36-402 DA Exam 2

James "Morgan" Hawkins (jmhawkin)

5/5/2023

Introduction

Receiving regular health checkups is desirable for public health in Vietnam. However, many Vietnamese people do not receive regular checkups. In order for the Ministry of Health to most efficiently mitigate this issue, they must know the primary reasons that people do not receive regular checkups. Knowing these primary reasons allows the Vietnamese Ministry of Health to create campaigns to change the public opinions around why people don't receive checkups.

(1) There are three questions most relevant to our issue. First, we would like to know how people generally rate the value and quality of medical service and information they receive in checkups. Second, we are interested in exploring what factors are most associated with someone being less likely to receive a checkup every 12 months. Last, we are interested in exploring if the quality of information patients receive in a checkup is an important predictor of whether patients get checkups, and whether this relationship differs for patients with and without health insurance.

(2) In this study we found that general consensus around the quality of information and value of service is negative. We found a job status of student, believing check-ups to be a waste of Time, believing check-ups aren't important, and a suitable frequency greater than 18 months to be the factors most associated with someone being less likely to have received a check-up in the last year. Although sentiment around quality of information is negative, we did not find any quality of information variables to be predictive of someones probability of receiving an exam in the past 12 months.

Exploratory Data Analysis

(1) In this analysis, we will treat the following variables as categorical: `HadExam`, `Sex`, `Jobstt`, `HealthIns`, `Wsttime`, `Wstmon`, `Lessbelqual`, `NotImp`, `SuitFreq`. The remaining variables will be treated as continuous.

