36-402 DA Exam 2

Sarah Li (sarahli)

5/5/2023

Introduction

(1) Access to medical care for serious diseases can be financially burdensome, especially in countries with no socialized medical systems. It would be ideal for patients to receive check-ups (GHEs) regularly in order to detect these diseases and get treated earlier. This approach would enhance public health and outcomes and lower costs. However, not a lot of patients receive check-ups regularly for a variety of reasons, potentially due to cost, inconvenience, lack of trust in doctors, or negative prior experiences. Public health researchers from Vietnam thus conducted interviews in Hanoi, Hung Yen, Vietnam, by traveling to schools, hospitals, companies, and government agencies in order to find out why people avoid medical checkups. In this report, we use the dataset of the resulting 2068 responses to address the following research questions:

1) On average, how do people rate the value and quality of medical services or information they receive in check-ups?

2) What factors appear to make a person less likely to get a check-up annually?

3) Does evidence suggest that this is an important predictor in getting check-ups, and does this have to do with having health insurance?

(2) After modeling and diagnostics, we observe people rate the quality of health check-up information on a scale of 1 to 5 and notice a majority of people found the quality of medical services and checkup information to be average or below average. We also found that job status, beliefs on the importance and effectiveness of health exams, and quality of information were associated with them getting a checkup in the last year. We also find

evidence suggesting that health insurance in regards to higher ratings of quality variables does not play an important role in predicting whether a respondent received an exam.

Exploratory Data Analysis

(1) The key variables we will explore in our dataset include `Jobstt` (job status), `HealthIns` (has health insurance), `Wsstime` and `Wstmon` (believes check-ups are a waste of time or money), `NotImp` (believes check-ups aren't important), `SuitFreq` (how often check-ups should be done in months), `SuffInfo`, `AttractInfo`, `ImpressInfo`, and `PopularInfo` (ratings on a scale of 1-5 in how a respondent views the information quality they receive in checkups). **(2)** Our response variable is `HadExam` (which indicates whether any of the factors played in role in having a health exam). We decide to treat Had Exam, Job Status, Health Insurance, Waste time, Waste money, Not important, Suit frequency as categorical variables. We will treat the quality of information variables as numerical even though they are ordinal categorical. (Note: all yes or no variables have been converted to 1 or 0).

