36-402 DA Exam 2

Madhuri Raman (madhurir)

05/01/2020

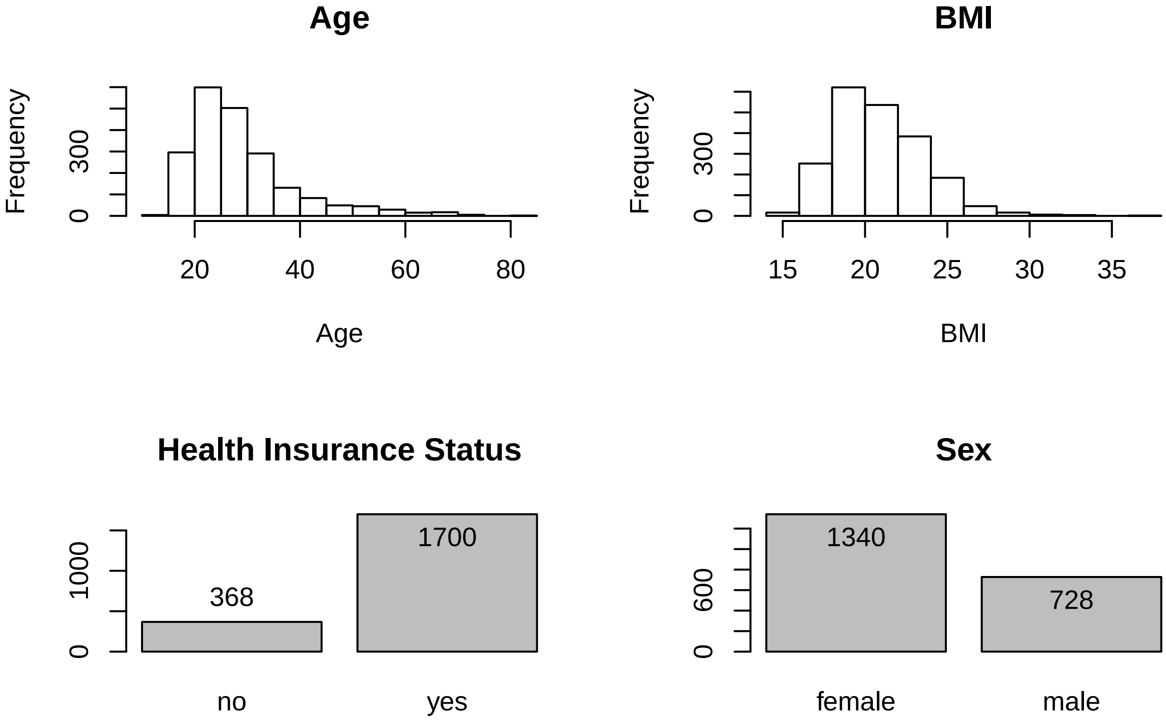
Introduction

(1) The purpose of this study is to investigate the potential factors that lead people in Vietnam to not sign up for an annual health exam. We are interested in three main questions. First, we want to understand how people overall rate the value and quality of medical service and the quality of information they receive in checkups. Based on this, the Vietnamese Ministry of Health can identify specific areas in which checkups may need to improve. Next, we want to identify the most important factors that appear to make a person less likely to get an annual check up, in order to help design advertising and public relations campaigns in a more targeted fashion. Finally, we want to focus on the quality of information received in checkups and understand if it an important factor contributing to whether or not patients sign up for a checkup. This may be different for people with health insurance and without health insurance, so we want to investigate if that is the case as well. These findings will help the Vietnamese Ministry of Health understand if they need to target their marketing campaigns around the quality of information received at checkups differently for different groups of people.

(2) We found that 50% of respondents feel that check-ups are a waste of time, but only one-third feel that they are a waste of money. Respondents also feel quite satisfied with the quality of equipment and personnel at the medical office. However, they are less satisfied what the hear, in terms of information, from the doctors themselves. We recommend that this is what a marketing campaign focus on. Respondents beliefs about the quality of information are not significantly different between people with and without health insurance; we do not see sufficient evidence that it would be worth targeting a marketing campaign to people based on their health insurance status. Finally, we found a few other important predictors of whether or not people get check-ups. These were: whether they thought it was a waste of time, their suitable frequency for check-ups, whether they thought a check up was important, their job status, and their health insurance status alone (not the interactions with quality of information).