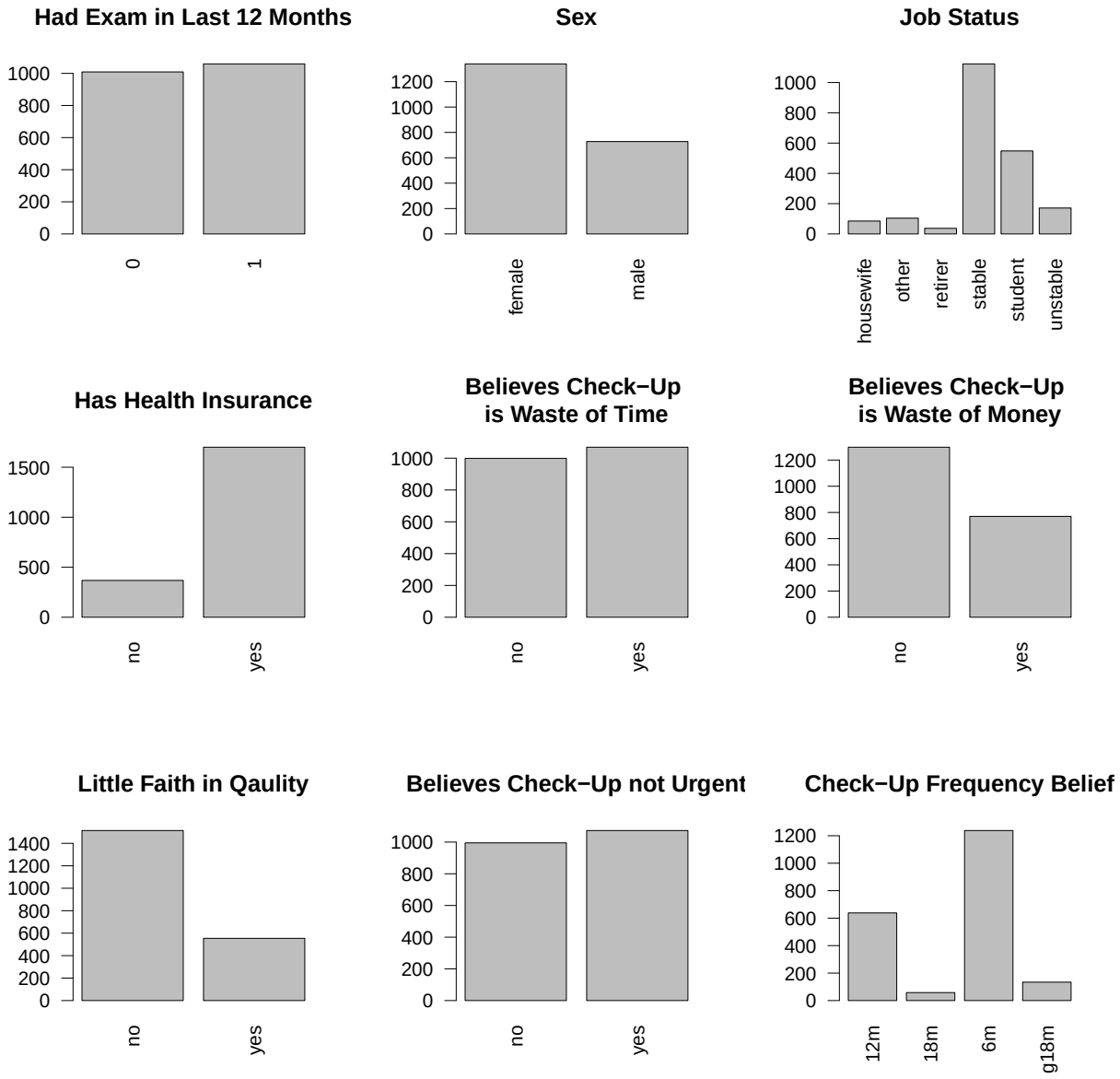

Figure 1: Categorical Variable Distributions

In Figure 1 we see that only around 51.2% of individuals in our data set have had a checkup in the past 12 months. We also notice that 64.8% of our data set contains female respondents

Table 1: Quality of Information Variable Correlations

	SuffInfo	AttractInfo	ImpressInfo	PopularInfo
SuffInfo	1.000	0.665	0.626	0.577
AttractInfo	0.665	1.000	0.686	0.608
ImpressInfo	0.626	0.686	1.000	0.639
PopularInfo	0.577	0.608	0.639	1.000

which is nearly two thirds. The most common job status is stable, and the least common is retired. Around 82.2% of respondents have health insurance. Around 51.7% think check-ups are a waste of time, but only 37.2% think they are a waste of money. 26.8% have little faith in the quality of medical service. 51.9% think that check ups are not urgent, but more than 90% of respondents believe checkups should be conducted every 12 months or sooner. Many different locations were surveyed, but around 79.2% of respondents were surveyed in Hanoi. The second most common survey location was Hungyen with only around 7.5% of respondents.

In Figure 2 we see that `Age`, `Weight`, and `BMI` have right skewed distributions. They have estimated means of 29.2, 54.7, and 20.9 respectively. They also have estimated standard deviations of 10.1, 9.7, and 2.7, respectively. We notice 2 outliers: a respondent with `Age` 83 and a respondent with a `BMI` of 37. `Height` has an approximately symmetric distribution with sample mean of 161.6 and sample standard deviation of 7.6. **(3)** Variables `Tangibles`, `Empathy`, `SuffInfo`, `AttractInfo`, `ImpressInfo`, and `PopularInfo` all take values between 1 and 5 inclusive. They have sample estimated means of 3.6, 3.5, 3.0, 2.7, 2.8, and 2.8 respectively. They also have estimated standard deviations of 1.1, 1.3, 1.2, 1.1, 1.1, and 1.3, respectively.

(2) Our response variable of interest in this analysis will be `HadExam` which indicates whether the respondent has had a check-up in the past 12 months. `HadExam` follows a Bernoulli distribution with an estimated p = .512 ($CI_{95\%} = [.490, .534]$).

We would now like to explore relationships between out continuous predictor variables. We see in Figure 3 that weight and body mass index appear to have a strong positive linear relationship with a correlation coefficient of .842. We also notice that height and weight have a positive relationship with a correlation coefficient of .697.

(4) Next, we see in Table 1 that our quality of information variables are quite correlated. It may be beneficial to sum these variables into one composite variable rather than include all variables individually in our final model.