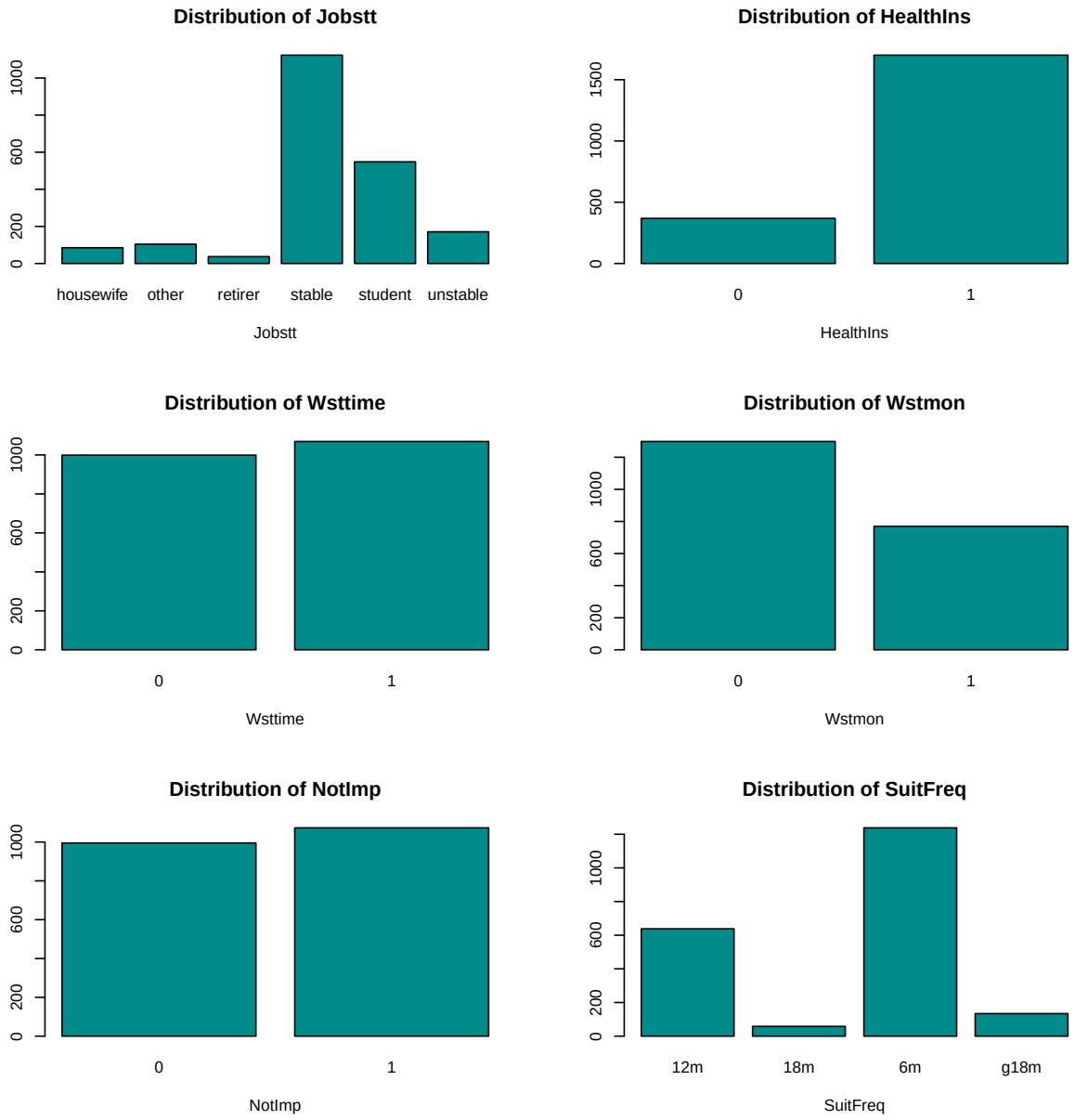

Figure 1: Univariate EDA on Categorical Variables

We see from the Figure 1 that most respondents had stable or student jobs, most had health insurance, around the same number of people thought check-ups were and were not a waste of time, most people actually did not think they were a waste of money, an even amount thought they were and weren't important, and most believe check-ups should be done every 6 months.