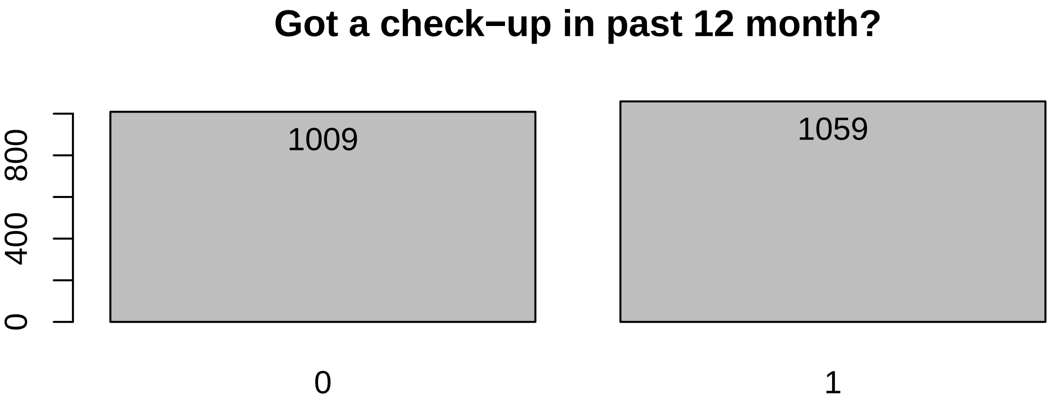
Exploratory Data Analysis

(1) To predict whether or not a respondent will get a check-up, we will use three main categories of predictor variables. These are demographic variables about the respondent, their beliefs about the value and quality of the medical service at check-ups, and their beliefs about the quality of information they receive at check-ups. We will exclude respondent id and place from the models because they are either 100% unique or nearly 100% the same factors for all respondents so they will not be useful. The continuous variables we will use are: age, height, weight, BMI, quality of tangibles score, empathy score, sufficiency of information score, attractiveness of information score, impressiveness of information score, and popularity of information score. The categorical variables we will use are: sex, job status (6 levels), health insurance status (0 or 1), waste time belief (0 or 1), waste money belief (0 or 1), little faith in quality of service (0 or 1), importance of check-ups (0 or 1), and suitable frequency for check-ups (4 levels). Below, we explore the univariate distributions of some demographic variables, continuous and categorical. Later, we will explore the value an d quality of medical service and quality of information variables.



We observe that the distributions of age and BMI for the respondents are both unimodally skewed right. There are more respondents in the data set that have health insurance than do not and there are more females than males in the data set. These do not violate the assumptions that go into the generalized linear models we will create because we do not need our predictors to be approximately normally distributed for our model to be appropriate.

