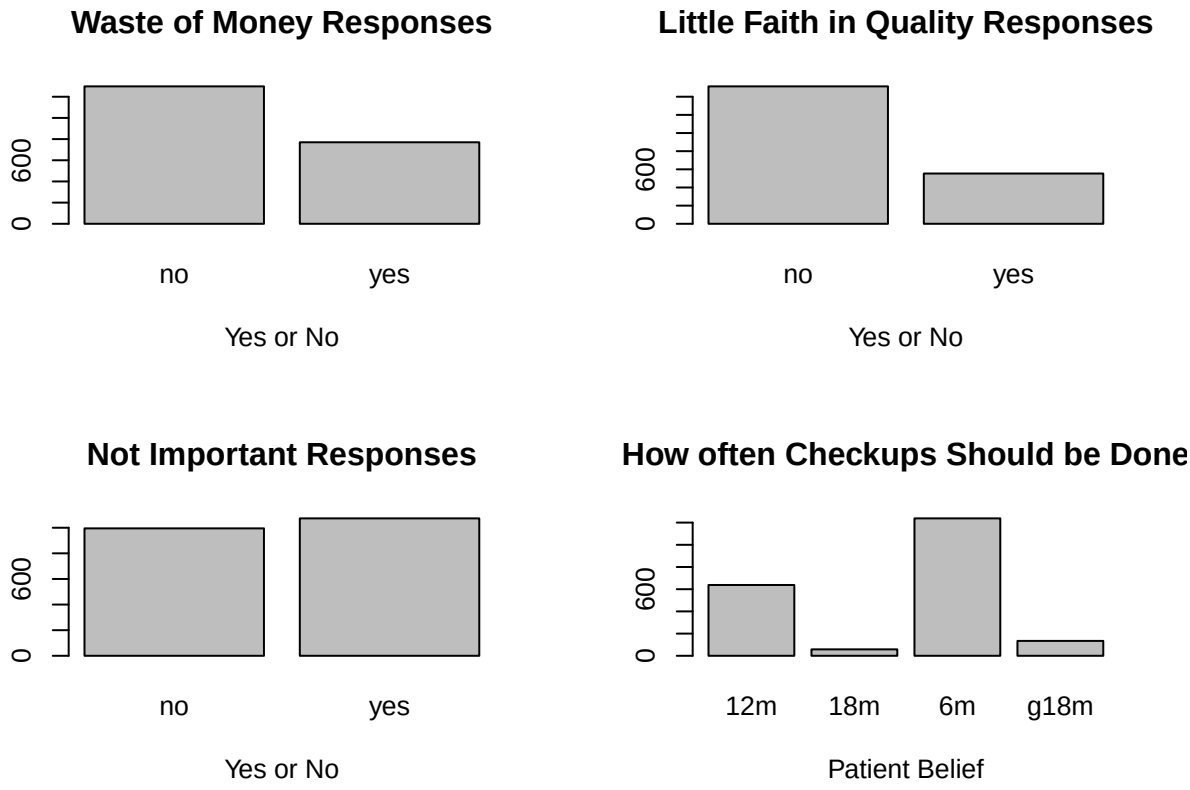

Figure 4: Barplots of Categorical Variables

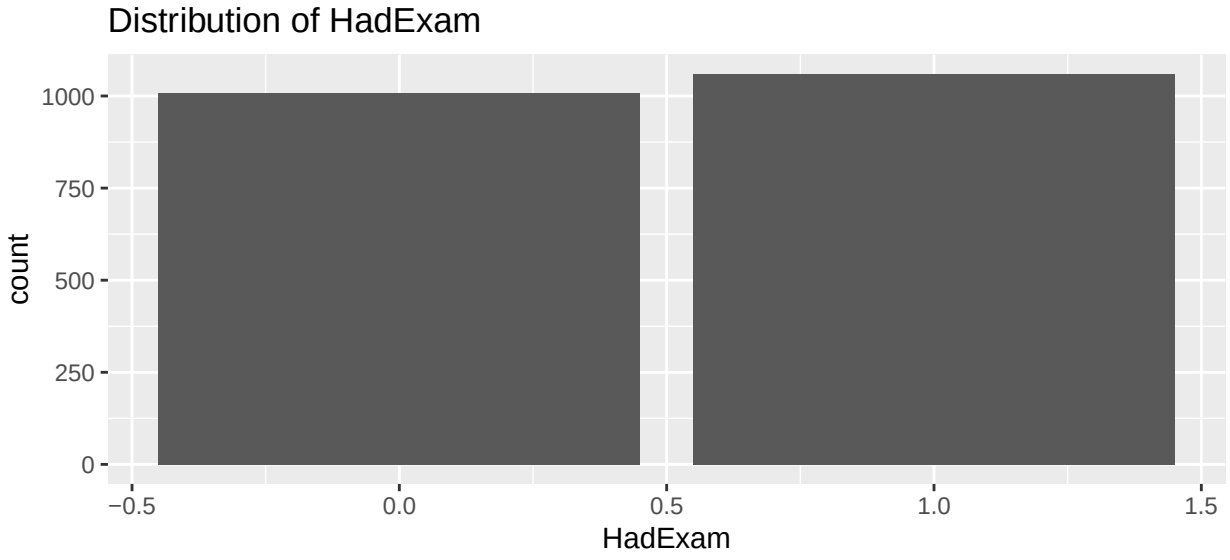

Figure 5: Barplot of Response Variable

(3) Next, we will look at two plots that summarize how people rate the value and quality of medical service, and the quality of information they receive in check-ups. The first plot in Figure 6 shows the mean responses for the various quality of information surveys. Since they were on a scale of 1-5, one can assume that a score of 3 would denote average quality. Based on the plot, one can see that the mean responses for perceived quality of tangible medical equipment and perceived empathy of staff were above average. Furthermore, the mean response for perceived sufficiency of information received was just about average. On the other hand the mean scores for perceived attractiveness of information, impressiveness of information, and popularity of information were all slightly below average.

Figure 7 shows the proportion of "yes" answers to the surveys regarding the value of medical service. Two things stick out about this plot. The first is that the slight majority of respondents found that regular check-ups were not important or a waste of time. The second is that the majority of respondents did not think that medical service was a waste of money and did not have little faith in the quality of medical service.