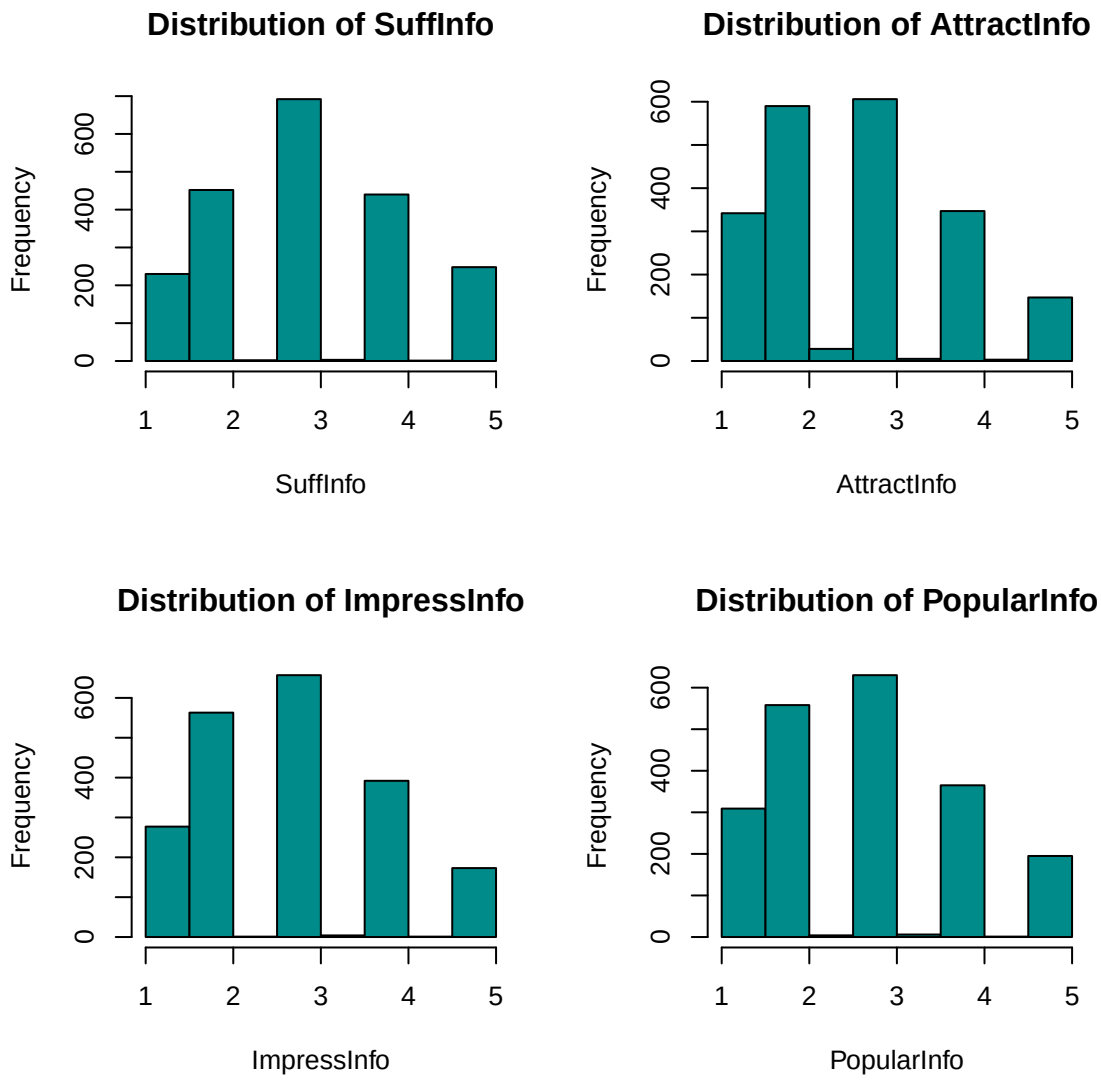

Figure 2: Univariate EDA on Numerical Variables

From Figure 2, we observe for the quality of information variables, that most people found check-up information having average and below average quality consistently, with the majority of ratings being 3 or lower and all the distributions being slightly right skewed.

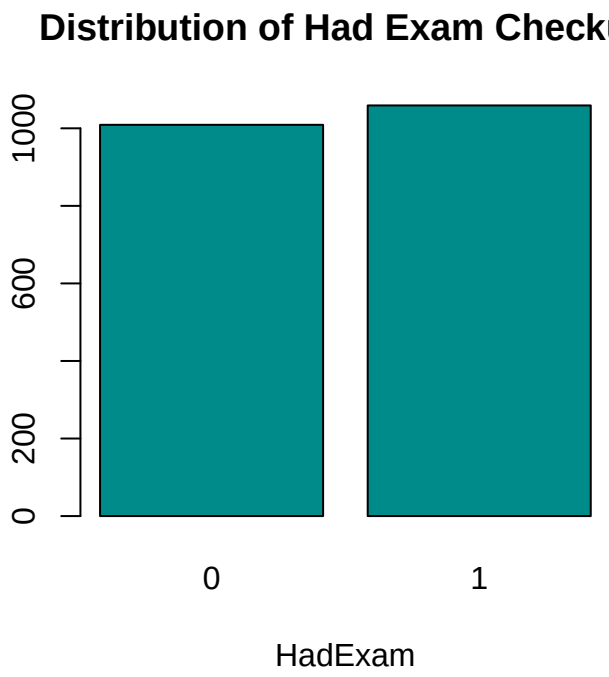

Figure 3: Distribution of Response Variable HadExam

(2) We observe in our response variable HadExam, 1059 respondents (a little over half) reported having a check-up within the last year, making the distribution roughly even and binomial.

