36-402 DA Exam 2

Chaiyatat (Chawaldit)

5/5/2023

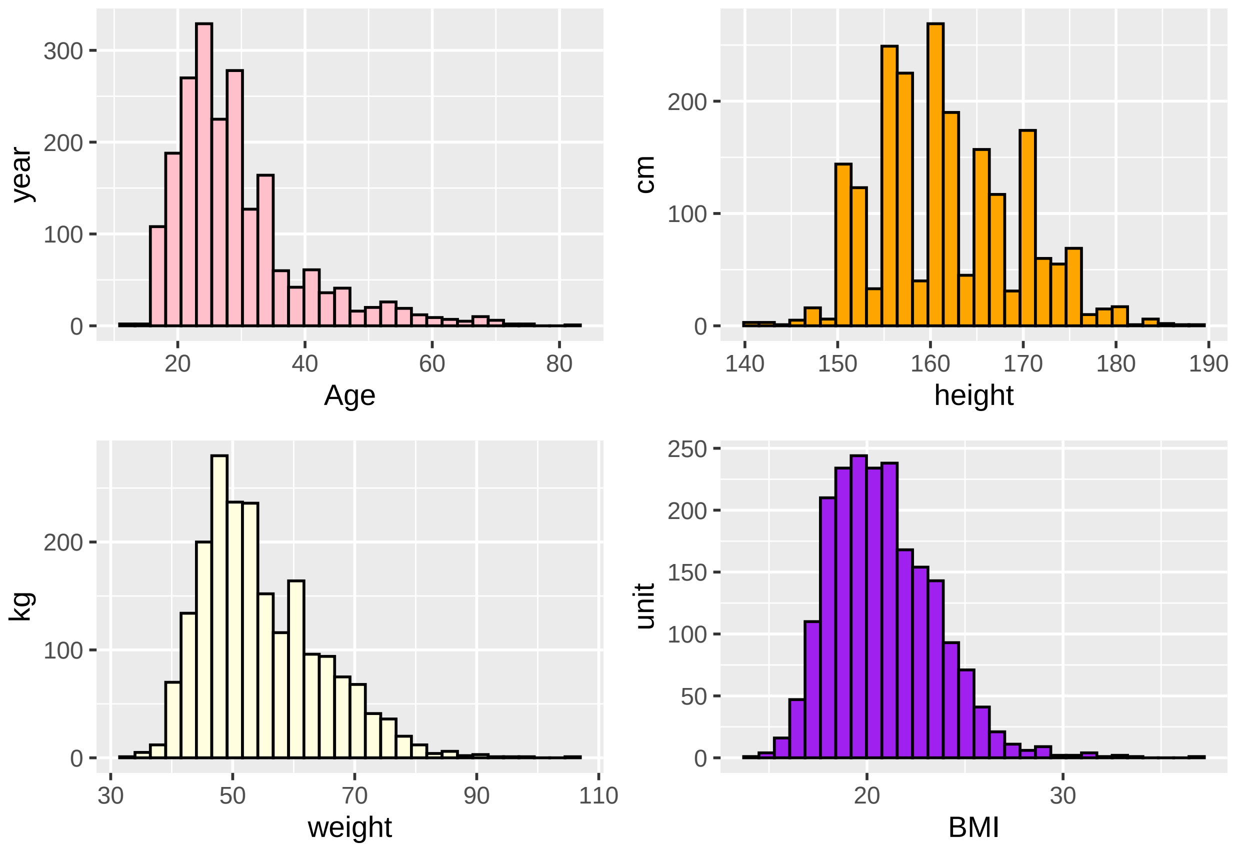
Introduction

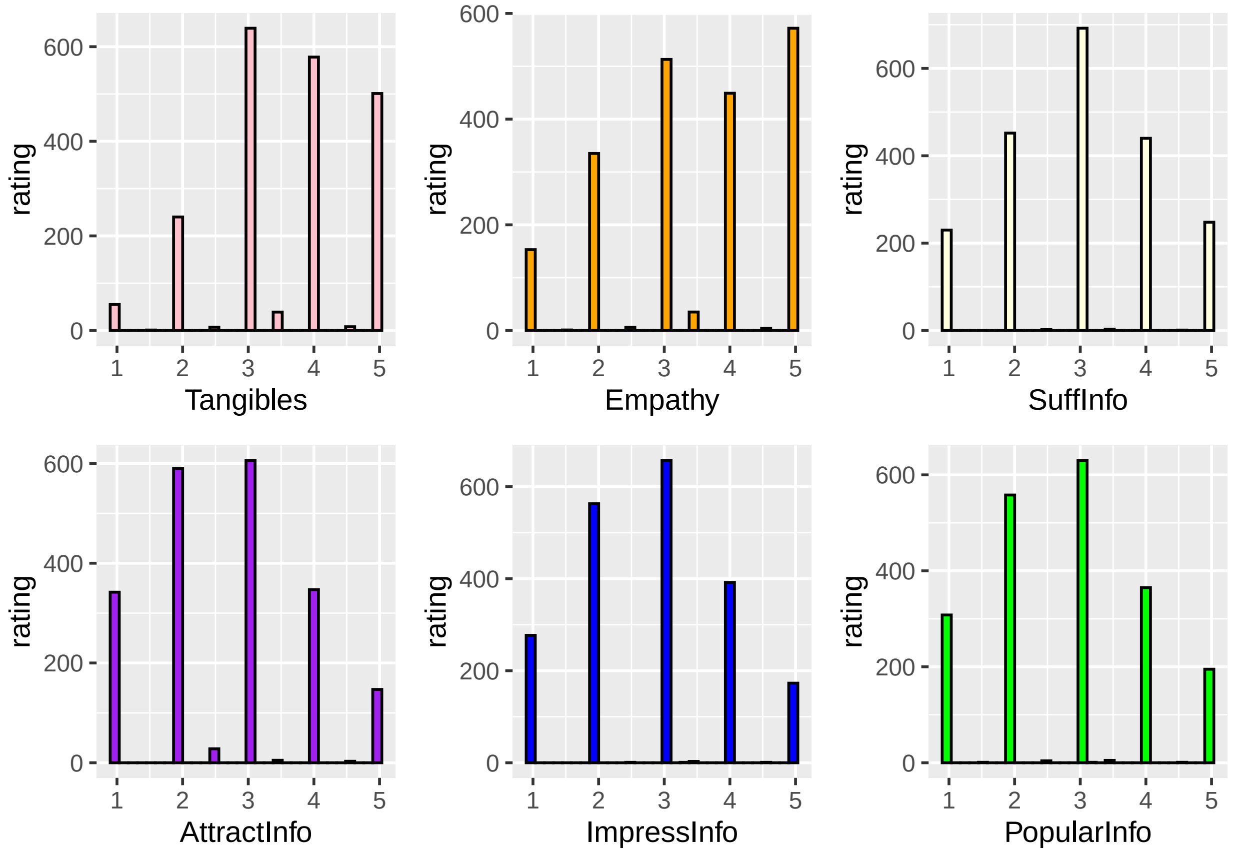
(1) Medical care for serious disease can be very expensive. Ideally, patients would get regular check-ups so that serious conditions could be detected and treated early. Public health researchers in Vietnam want to improve the rate of medical check-ups by answering three research questions: 1.) Overall, 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.) Does the evidence suggest quality of information is an important predictor of whether patient gets check-ups, and does this depend on whether the people has health insurance?

(2) Briefly mention your final findings. Connect these to the Assistant Minister of Health's substantive questions.