Lastly, we would like to explore the relationships between our response `HadExam` and our

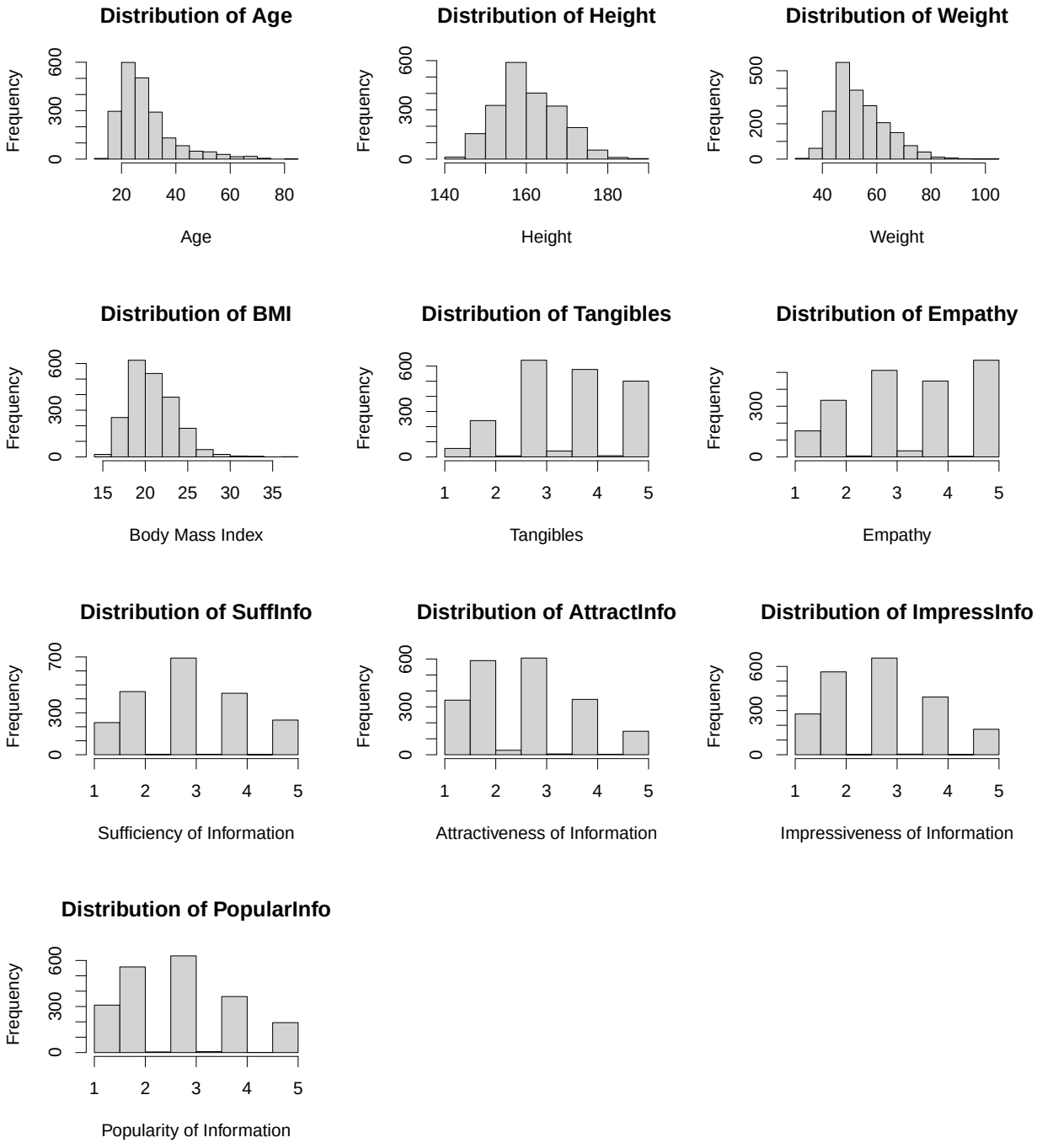

Figure 2: Continuous Variable Distributions

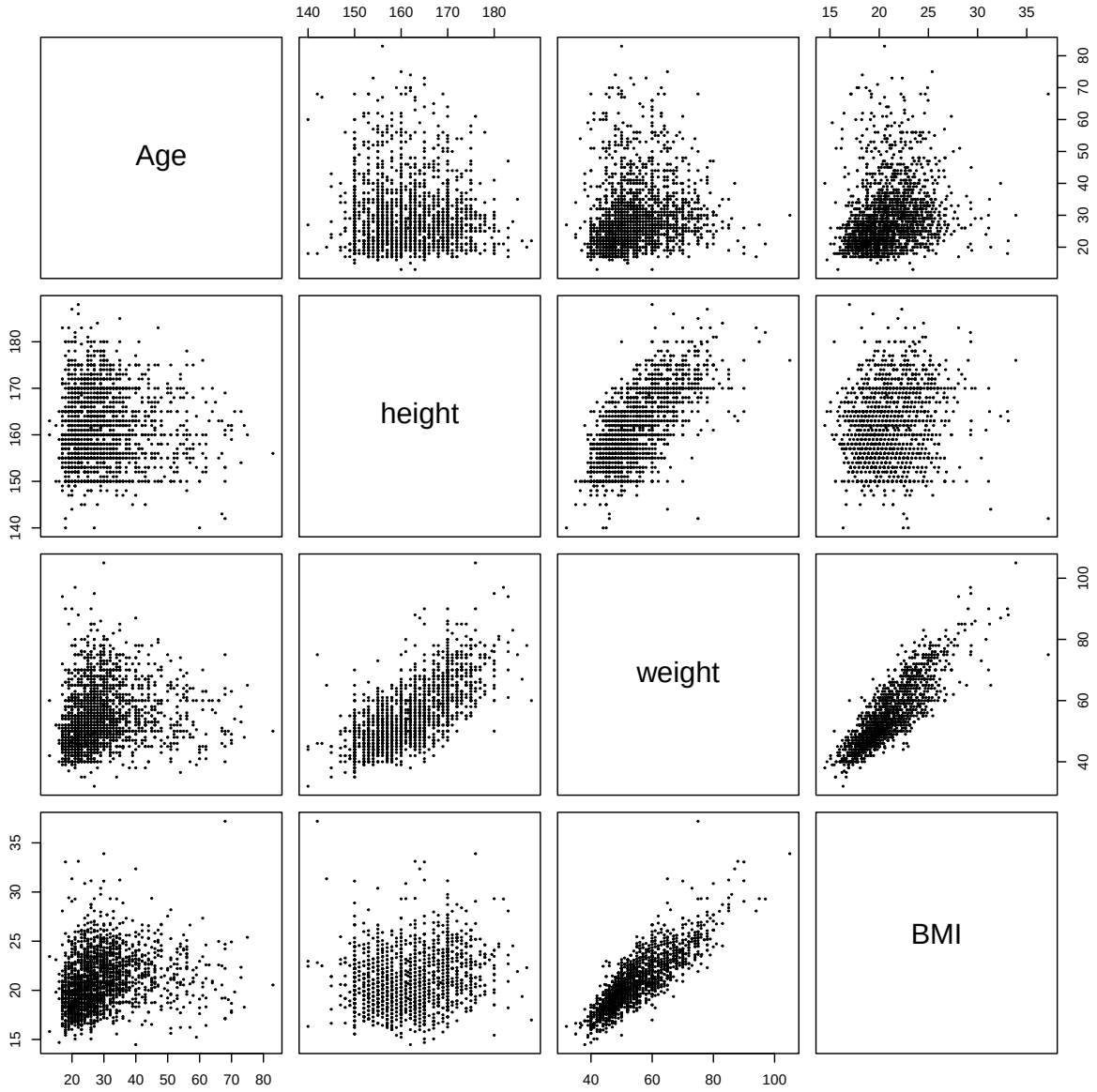

Figure 3: Pairs Plot of Demographic Variables

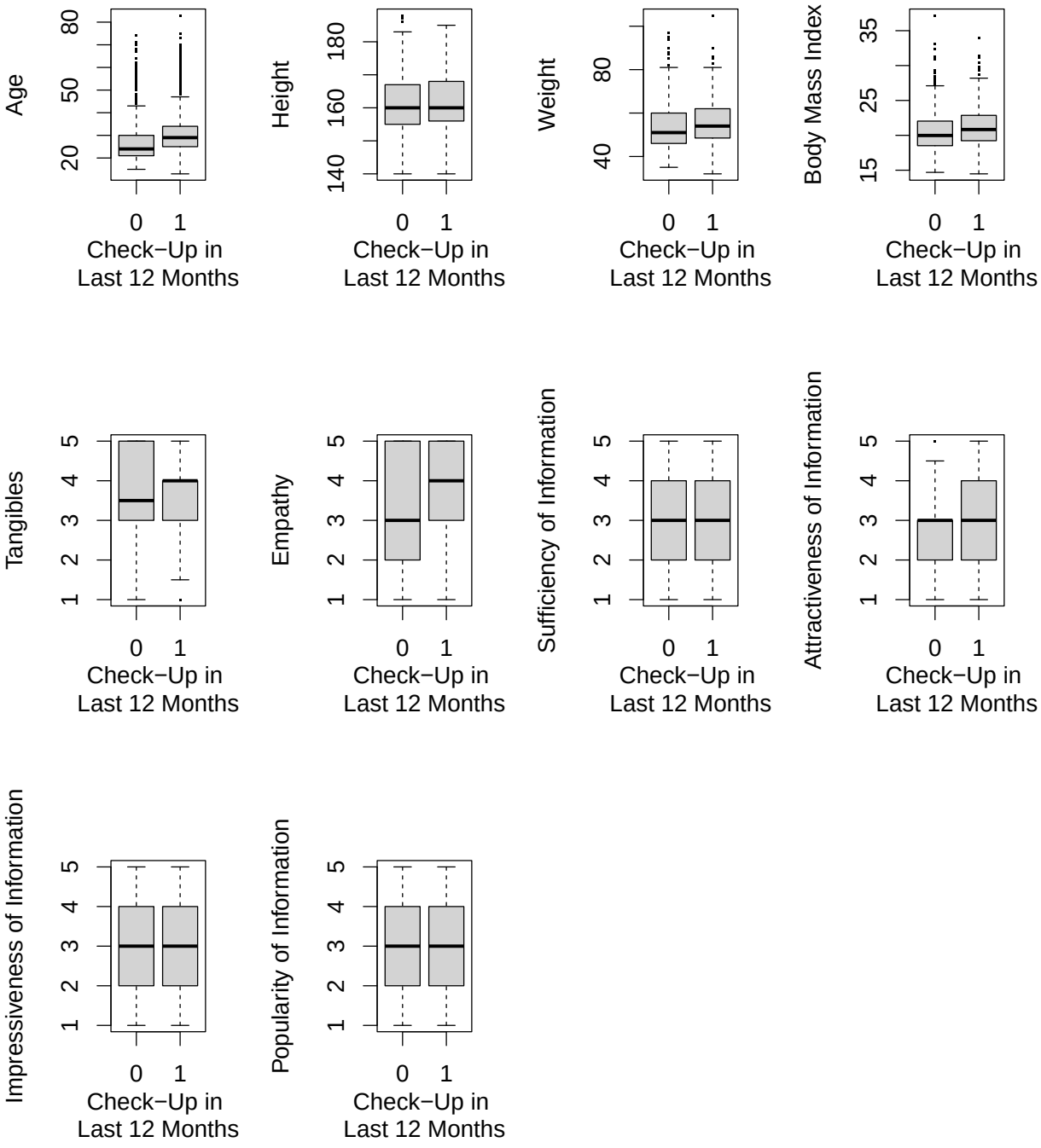

Figure 4: Continous Variable Relationships with Response

continuous predictors. We see in Figure 4 that our patients who have had a check-up in the last 12 months appear to have higher reported values for Age, Weight, BMI, Tangibles, and Empathy. Age and Empathy appear to have the strongest relationship.

Initial Modeling and Diagnostics

(1) We will begin our initial modelling by fitting a generalized linear model (model 1) predicting our response, `HadExam`, from all demographic