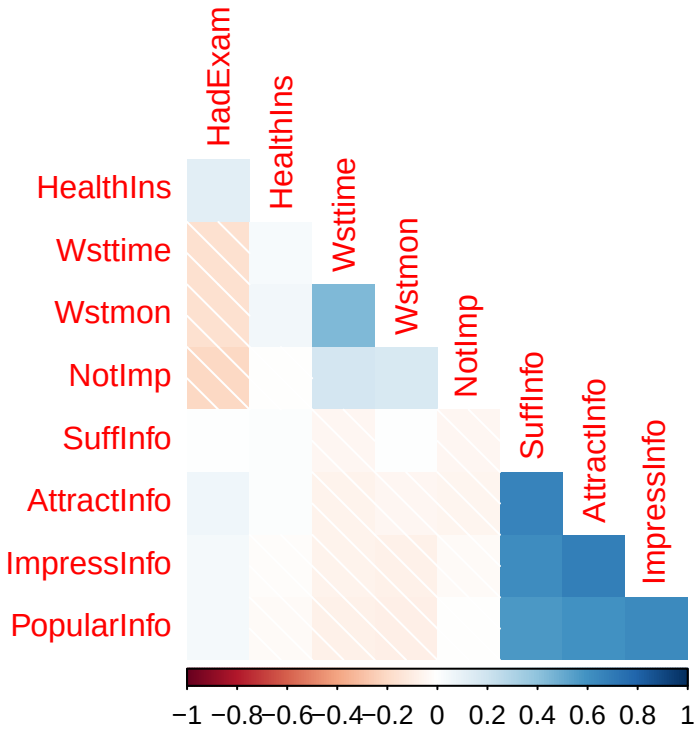

Figure 4: Bivariate Correpation Plot of Numerical and Binary variables

(3) From bivariate EDA in Figure 4 on the variables of interest in the correlation plot, we observe most positive correlation among the quality of information variables. It seems that respondents had similar sentiment regarding how sufficient, attractive, popular, or impressive their check-up information were. In general, the quality of medical service also shows to have positive correlation, with those who believe check-ups were a waste of time also roughly believe they were a waste of money and not important.

(3) cont We observe in the next two Figures 5 and 6 that the distribution of ratings for quality of medical serve, that the average sentiment for non-importance and waste of time was neutral with even distributions; however, even with this in mind, more people still believed check-ups were not a waste of money than worth the pay. As for the quality of information, the attractiveness, impressiveness, popularity, and sufficiency all followed roughly the same marginal rating proportions. Overall, most respondents believed that the quality of information was average and not the best (indicated by proportions of <=3 ratings lying above the 0.5 line).