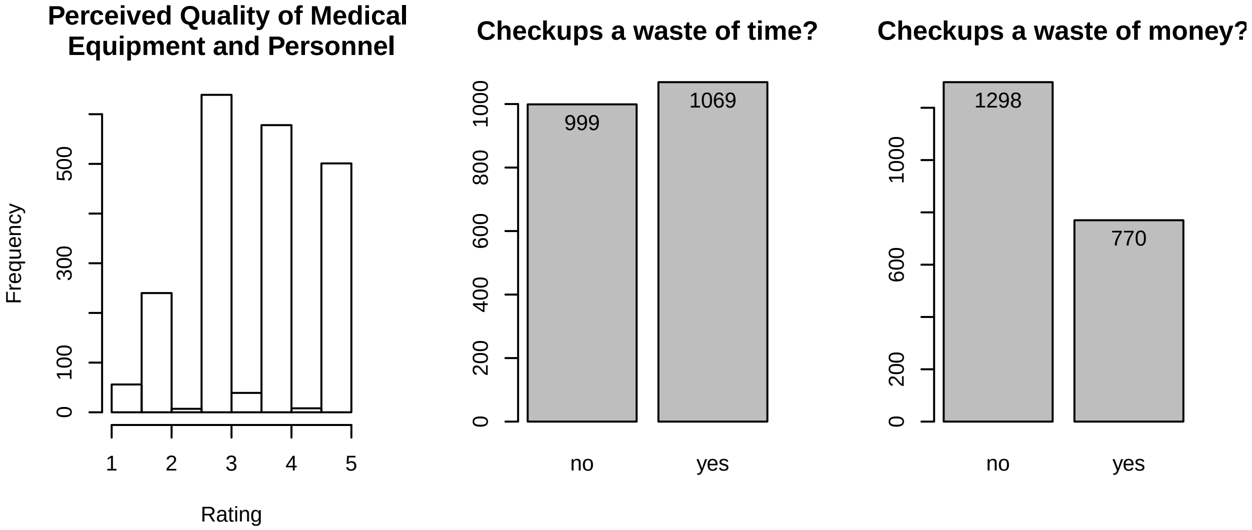
(2) The response variable we will use in our models is `HadExam`, which is a binary variable representing whether or not the respondent had a checkup in the past 12 months. A value of 1 represents yes and a value of 0 represents no.


Got a check−up in past 12 month?

It is interesting to note that about half of respondents did get a check-up in the past year and the other half did not. Our prediction classes are indeed balanced, not skewed, so we can proceed with our models without a transformation.

(3) Since the Ministry of Health is specifically interested in how people rate the value and quality of medical service and the quality of information they receive at check-ups, we will now look at the distribution of ratings for some of these variables.

Value and quality of medical service variables:

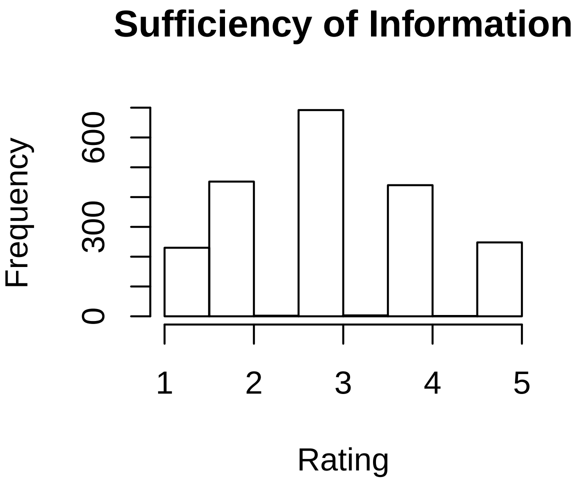

Perceived Quality of Medical Equipment and Personnel

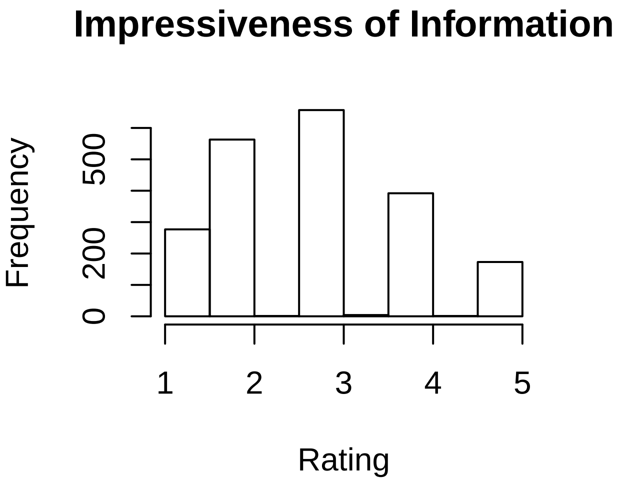
Checkups a waste of time?

Checkups a waste of money?

The histogram of quality of tangible medical equipment personnel, which is on a scale of 1-5, is centered around 3 with some left skewness. The majority of respondents rate this a 3 or above, which is good news for the Assistant Minister of Health. On the question of whether check-ups are a waste of time, it is interesting that respondents are split about 50-50 with their answers of yes or no, but when asked if check-ups are a waste of money, about two-thirds of people say no and one-third of people say yes. This is good for the Assistant Minister to know, because he can now target improvements of the check-up process to the content of exams rather than a monetary aspect.