Exploratory Data Analysis

1. Explore the key variables you need to conduct your analysis. Describe them with any necessary univariate EDA. Specify which variables you will treat as continuous and which ones you will treat as categorical in your analysis.


year
300
200
100
0
20
40
60
80
Age
cm
200
100
0
140
150
160
170
180
190
height
kg
200
100
0
30
50
70
90
110
weight
unit
250
200
150
100
50
0
20
30
BMI



(1) The obvious continuous variables are place, Age, Sex, height, weight, and BMI. The obvious categorical Variables are Sex, Jobstt, Healthins, Wsttime, and Wstmon. There is a few ordinal variables namely Tangibles, Empathy, SuitFreq, SuffInfo, AttractInfo, ImpressInfo, PopularInfo. All of these except SuitFreq will be treated as continuously variable for our analysis. **(2)** The response variable for our research question is HadExam, which is also a binary variable.

(3)

(4) In terms of how people rate quality of medical service, Wsttime and NotImp distributions between “yes” and “no” are roughly even where as Wstmon and Lessbelqual distributions skew more toward “no” at 73% and 63% respectively. Most response prefer a check-up frequency of 12 months with a significant proportion preferring every 6 months. The rating mode is three for all rating based medical service predictors. In terms of hoe people rate quality of information, the rating mode is also 3 for all rating based predictors. Out these features, it seems all predictors in Quality of Information have roughly the same distribution.