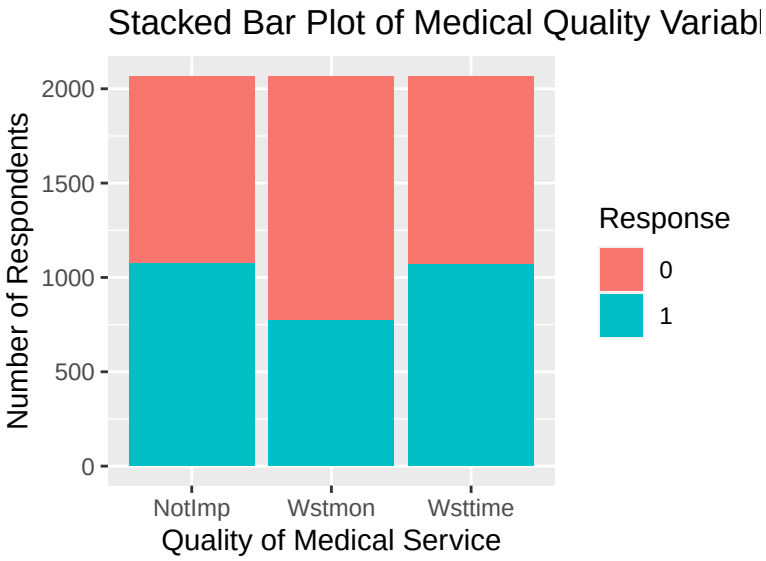

Figure 5: Stacked Barplot of Medical Quality Variable Ratings

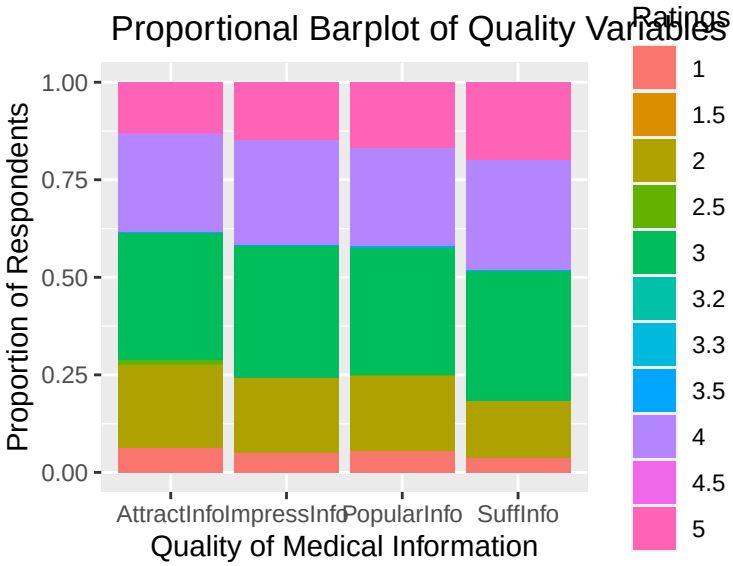

Figure 6: Proportional Barplot of Information Quality Variable Ratings

(4) Last for EDA, we want to address some of the findings we may consider while modeling. First we note the continuous scale of ratings that included decimals, which is why we treated the ordinal values as continuous. Additionally, the bivariate comparisons seem to suggest that there may be correlation among the variables regarding quality of medical services and quality of check-up information among themselves. We will do our best to keep this caution in mind when undergoing the assumptions of our modeling. Another important note concerning our response variable HadExam is that, from the correlation plot, we observe no significant correlation between itself and any other predictor, which we did not expect. Further analysis would have to be done in modeling to see what kind of relationship holds for predicting our response.

Initial Modeling and Diagnostics

(1) After initial data analysis, we move on to modeling and formulate a generalized linear model (call Model 1) that predicts HadExam as a function of all the demographic variables in the dataset and quality of medical service variables, excluding health insurance. (We note that this model includes variables we omitted in the EDA as they will be eliminated in the future.)

From the summary output of Model 1, we observe many high p-values for most of the predictors, indicating that they do not contribute much to the model. From the correlation matrix and vif (variance inflation factors) test, we observe evidence of collinearity among some predictors, with height, weight, and BMI having high inflation values over 5. Again, these demograhic variables were omitted during EDA but still contained within the dataset. As for model diagnostics, since this is a logistic regression model, the residuals do not have to meet the normality assumption. Instead, we note that the assumptions met are that the variables are independent but one that is not met is multicollinearity. Thus, our initial model is not the best and we proceed to stepwise elimination.

(2) We perform stepwise regression using the backward direction until the AIC error is minimized. The final model (call Model 2) decides to drop variables Age, Sex, Height, Weight, BMI, Wstmon, Lessbelqual, Tangibles, and Empathy. The resulting variables that contribute most to model significance with low p-values include retired, stable, or student job statuses (Jobstt), belief that check-ups are a waste of time (Wsttime), belief that check-ups are unimportant (NotImp), and the belief that checkups should be done less than equal to 18 months (SuitFreq). Using the vif test again, we observe no multicollinearity concerns among any of the predictors, making this model meet our logistic regression assumptions

given independent samples.

(3) Interested in seeing how Health Insurance may play a role in having an exam, we add this to the model and the quality of information variables with their interactions. From Model 3, we include the interaction between whether or not the respondent had health insurance and variables indicating the quality of information they received. From the vif test, we are not concerned much about multicollinearity of the newly added interaction terms of the model but there are many newly added terms with relatively high p-values in the summary output now.