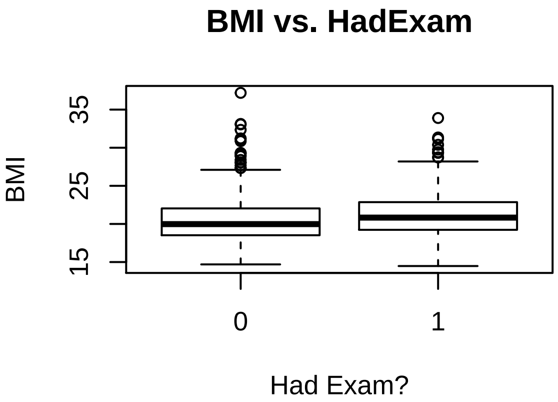
Quality of information variables:

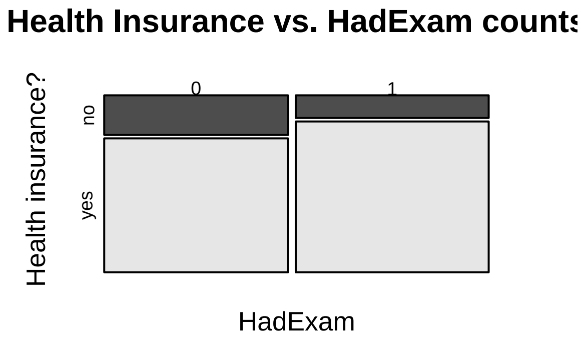


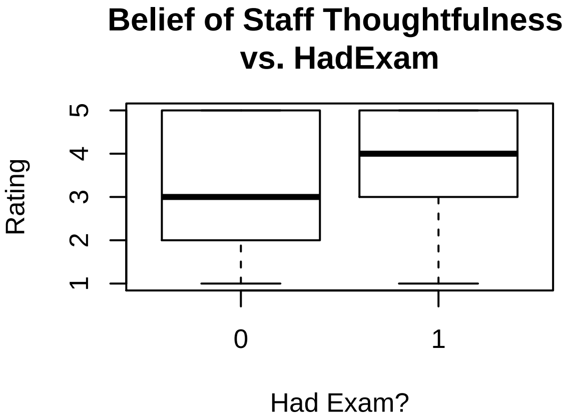


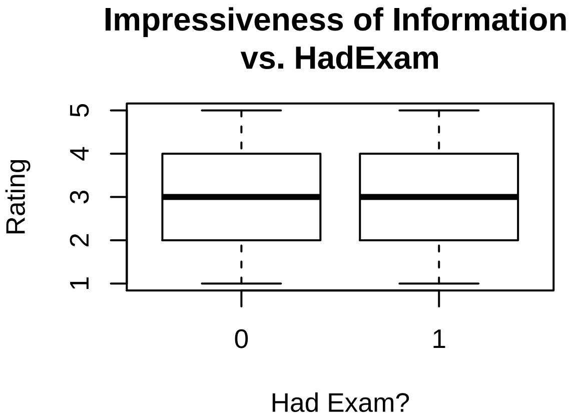
Since we saw that more people think checkups are a waste of time than waste of money, we want to get a deeper understanding of how respondents really view the information they receive at checkups. In the plots above, both of these variables have a mode of 3, meaning that a rating of 3 out of 5 had the most votes. For sufficiency of information, the distribution of respondents was very Gaussian-shaped around 3 with no skewness. However, for the impressiveness of information metric, the distribution has more ratings of 2 and 3 and less ratings of 4 and 5; it is right skewed. This is a sign for the Assistant Minister that for some reason, many people are not impressed with what they hear at checkups, so this factor might be a good area to target in the marketing campaign.

(4) Since we aim to predict whether or not respondents will get a check-up from many different variables, we will finish our EDA by exploring the relationships between `HadExam` and one variable of each variable grouping (demographic, value and quality of service, and quality of information). We will also look at health insurance status since the Ministry of Health is specifically interested in its potential effect.