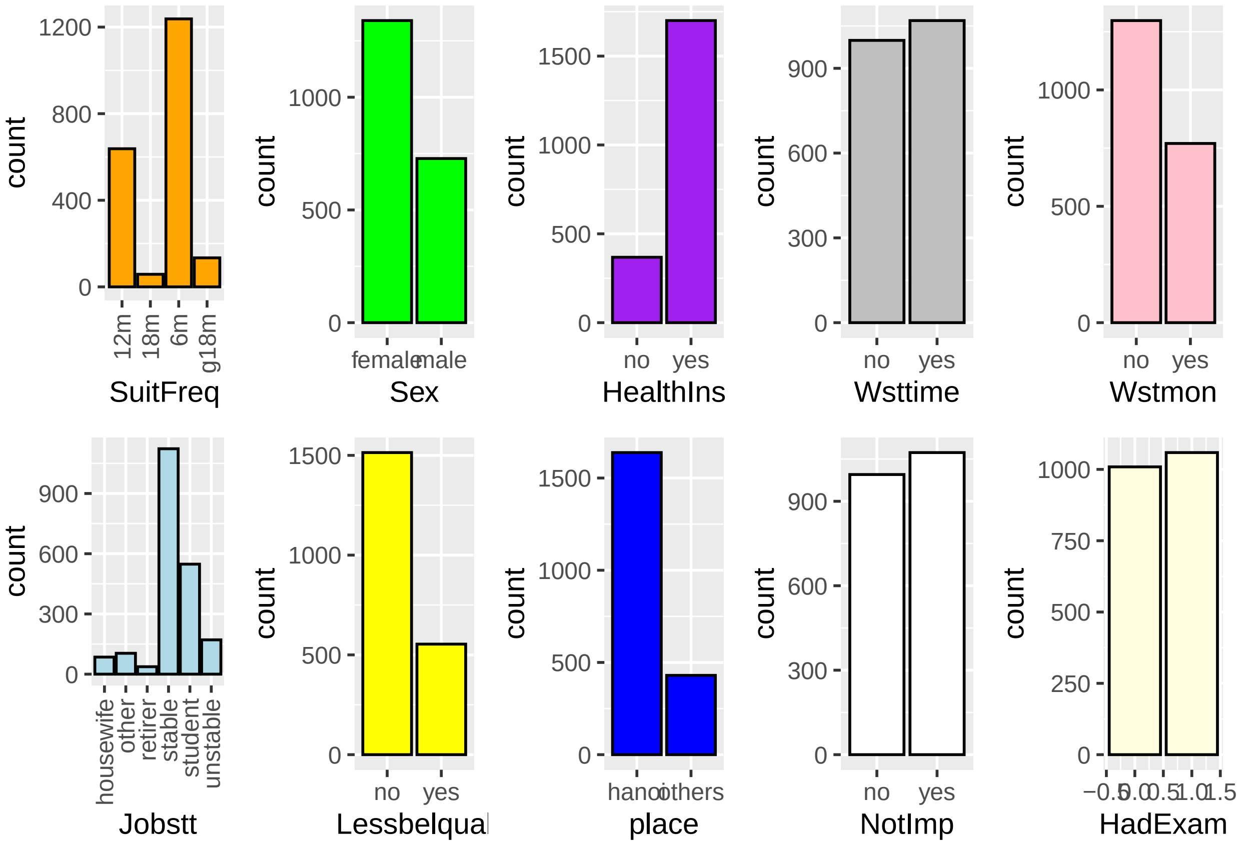


Figure 1: Figure 2. Categorical Variable Distribution

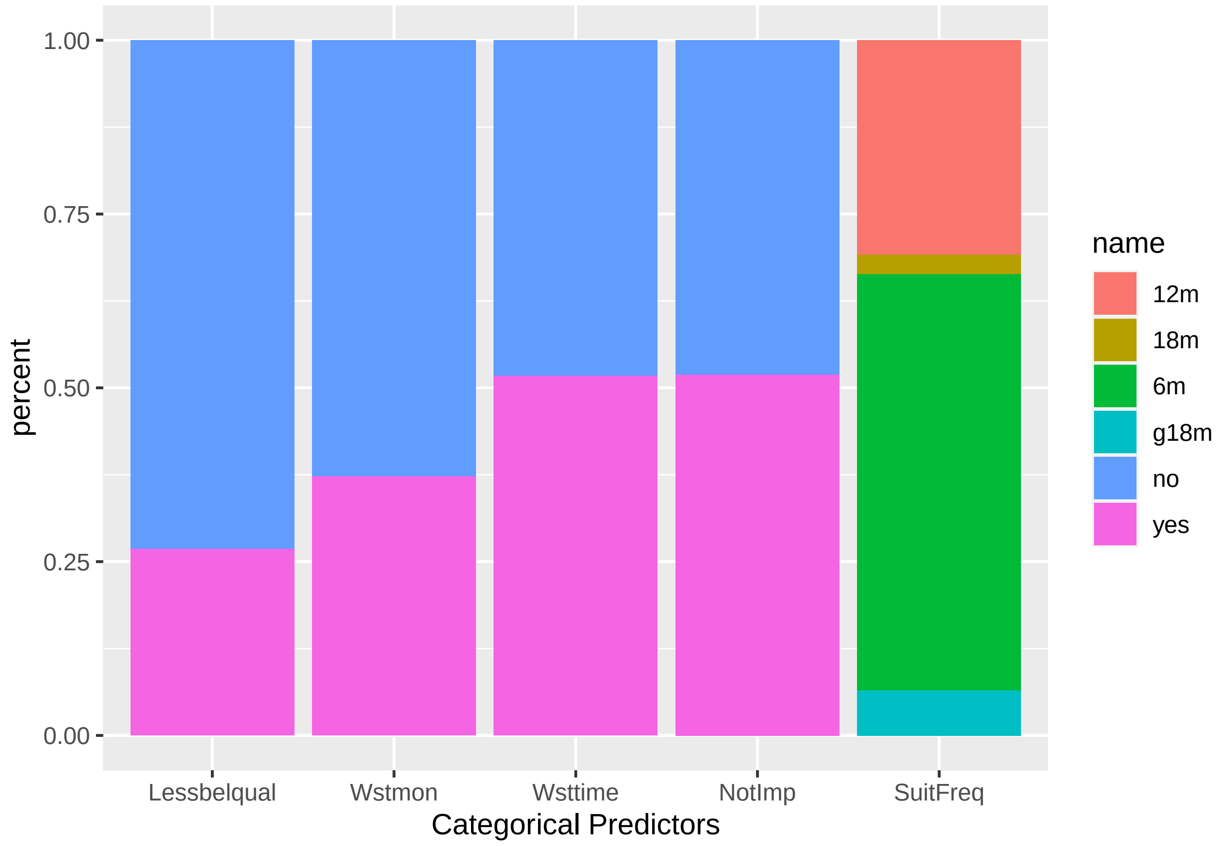


Figure 2: Figure 3. Medical Service Value and Quality Categorical Variable Distribution