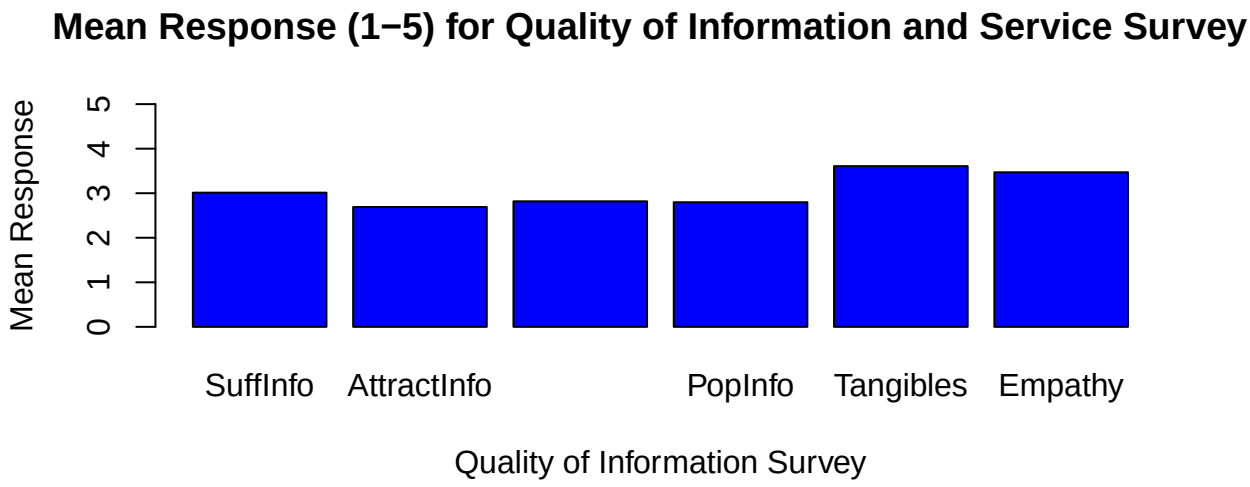

Figure 6: Comparison of Mean Responses for Different Surveys

(4) Based on this EDA, there are a couple of interesting results that might suggest future findings in our analysis. These include people thinking that checkups are a waste of time and not important. Furthermore, based on Figure 4 we can see that many people weren't particularly impressed or didn't like the information they receive. One can guess that the people who answered this way might be less likely to have gotten a check-up in the past year.

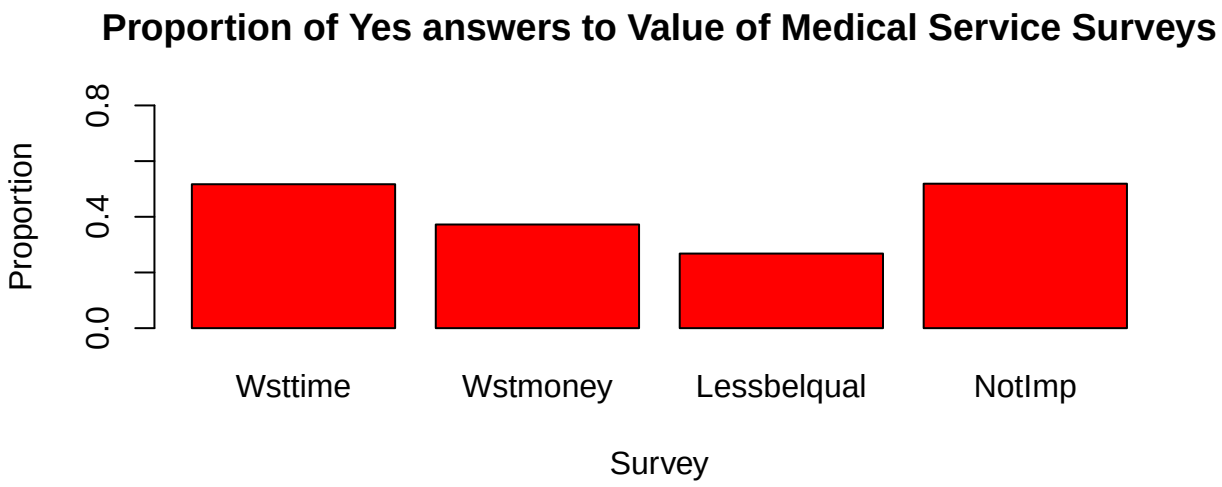

Figure 7: Comparison of Yes Answers to Medical Service Surveys

Initial Modeling and Diagnostics

(1) Building off of this EDA, we will start our initial modeling by building a generalized linear model that predicts our binary HadExam variable as the response for all of the demographic variables and the variables regarding the value and quality of medical service. We will call this Model 1. One thing to note is that we are not including the health insurance variable or the variables about the quality of information presented in the check-ups due to the fact that we will be testing the significance of these variables in later models.

(2) Even though our Model 1 doesn't use those specific variables, it still uses thirteen predictor variables. It is reasonable to believe that not all of these variables are needed to help predict the outcome. Therefore, we will create a second model, Model 2, that uses a backward stepwise selection procedure to remove variables that don't help predict the outcome by minimizing the model's AIC error estimate. After completing this backwards selection, we are left with a second model that contains only four predictor variables. These variables are Jobstt, Wsttime, NotImp, and SuitFreq. It makes sense that the Wsttime and NotImp variables would be significant based on our EDA. Furthermore, including Jobstt also makes sense because whether or not people have a job could greatly impact they're ability to get doctor's check ups. SuitFreq makes sense as a predictor as well because if people believe that one needs a checkup less often than once a year, it would make sense if they didn't get one in the last year.