Since our response variable is binary, we created boxplots and mosaic plot to help us visualize the bivariate relationships in our data. For the demographic variable BMI, the distributions look quite similar, although the median BMI of people who got an exam is very slightly higher than for people who did not. However, the range of BMI for people who did not get an exam is larger, so in general, we cannot identify a specific BMI range where people definitely will or will not get an exam. The Ministry of Health is specifically interested in whether health insurance is an important predictor of getting an exam, so in exploring that relationship, we found that the majority of people in the data set do have health insurance. We already know that about 50% of people get an exam and 50% do not, but we can visualize the fact that a higher percentage of people who did not get an exam also do not have health insurance, compared to people that did get an exam. This is a sign that health insurance status may help explain whether people will sign up for an exam. Finally, looking at the two boxplots, we see that impressive of information ratings are almost identically distributed for people who did and did not get an exam, but people that did get an exam have a median rating of staff thoughtfulness and responsibility of a 4 while people who did not have it a median rating of 3.

Initial Modeling and Diagnostics

(1) Since we want to predict the probability of the people signing up for a check-up with the binary variable `HadExam`, we will first construct a generalized linear model, specifically a logistic regression model. Here, we will use demographic variables about each person as well as variables about their opinions on the value and quality of medical service to predict whether or not they got an exam. This is "Model 1".

(2) Even though we excluded health insurance and quality of information variables from Model 1, there are still many variables in the model and we don't if they are all necessary to have. We will use a stepwise selection process with the AIC error estimate to remove the variables in Model 1 that do not help predict `HadExam`. Specifically, we will use backward elimination with the `step` command for variable selection. The AIC criterion is a measure of the difference between goodness of fit and model complexity, and this is useful to use because it helps us select variables based on not only the model's improved fit, but also a goal of minimizing the model's complexity. For example, we could have an extremely complex model with over a 100 variables predicting `HadExam` with a near-perfect fit, but that is not what we want because many variables are unnecessary. Our reduced model after this stepwise selection will be "Model 2".

After performing stepwise backward elimination, we reached a Model 2 that is based on only four predictors (not including the encoded dummy variables for each): job status, whether check-ups are a waste of time, whether checkups are considered important, and how often people believe check-ups should be done. Note that Model 2 is a logistic regression model (which is a generalized linear model with binomial family) just like Model 1.

(3) Next, we will create a third model, "Model 3", that adds variables to Model 2. We now want to include health insurance status and variables about the quality of information received at check-ups to our "reduced" Model 2 to see how those features contribute to predicting whether or not a person sign's up for a check-up. Additionally, since we will also be interested in potential associations between the health insurance variable and the four quality of information variables, we will add four interaction terms between health insurance and each quality of information variable to our model as well. We will keep the type of model, logistic regression, the same as Models 1 and 2.

(4) To test the goodness of fit of Model 3, we will conduct a test that compares Model 3 to Model 2 which did not include the health insurance or quality of information variables. Since we know that Model 2 was chosen from backward elimination from Model 1 based on