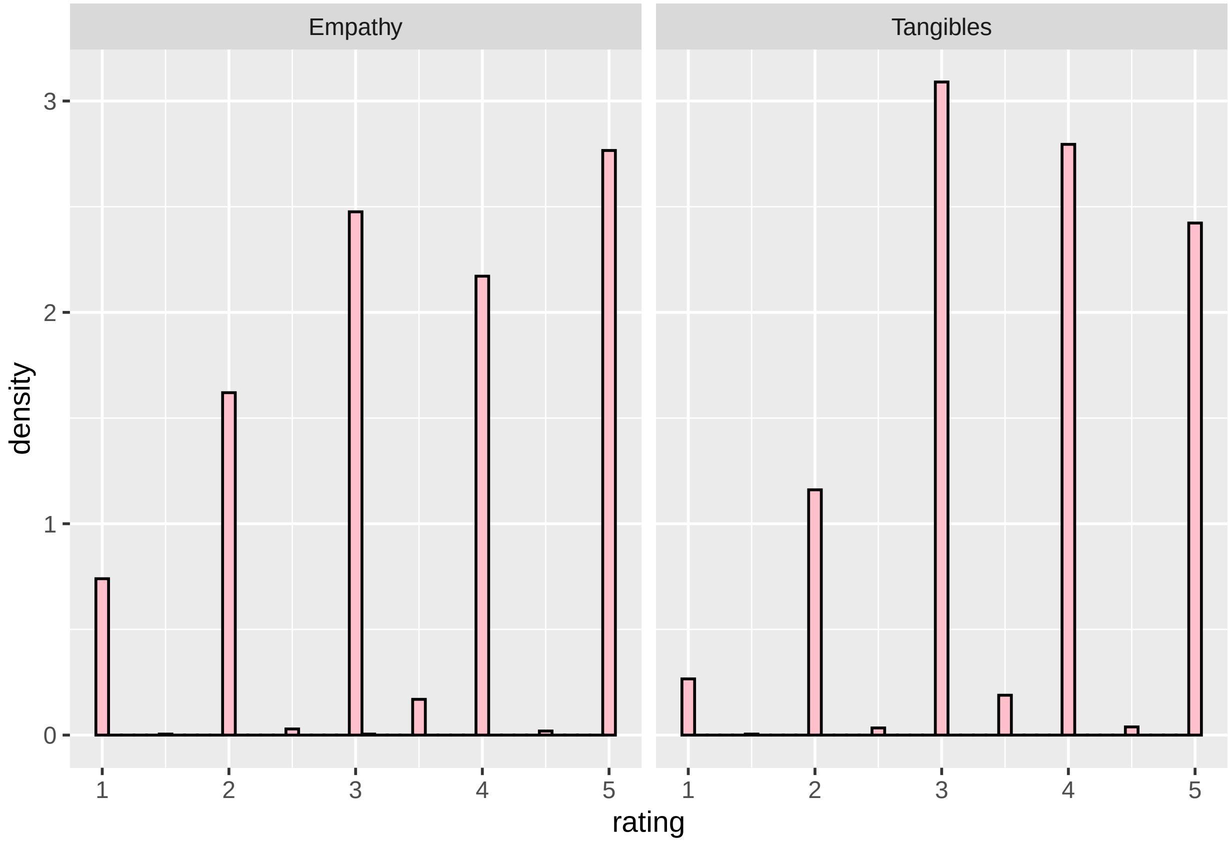


Figure 3: Figure 4. Medical Service Value and Quality Continuous Variable Boxplot