variables and quality of information variables (excluding whether the respondent had health insurance). Fitting this model, we see that our model 1 has significantly more predictive power than the null model ($\chi_{19} = 98.64,\ p = 9.43e-13$).

We see that many variables have insignificant coefficient values so we would now like to use a backwards step wise search to remove variables that do not add predictive power to our model. **(2)** Conducting this step wise search, we see that our model (model 2) now includes just `Jobstt`, `Wsttime`, `NotImp`, and `SuitFreq` as predictors. This is a significant reduction in model complexity. Model 2 has 11 degrees of freedom instead of the 20 degrees of freedom in model 1.

(3) Now that we have obtained a better model through a step wise search, we will add quality of information variables as well as their interactions with the variable `HealthIns` to model 3. This will allow us to explore the relationship between our quality of information variables and our response, as well as if this relationship changes between patients with and without health insurance. **(4)** After fitting, we see that our model has a deviance of 410.739. The deviance of our model follows a χ_{2048} distribution. The p-value of our deviance test is approximately 0 ($\chi_{2048} = 410.739,\ p < 2.2e-16$) so we see that our model fits well because we expect better fitting models to have lower deviance values. However, we also notice through a deviance test that this model does not perform significantly better than model 2 ($\chi_9 = 8.924,\ p = .444$).