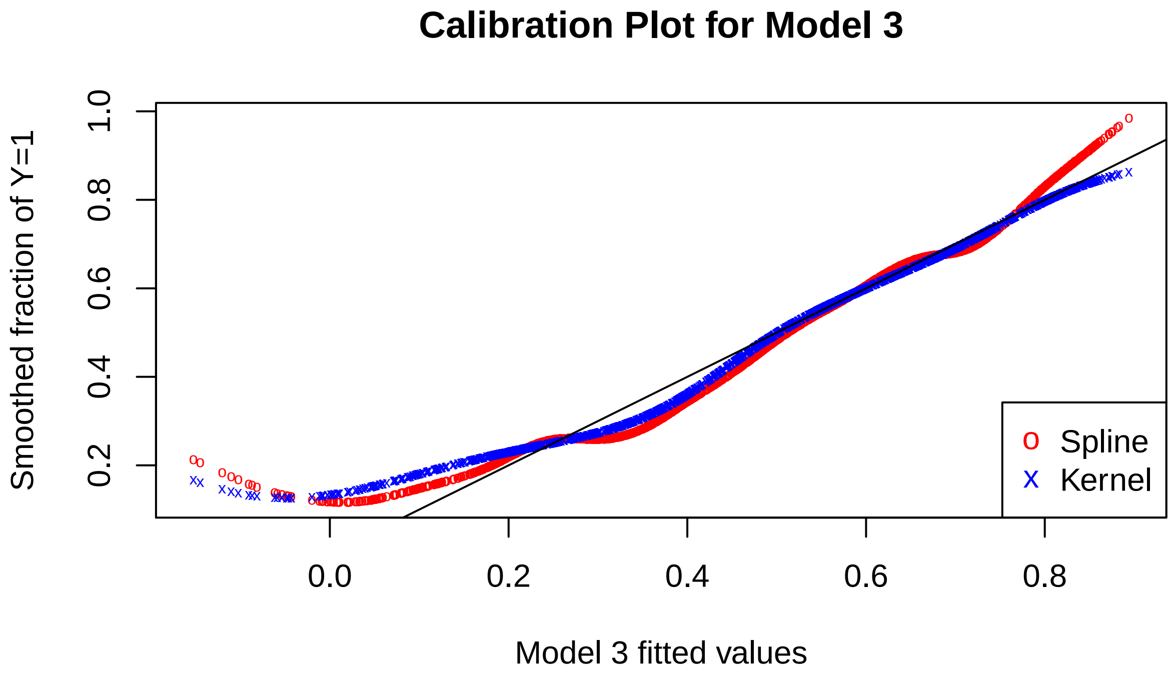
AIC, is considered a separate model from Model 3. Thus, it is safe to assume that in the deviance test comparing Model 2 with Model 3, the deviance test statistic will be chi-squared distributed with 9 degrees of freedom. We can proceed with the deviance test with `test = "Chisq"` as usual. The following are our hypotheses for the deviance test:

H_0 : Model 2 is correct

H_A : Model 3 is correct

After performing the test, we get a chi-squared test statistic for the drop in deviance of 2034.7 and p-value of 2.2e-16, so we have sufficient evidence to reject the null hypothesis that Model 2, the reduced model that does not include health insurance or quality of information variables, is the correct model. This suggest that it is worth including at least one of the health insurance or quality of information variables (or one of their interactions) into Model 2 for our prediction of whether a person signed up for a health check-up.

(5) Even though Model 3 is considered significant over Model 2, we still want to measure how well it really fits our data. To do this, we want to see how well-calibrated Model 3 is, meaning that we want to see how close its estimated probabilities are to the fractions of Y = 1 cases. This can be done by comparing the fitted values of Model 3 to the fitted values of various smoothers to see how they map estimated probabilities to fraction of Y = 1 cases. We will look at a spline and kernel regression smoother fit on the fitted values of Model 3.



In the above calibration plot for Model 3, we plotted the fitted values of the kernel regression smoother and spline against the fitted values of Model 3. We want to see how close these plots are to the y = x line. By eye, the smoothers essentially fall on the y = x line which is great. Model 3 appears to be quite well-calibrated. We do notice that around a fitted value of 0.4, the fraction of Y = 1 seems too low for the fitted values, and that at the extreme fitted values, the smoothers stray farther from the y = x line. This is okay, as we expect this behavior from the tail ends of the distribution of fitted values where there are less data points. Additionally, using a more quantitative method, we can compare the average distances between the two smooth curves to the y = x line. We will look at the distances at the 300th smallest and 300th largest fitted values so as to avoid the end effects. For the two smoothing functions, the average distances to the y = x line are 0.01313 and 0.0226, which are quite small and close to 0. We can conclude that Model 3 is well-calibrated and fits the data pretty well, so we do not need to perform any modifications.

Model Inference and Results

(1) Recall that in Model 3, we included interaction terms between health insurance indicator and the four variables about quality of information received at check-ups. We will interpret each of these terms' coefficients in Model 3 in order to understand the difference, if anything, between people with and without health insurance for each of these variables.

- Health insurance Yes * Sufficiency of information: 0.03953

For people with health insurance, a 1 unit increase in a person's sufficiency of information score predicts a `exp(-0.048676 + 0.039573 * 1) = 0.9909383` times increase in their odds of signing up for a check up compared to a person without health insurance, keeping all other variables fixed.