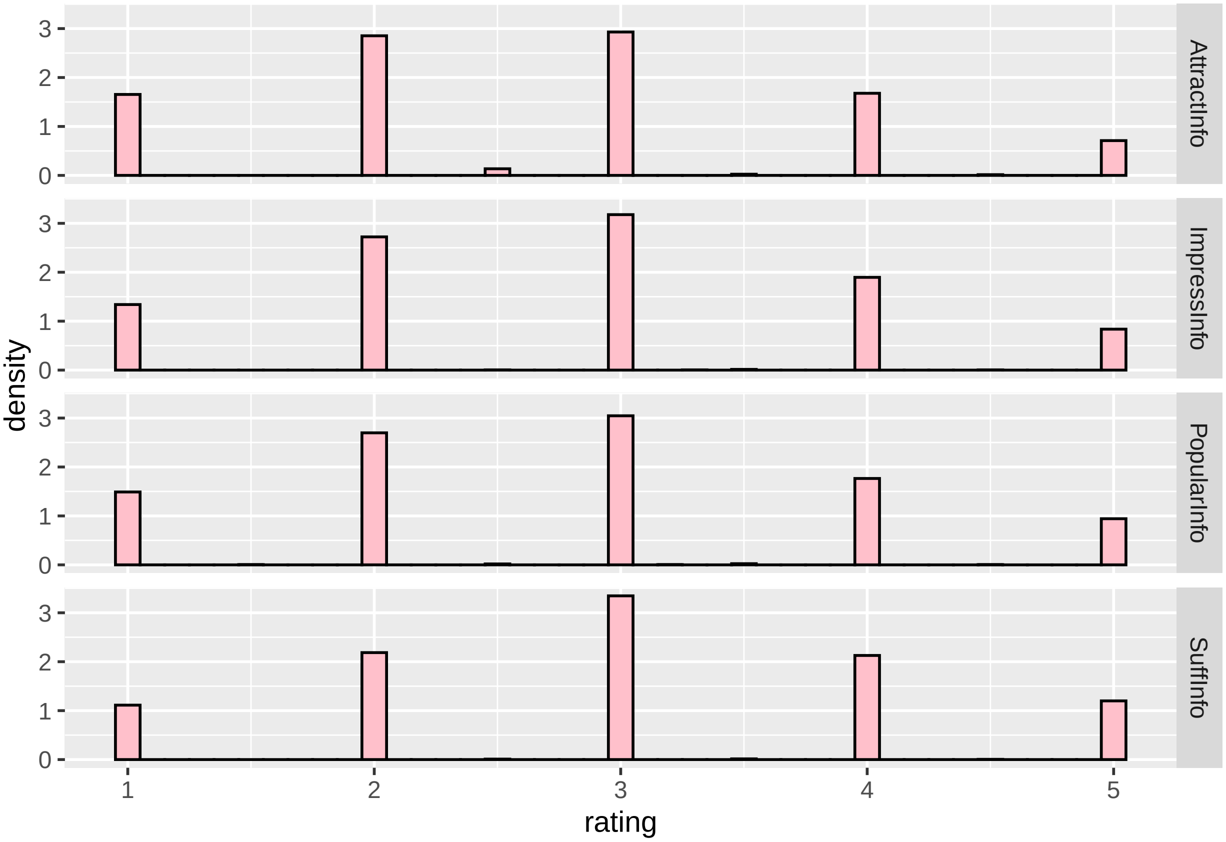


Figure 4: Figure 5. Quality of Information Continuous Variable Histogram

Initial Modeling and Diagnostics

1. Formulate a generalized linear model that predicts the response variable as a function of all the demographic variables, and the variables regarding the value and quality of medical service. Do *not* include health insurance or the variables about the quality of information presented in check-ups. Call this Model 1.