We would now like to investigate how well calibrated our model is. **(5)** In Figure 5, we see that our model appears to be well calibrated. Our smoothed curve between observed values and fitted values appears to follow the line which we expect it to follow for calibrated models. For fitted values below .2, we notice that our smoothed curve strays from the line a little. However, this could be due to the fact that their aren't many fitted values below .2, so our our kernel smoother has higher variance in this area.

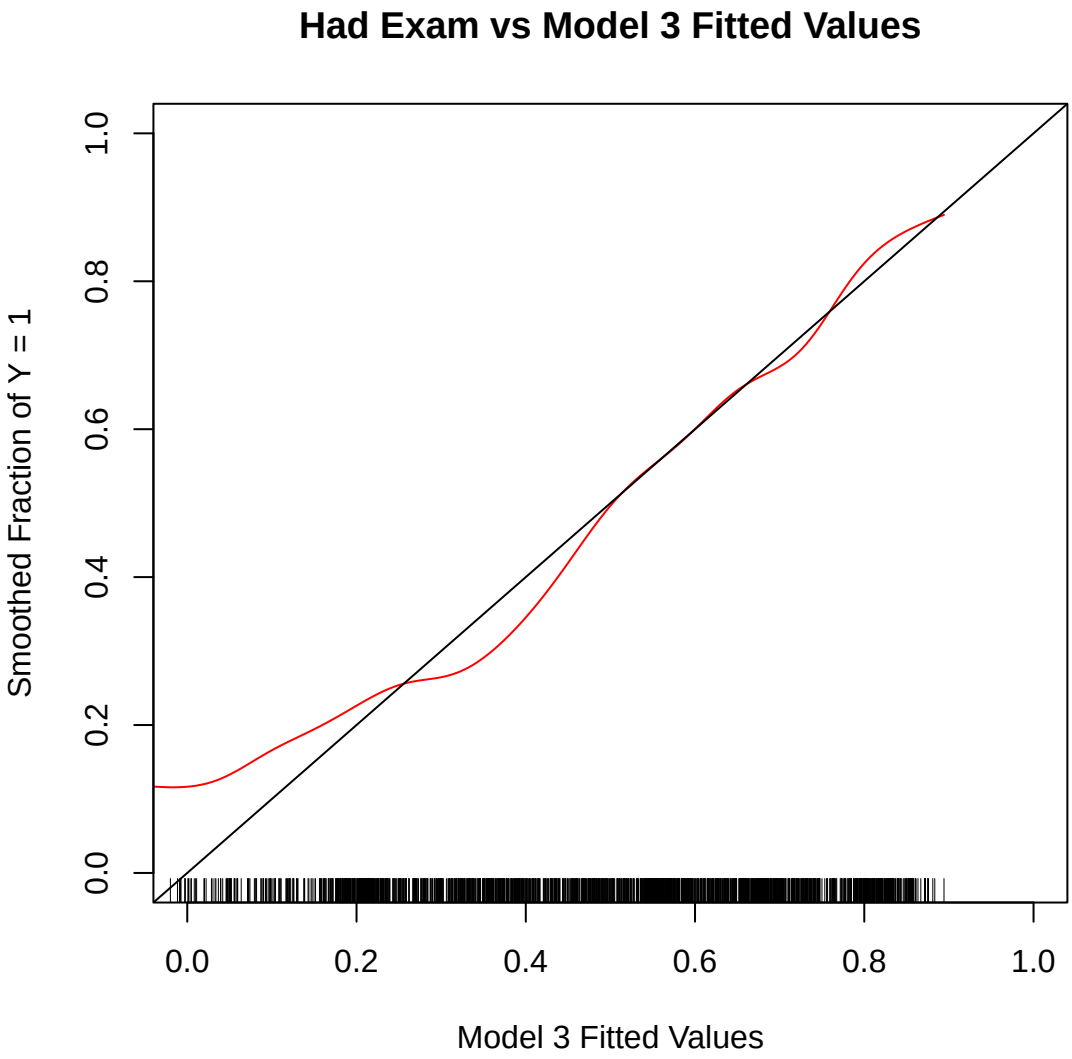

Figure 5: Model 3 Calibration Plot

Model Inference and Results

(1) Our fit in model 3 estimates that, all else constant, individuals with health insurance have an associated odds of having a check-up in the last 12 months that is 1.13 times greater than those who do not have health insurance. The associated increase in the log odds for a unit increase in `SuffInfo`, `AttractInfo`, `ImpressInfo`, and `PopularInfo` is estimated to change for respondents with health insurance by .0396, -.0040, -.0091, and -.0154, respectively. We notice that none of these estimates are significantly far from 0, so we do not have evidence that people with and without health insurance's probability of receiving check-ups have different relationship with the quality of information variables.