- Health insurance Yes * Attractiveness of information: -0.004023

For people with health insurance, a 1 unit increase in a person's attractiveness of information score predicts a `exp(-0.028849-0.004023 * 1) = 1.025137` times increase in their odds of signing up for a check up compared to a person without health insurance, keeping all other variables fixed.

- Health insurance Yes * Impressiveness of information: -0.009051

For people with health insurance, a 1 unit increase in a person's impressiveness of information score predicts a `exp(0.007458-.009051 * 1) = 0.9984083` times increase in their odds of signing up for a check up compared to a person without health insurance, keeping all other variables fixed.

- Health insurance Yes * Popularity of information: -0.015447

For people with health insurance, a 1 unit increase in a person's popularity of information score predicts a `exp(0.020422-0.015447 * 1) = 1.004987` times increase in their odds of signing up for a check up compared to a person without health insurance, keeping all other variables fixed.

In general, based on our interpretations of the interaction coefficients, we predict that people with health insurance have a slightly higher increase in their odds of signing up for a check-up than people without health insurance for a one unit increase in any of the quality of information variables' scores.

(2) Our coefficients on the interaction terms, however, still looked pretty small and did not seem to be significant in the Model 3 t-tests. We will conduct a deviance test comparing Model 3 to a model with all of the same variables except with no interaction terms. We will call this Model 4, and note that Model 4 is a "reduced" version of model 3. In comparing Model 3 and Model 4, we seek to answer the question: are the interactions necessary?

It is safe to assume that the deviance test statistic in comparing Model 3 and Model 4 will be chi-squared distributed with 4 degrees of freedom. We can proceed with the deviance test with `test = "Chisq"` as usual. The following are our hypotheses for the deviance test:

H_0 : Model 4 is correct

H_A : Model 3 is correct

After performing the test, we get a chi-squared test statistic for the drop in deviance of 0.32577 and p-value of 0.8044, so we fail to reject our null hypothesis. There is not enough evidence to suggest that Model 4, the reduced model that does not include any interactions between health and quality of information variables, is not the correct model. With a p-value of 0.8044 that is much greater than $\alpha = 0.05$, the interaction terms are clearly not significant and are not necessary to include in our model for predicting whether a person signed up for a health check-up. The relationships between the log odd of signing up for a check-up and the four quality of information variables are not significantly different for people who have and who do not have health insurance.

(3) Now let's do some inference using Model 4, the model with no interaction terms. We will compute the ratio between the odds of having a checkup for people with the most belief in the quality of information (rated each item 5) and the odds for those with the least belief in the quality of information (rated each item 1). Note that we already considered the relationships between these variables and health insurance status not significant, which is why we are working from Model 4. Remember, we have four quality of information variables in our model: sufficiency of information, attractiveness of information, impressiveness of information, and popularity of information.

The odds ratio we will calculate is:

$$\frac{odds\_of\_checkup\_given\_scores\_5,5,5,5}{odds\_of\_checkup\_given\_scores\_1,1,1,1}$$

The odds ratio is 1.069964 $\approx$ 1.07. This means that the odds of getting a check up is 1.07 times higher, or 7% higher, for a respondent with the most belief (rated all 5s) in the quality of the information compared to one with the least belief (rated all 1s).

(4) Finally, we will compute a 95% confidence interval for this odds ratio. This is shown in Table 1 below. We are 95% confident that the true ratio of the odds of getting a check-up between people who rated all 5s and people who rated all 1s is between 0.9888 and 1.1577. Since 1 lies in this interval, there is not enough evidence to suggest that, people who rated all 5s have a significantly higher odds of getting a check-up than people who rated all 1s.

Table 1: 95% Confidence Interval for Odds Ratio

2.5%	97.5%
0.988846	1.157736

Conclusions

(1) Based on our analysis, we found that people generally seem satisfied with the quality of tangible medical equipment and personnel when they get a check-up. Most people rate this a 3, 4, or 5. Half of people feel checkups are a waste of time, but only a third of people feel that they are a waste of money. Finally, people find the sufficiency of information average, but are less satisfied with the impressiveness of information they receive. Thus, we recommend that the Ministry of Health work to improve the way that doctors and nurses communicate with patients to make them feel more confident about the actual information they are receiving.