(1) To better understand our second and third research question, we fit an initial model to predict `HadExam`, which uses all demographic variables except `health insurance` and value and quality of medical service variables. **(2)** Because there are too many variables, we use AIC error estimate stepwise selection to remove statistically insignificant variables. As a result, our model 2 has dropped `Age`, `Sex`, `height`, `weight`, `BMI`, `Wstmon`, `lessbelequal` `Tangibles`,`Empathy`, and `jobstt_unstable`. **(3)** We next create a third model which factors in health insurance and its interactions with quality of information variables to later check if the quality of information variables have different associations between patients with and without health insurance. **(4)** To test goodness of Model 3 fit, we check the null deviance which is 2865.6 on 2067 degrees of freedom, and compare it to the residual deviance is 2403.6 on 2048 degrees of freedom. Because both values are quite large, the null and residual model might not best explain the data. **(5)** We use calibration plot to check that model 3's observed probabilities match the observed proportions of outcome from the data. From the plot, probabilities greater than 0.5 smoothly matches the observed probability while probabilities lower than 0.5 are less in accordance.

Model Inference and Results

(1) The four interaction terms from Model 3 are respondents with health insurance with rating of the attractiveness of information received, rating of the impressiveness of information received, rating of the popularity of information received, and rating of sufficiency of information received. There respective coefficient and p-values are (-0.009, 0.96), (-0.044, 0.80), (-0.075, 0.63), and (0.173, 0.28). Since all four coefficients have a p-value greater than .05, we fail to reject their null hypothesises that the interaction terms' coefficient value is zero, and thus do not have a significant effect on response variable. **(2)** Equivalently, we did a chi-square deviance test between model 3 against a simplified version of model 3 without the interaction terms, and the difference in devaince is 1.2818 while the p-value is 0.86, which is greater than 0.05. Therefore, the chi-square deviance test fails to reject the