(2) We would now like to conduct a formal hypothesis test to determine if there is a significant difference between model 3's predictive power and the predictive power of the same model without interaction terms. To test this we will conduct a deviance test at confidence level $\alpha = .05$. Our null hypothesis is $\text{Deviance}_{red} = \text{Deviance}_{full}$ and our alternative hypothesis is $\text{Deviance}_{red} > \text{Deviance}_{full}$. Conducting this test, we fail to reject the null hypothesis and conclude that we do not have sufficient evidence that $\text{Deviance}_{red} > \text{Deviance}_{full}$ ($\chi_4 = .326$, $p = .988$). In the context of modelling, this means that the relationship between the quality of information variables and our response `HadExam` does not change for respondents who do and don't have health insurance.

Since we've determined through our hypothesis test there isn't evidence that the relationship between our quality of information variables and our response changes if someone does or doesn't have health insurance, we can calculate the ratio between the odds of having a checkup in the last 12 months for people with the most belief in the quality of information and the odds for respondents with the least belief in the quality of information without including interaction terms. **(3)** This ratio is calculated to be 1.070. This means that the odds of receiving a check-up in the last 12 months for those with the most confidence in the quality of information is 1.070 times greater than the odds of those with the least confidence.

We would now like to create a confidence interval for our computed estimate of the ratio between odds. Assuming that our coefficients are normally distributed, **(4)** we believe that there is a 95% chance that if we were to recreate this experiment with new data, our estimate for the ratio between the odds of having a checkup in the last 12 months for people with the most belief in the quality of information and the odds for respondents with the least belief in the quality of information would fall in the range [0.989, 1.158]. We note that 1 falls is this range, which does not inspire confidence that this ratio is actually greater or less than than 1.

Conclusions

After exploring our data and modelling the probability of having an exam in the past 12 months through generalized linear models, we have made the following findings. **(1)** Firstly, the public generally does not believe the quality of information being provided by check-ups is valuable at the moment. All four of our quality of information variables had means very near 3 or below 3. However, these are not the strongest predictor of `HadExam`.

Although 49.6% rate the empathy of their doctors as a 4 or higher, 51.7% still think check-ups are a waste of time and 37.2% think they are a waste of money. Our demographic and quality of service variables are much stronger predictors of `HadExam` than our quality of information variables. The set of variables that we found to be most predictive of our response was Job Status, Waste of Time, Not Important, and Frequency.

(2) We noticed that adding waste of money did not improve our AIC and this is likely because those who believe check-ups are a waste of time are more likely to believe they are a waste of money (correlation = .444). We also notice that job status is a significant predictor of our response. This is possibly because job status is predictive of income which would contribute to someone's ability to receive a checkup. We also note that this likely contributes to the exclusion of waste of money from our model because those who earn less would be more likely to believe check-ups are a waste of money. Lastly, we notice that the frequency with which someone believes they should receive a checkup is predictive of if they received a check-up in the past 12 months. This is expected because, for example, someone who believes you should receive a check-up every 18 months would be less likely to have received a check-up in the last 12 months if asked at a random time.

There are some limitations to our analysis that stem from the data available for our use as well as the nature of this study. **(3)** Firstly, there are additional demographic variables that we would like to control for. We expect that distance to nearest clinic would be a strong predictor of our response. This is not available to us, so we are unable to asses whether setting up more clinics would help alleviate the issue of people not receiving check-ups. Also, we expect variables such as age to be correlated with distance to nearest clinic (harder to receive medical attention) so these correlations may be causing our estimates to be inaccurate as they are capturing effects caused by distance to nearest clinic. Secondly, this is an observational study. Our predictor variables were not randomly assigned to respondents so we cannot make any causal conclusions. This means that confounding variables may be the cause of our observed relationships and there may not be a causal relationship between the predictors we used and the response.