(4) To check whether our model is a good fit of the data, we conduct a chi-squared goodness of fit test with a 2403.6 residual deviance on 2048 degrees of freedom.

H_0 : The model is correct. The Residual deviance follows X^2_{n-q}
H_A : The model is not a good fit.

Since the result of the GOF test is very small $pval = 6.6856e-08$, we reject the null hypothesis that the model is correct and conclude that the fit of Model 3 is not good. This may be due to the abundance of predictor variables in the model creating noise.

(5) Next, we check if Model is well calibrated in Figure 7. The proportions of respondents who had taken an exam within the last year equal the average prediction (0.5121), making the model trivially well-calibrated. To check how well-calibrated Model 3 is, we consider using a kernel smoother with a bandwidth of 0.05 to predict HadExam using the estimated probabilities and check how close the values fit the line y=x.

The kernel smoother seems to follow the line relative closely, making the model decently well-calibrated and its predicted probabilities matching the observed proportions of outcomes in the data. However, our model did not pass the goodness-of-fit test. Since the interaction model includes a lot of factors, we may consider running stepwise regression once again (AIC) to eliminate variables that don't hold much predictive power. The summary output also demonstrates many of the predictors having high p-values and relatively low contribution to the model. This would help remove noise and potential collinearity.

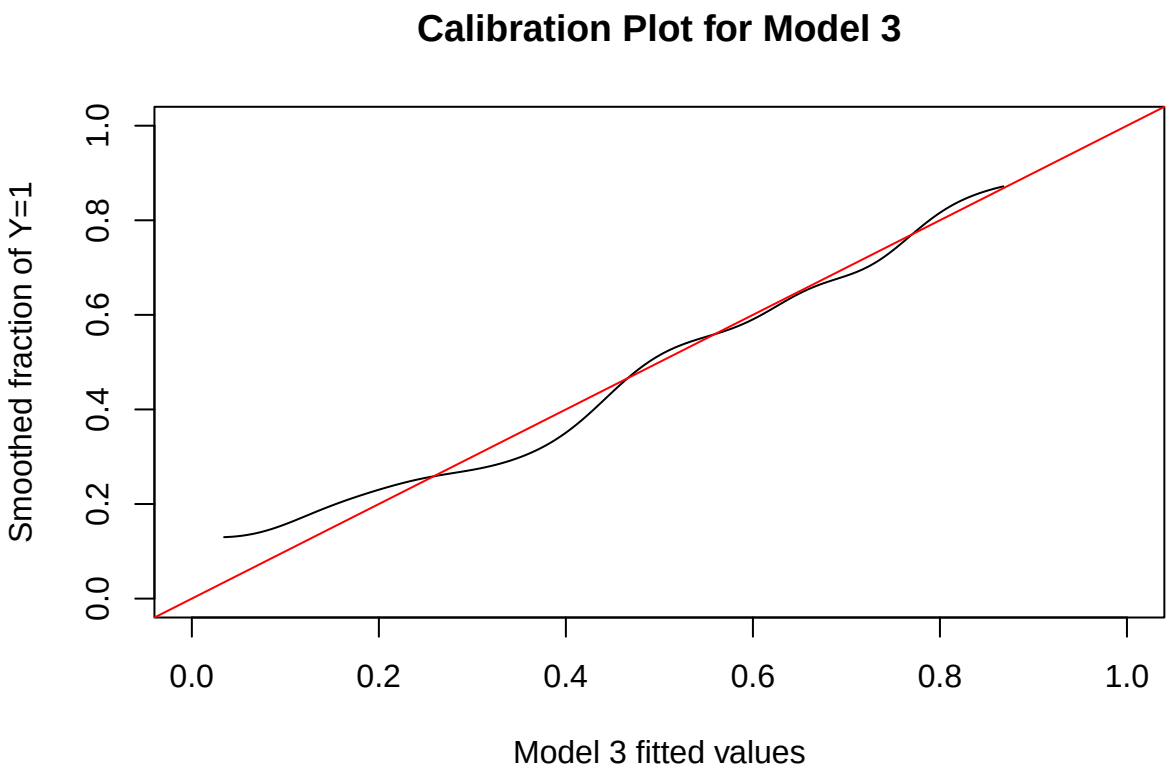

Figure 7: Calibration Plot of Model 3 with Interactions

Model Inference and Results

(1) When keeping other constants fixed, the respective multiplicative odds of rating SuffInfo, AttractInfo, ImpressInfo, or PopularInfo one unit higher for a person with health insurance would have been around 1.189, 0.991, 0.967, and 0.928 compared to a person without health insurance. These odds are all relatively close to 1, making the difference trivial. Additionally, we note that none of the added information quality variables and their interactions with health insurance hold much predictive power in the model, with all their p-values being large in the summary output. Thus, there is no notable difference regarding the increase of check-up information ratings for those who had and did not have health insurance.

(2) To confirm this, we conduct a chi-squared model selection test on Model 3 with and without interactions with the following hypotheses:

H_0: The reduced model is true (Model 3 without interactions)
H_A: The full model (Model 3 with interactions) is true and is a significant improvement on the reduced model

We observe a test statistic of $Dev = 1.2818$ and a p-value of $P(X^2 \geq 1.2818) = 0.8644$. Since the p-value is very large (>0.05), we fail to reject the null hypothesis and cannot conclude that the interactions between having health insurance and ratings of checkup information qualities made a significant difference or improvement on Model 3. That is, when observing increasing ratings of the information quality obtained from check-up, whether the respondent had or did not have health insurance does not play a role in determining whether he or she had taken an exam in the last year.