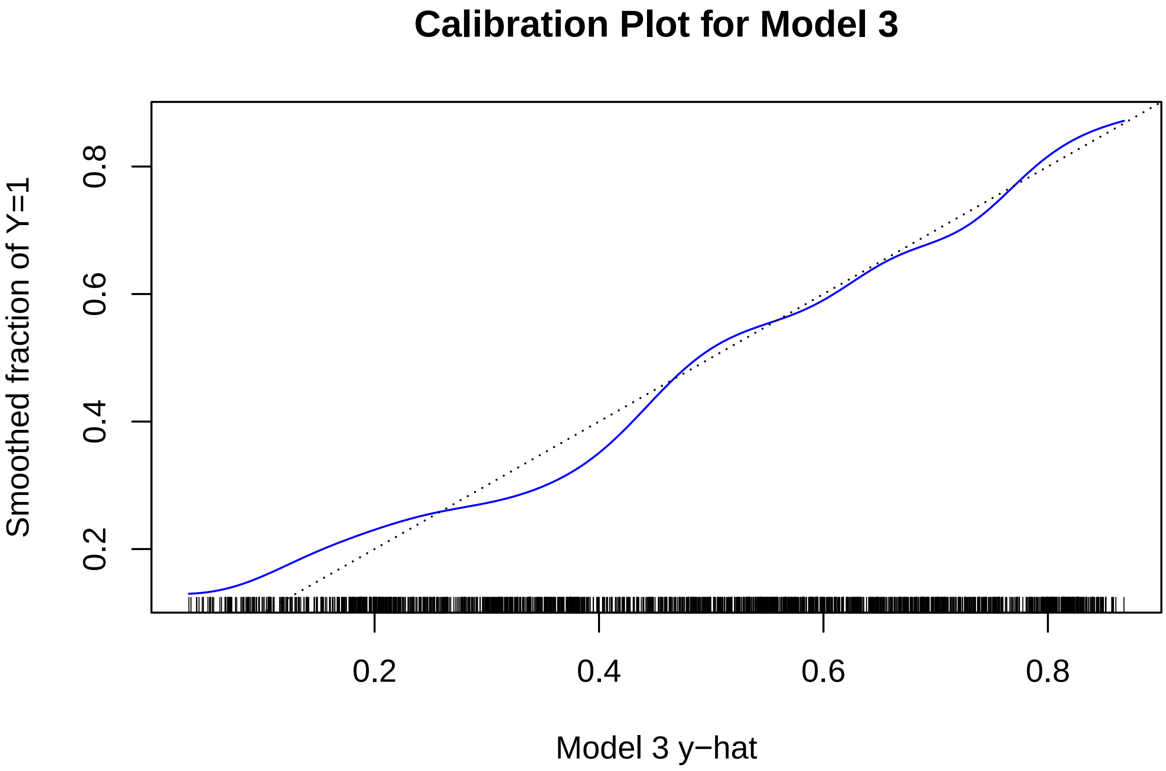


Figure 5: **(5)** Model 3 Calibartion Plot

null hypothesis that model 3 is an improvement of the simpilfied version without interaction term. This means the interaction terms are not useful in predicting of `HadExam`.

(3) Following this, we decided to modify model 3 by removing the interaction terms from the model, call model 4. The sample ratio between the odds of having a checkup for people with the *most* belief in the quality of information (rating each item 5) and the odds for those with the *least* belief in the quality of information (rating each item 1) is 1.508. Because model 4 does not include interaction terms between health insurance and any of the quality of information predictors, this ratio does not depend on whether each person has a health insurance or not. **(4)** The 95% confidence intervals for the odds ratio is (0.1244, 18.2906). This means we are 95% confident that the true ratio between the odds of checking up given most belief and the odds of checking up given least belief in quality of information of check up lies with in this range.

```
## Waiting for profiling to be done...
```

Conclusions

(1) To answer our original research questions, from our final model, model 4, the factors that appears to make a person less likely to get a check up are Jobstt_student, Jobstt_unstable, Wsttimeyes, NotImpyes, SuitFreq18m, SuitFreqg18m, ImpressInfo. Of these factors, the statistically significant factors are Jobsttstudent (p = 0.0009), Wsttimeyes (p = 2.17e-5), NotImpyes (p = 3.09e-14) and SuitFreqg18m (p = 1.86e-5), and SuitFreq18m (p = 0.02). From the analysis, being a student increase the log odds of not checking up by 0.84, believing the check up is a waste of times increase the expected log odds of not checking up by 0.43, believing the check up is not important increase the expected log odds of not checking up by 0.77, and believing the suitable check up frequency is greater than 18 months increases the expected log odds of not checking up by 1.11.

(2) The evidences do not suggest the quality of information is an important predictor of check-ups, because none of all of quality of information predictors has p-value lower than 0.05, thus is not statistically significant. This statistical insignificance does not depend on whether the person has health insurances either as their interaction terms are also statistically insignificant. People who view medical check ups as not important or waste of time do not value medical check ups, and thus do not frequently come. Lastly, people who believe a suitable check up frequency is greater than 18 months are less likely to have a check within the past 12 months, because their people are more likely to prefer to wait up to their suitable check up time before going for a check up, which can be any number of months greater than 18. It is interesting to see no statistical significance between checking up in the past 12 months and the quality of information from the medical check up and whether or not it depends on the person having health insurance. This finding suggests that there is no association between increasing the quality of information or promoting quality of information on the rate of people doing check ups. A possible explanation is because people do not value the quality of information in the check up.

(3.) A limitation on this analysis is the study is an observational study on patients, and was not experimentally designed to test the effect of each predictor on whether the patient has check-up while control the different predictor. This will allow us to make causal inference on the predictor's impact on our response. Another limitation is we can not randomly assign different predictor values (for example assigning job stability for each patient) to each patient, which helps make the predictors independent from each other, and also allows causal inference in our analysis.