Now we address the Assistant Minister of Health's second question, which asks about what factors appear to make a person less likely to get an annual check-up. The best model we created was one that included the following variables: waste of time (0 or 1), suitable frequency for checkups (4 options), importance of check-ups (0 or 1), job status (6 options), health insurance status (0 or 1), sufficiency of information (1-5 score), attractiveness of information (1-5 score), impressiveness of information (1-5 score), and popularity of information (1-5 score). To summarize this, all of the quality of information variables are included, as well as three value and quality of medical service variables and two demographic variables. However, when we looked at the ratio between the odds of getting of a checkup for people who gave all 5s and all 1s for quality of information, our estimate of the true ratio was 1.07 but we found that with 95% confidence, the true ratio lies between 0.9888 and 1.1577. Since this spans below and above 1, we cannot be sure that one group (either all 5s or all 1s) really has a higher odds of getting a check-up than the other. So, even though we included these quality of information variables in the model, and even though we can identify points of improvement from exploring the variables' distributions, we cannot identify a statistically significant difference between the people that rate all 5s and all 1s on quality of information. We have answered the Assistant Minister's third question. Finally, the Ministry was interested in whether the relationship between quality of information ratings and getting a check-up is different for people with and without health insurance. Although, in our exploratory analysis, we saw a slight difference in the proportions of people with health insurance for people who got a check-up vs. people who did not, our statistical analysis in comparing models suggested that there is not a statistically significant interaction between the quality of information ratings and whether or not a person has health insurance. We do not recommend that a marketing campaign about the quality of information at check-ups be targeted towards particular groups of people with and without health insurance. There is not enough evidence that it will really be worth creating.

(2) We did not have any strong hypotheses when we began this study about what factors we thought would really contribute to whether or not someone gets a checkup. Perhaps, this is because as researchers in the U.S., we are not as familiar with Vietnamese cultural "norms", so it is more difficult to intuitively infer the most contributing or important factors to getting a check-up in Vietnam. As future undertaking, it may be interesting to compare how models and important predictors may vary for Vietnamese and American citizens. Ultimately, we found that quality of information variables that we suspected would be important from the initial exploratory analysis did not give a statistically significant contribution to the odds ratio of respondents with the most belief vs. the least belief getting a checkup. Perhaps, this could mean that beliefs about the value and quality of medical service or other variables unrelated to quality of information be more indicative of the odds of respondents getting a check-up. This is an interesting path for future research.

(3) We must remind the reader that the conclusions resulting from our analyses are made on the basis of statistical and mathematical evidence. We cannot magically conclude any relationship between variables that imply causation because we did not perform a controlled study to collect this data. Instead, we can only state that there are statistically significant associations, or lack thereof, between variables in our model and in our data set. Another limitation of this study involves the actual variables measured to collect the data set. It is possible there are confounding, but critical, variables such as “distance from doctor’s office”, “recent medical history”, or “income tax bracket” that could also be important in explaining the odds of people getting an annual check-up. We must acknowledge limitations in our modeling and inference. We only constructed models with specific combinations of variables and variable groups, but not all combinations. Also, we only considered whether there was an interaction between quality of information ratings and health insurance status and did not find a statistically significant association, but it is possible there is a significant association with a different demographic variable such as job status for example. These issues come with an inherent restrictiveness in the types of models we can create. Finally, it is worth noting that our inference (the odds ratio and confidence intervals) stems from the model that we found performed based for predictions. However, the best predictive model does not necessarily imply that we are using the best inferential model, so it is possible that we could have come to different specific results, for better or for worse, depending on the model we used. This is why it is important for future research on this subject to evaluate more models than just what was used in this study. This also explains why we kept the quality of information variables in our final model, even though they *themselves* did not return a statistically significant odds ratio during inference. It may be worth considering other metrics to investigate for inference in any future work of this subject.