(3) To summarize this finding in a more high-level manner for the Assistant Minister, we calculate the ratio between the odds of having a checkup for people with most belief versus the least belief in the quality of information. Since this does not depend on a person's health insurance, we use the coefficients from the reduced model and report that, on average, respondents with most belief in information quality are 1.509 times more likely receive a check-up than those with least belief, holding other variables constant. **(4)** We are 95% confident that the true odds ratio of people with the most belief compared to people with the least belief in the quality of check-up information lies within the interval (0.9095, 2.5021)

Conclusions

(1) Even though our models did not indicate a relationship between Health Insurance and the belief in the quality of medical service or medical information having an effect on whether respondents took a health exam, we still discuss some notable findings. Particularly from EDA, we notice that most people had neutral or below average sentiment on the quality of medical services or information: on average, they believed that the sufficiency, attractiveness, impressiveness, and popularity of information received was very average or below average. They didn't particularly believe checkups were a waste of money however, just that that were not that valuable. Additionally, looking in closer detail at the coefficients of summary outputs, we observe in the reduced Model 3, that the coefficients for Waste of time, Not important, Sufficient information, Impressive information, Student/Unstable job status, and the belief that checkups should be done once every 18 months or greater were all negative. These variables contribute to how likely someone is to get an annual check-up. **(2)** These intuitively make sense as those who don't believe in or can afford checkups will not go (young people or not financially adequate people). Plus, those with beliefs in the quality of services or information will be influenced to go or not as well; many people do not trust the input of doctors and have differing beliefs about their own conditions. Health care can also get very expensive nowadays, which is why we might observe that unstable job status are correlated with lower odds of getting a checkup. Thus, we can reason how these factors played a role in lowering someone's willingness to get a health exam and why health insurance might not play a role in this.

(3) Lastly, we discuss the limitations of our findings and modeling. Our dataset contained many variables to begin with and thus could have a lot of added noise or hidden relationships among the variables (with potential of confounders). Our data is also not representative of how people generally view the healthcare industry and what may influence them to get an annual checkup as different regions have different policies and variations of healthcare. We also did not try interactions outside the questions or run stepwise regression multiple times in order to obtain a simpler and better model. Additionally, we did not plot every single linear relationship between the independent variables and their log odds in order to confirm our logistic regression assumptions; we will do that in the future. Since the Assistant Minister also asked a causual inference question related to having health insurance, we are unable to detect that relationship but can only provided an association among the quality variables and the response. We would have to conduct more studies and research to answer his question confidently.