(3) Next, we will build off this second model and make Model 3. Model 3 is the same model

but adding the health insurance and quality of information variables. Furthermore, the interactions between health insurance and these variables are also added to the model so that we can later check whether or not these variables have different associations between patients with and without health insurance.

(4) Next we will test the goodness of fit of Model 3 by conducting a chi-squared test to see whether or not there is a significant difference between Model 2 and Model 3. The null hypothesis of this chi-squared test is that there is not a significant difference between the two models, while the alternative hypothesis is that there is a significant difference between the two models. This test resulted in a deviance score of 41.87 on 9 degrees of freedom, which corresponds to a p-value below the 0.05 significance level. Therefore we can reject the null hypothesis and say that there is a significant difference between the two models and the added variables in Model 3 help better fit the data.

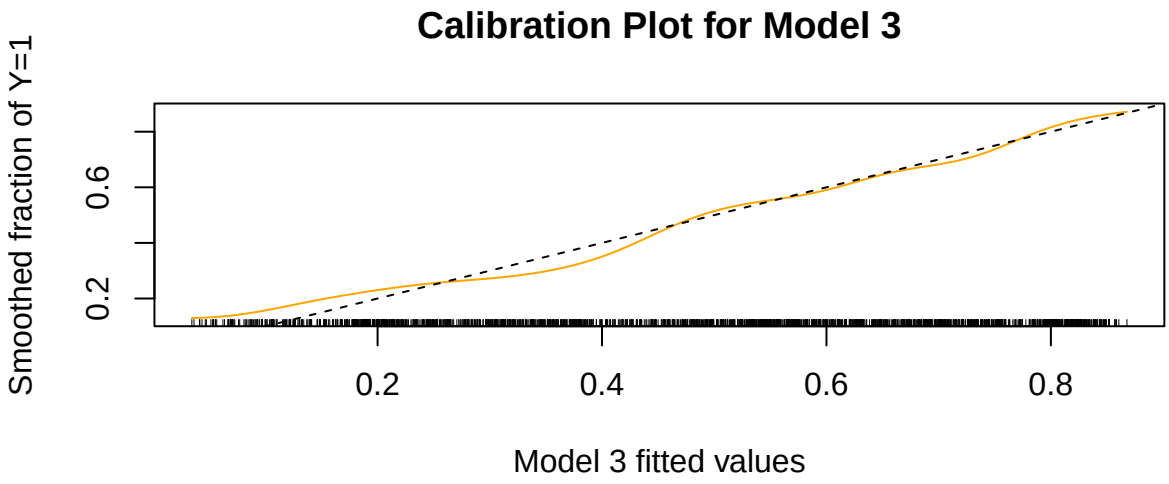

Figure 8: Calibration Plot based on Kernel Smoother

(5) Next, we will check whether or not Model 3 is well-calibrated. In this sense, well calibrated means that its predicted probabilities match the observed outcomes of proportions in the data. In order to do this, we will fit a kernel regression to smooth the HadExam variables with the fitted values from Model 3 as the predictor. This smoother will approximately give us the correspondence between our predicted probabilities and the true rate of Y=1. We will then plot the fitted values for this smoother against our fitted values from Model 3. If the model is well calibrated, the plot should look like y=x. The plot is shown in figure 8. While the fit isn't perfect, it is still pretty good. The line generally follows the same trend as y=x, although it is wavy and has small peaks and valleys at certain points. One potential

reason for these peaks and valleys is that some of the added variables and interactions may not be significant or warranted in the model.

Model Inference and Results

(1) To start our Model Inference we will look at the summary and coefficients for Model 3. Specifically, we will look at the interaction terms between health insurance and the quality of information variables. First, the estimate for the coefficient of the interaction term between HealthIns and SuffInfo is positive at 0.173. This means that when someone has health insurance, their log odds of having had a health exam within the last year increases by 0.173 every time their score on the sufficiency of information survey increases by 1. On the other hand, the coefficients for AttractInfo, ImpressInfo, and PopularInfo are all negative respectfully at -0.009, -.044, and -0.075. Therefore, when people have health insurance, the log odds of them having had an exam in the past year decreases by these coefficients with one unit increase on their responses to these surveys. One thing to note is that all of these interaction terms have very high p-values. This suggests that they might not actually be significant predictors in this model.

(2) In order to fully test whether these interaction terms are necessary, we will conduct an Analysis of Deviance Chi-Squared test on Model 3 and the same model without the interactions. Our null hypothesis for this test is that there is no significant difference between the two models. Our alternative hypothesis is that there is a significant difference between the two models. The test resulted in a deviance score of 1.2818 on 4 degrees of freedom, which corresponds to a p-value of 0.8644. Since this p-value is above the 0.05 significance level, we cannot reject the null hypothesis and we can conclude that there is no significant difference between the two models. Therefore we will move forward with the simpler model. This finding means that while the quality of information a patient receives could still be a good predictor on whether or not they had an exam, the responses on these surveys do not differ significantly between people with and without health insurance.

After completing another stepwise selection on this fourth model, we ended up with our final model. This model contains 5 predictors: Job description of the respondent, whether or not they thought health exams were a waste of time, whether or not they thought health exams were important, the frequency with which they thought exams should be administered, whether or not they had health insurance, and how they ranked the attractiveness of the information they received. Interestingly enough, the only variable concerning the quality of information in this model is AttractInfo. **(3)**Using the model's coefficient for this variable,

we find that the ratio between odds of having a checkup for people with the most belief in the quality of information (rating 5) and the odds for those with the least belief in the quality of information (rating 1) is 1.47. **(4)** A 95% confidence interval for this ratio is (1.045, 2.07). This means that the rate of people who have gotten an exam is 4.5% to 100% higher for people who ranked the attractiveness of information as a five compared to a one.

Conclusions

(1) Overall, our analysis yielded a couple of main findings. The first is that most people rated the quality and value of information they received around average (3 out of 5). Furthermore, over 50% of participants found that health checkups were not important or a waste of time. These aren't the best results one might want to see because you would want your citizens to think highly of your healthcare system. Another main finding was that the most important predictors of whether or not someone got a check up were whether or not they thought health exams were a waste of time, whether or not they thought health exams were important, the frequency with which they thought exams should be administered, whether or not they had health insurance, and how they ranked the attractiveness of the information they received. Of these predictors, only one of them is a response regarding the quality of information that patients receive. Furthermore, these responses didn't depend on whether or not the respondent had health insurance. From these results, the Assistant Minister of Health could focus on these 5 areas for their advertising. Specifically, they could work on advertising how attractive the information one could receive is and also push people to get health insurance.

(2) There are a couple possible reasons for why these variables would be the most significant.First, if people thought that check-ups were unimportant or a waste of time, it would make sense that they aren't getting them. Furthermore, including job description also makes sense because whether or not people have a job could greatly impact they're ability to get doctor's check ups. The response about frequency of check-ups makes sense as a predictor as well because if people believe that one needs a checkup less often than once a year, it would make sense if they didn't get one in the last year. As for health insurance, one would believe that people who have health insurance might put more effort into making doctors appointments. Finally, the attractiveness of information could matter because people want to think that they information they will receive is valuable to them.

(3) There were a couple different limitations in this analysis. One limitation is in the way that the survey was conducted. In general, many of the questions were framed in a leading manner that could lead to skewed results. For example, asking respondents if they thought

their checkups were a waste of time or a waste of money could lead to different results as compared to framing the questions in a more objective manner. Another limitation is that all of the models that we looked at, including the final model, included variables with high p-values. Therefore the models used in the analysis might not fit the data in the best possible way. Finally, our final model establishes a relationship between the predictors and the response, but it does not establish cause and effect. This could be an issue since the Minister of Health was looking for causal relationships.