

# Health exams in Vietnam Data Analysis Report

*Qiao Su*

*19/11/2019*

## Executive Summary

Medical care for serious diseases can be very expensive, and in countries without established socialized medical systems it can impose serious hardships on patients. Ideally, patients would get regular check-ups (or “general health examinations,” GHEs) so that serious conditions could be detected and treated early, before they cause serious problems. This would save money and improve public health.

However, there are many possible obstacles to getting everyone to go to regular check-ups. They might be too expensive, or too difficult to schedule; some people may not trust

doctors or believe that check-ups have any value; or some people may have had bad experiences when they previously tried to get check-ups.

Public health researchers in Vietnam wanted to explore these reasons and determine

what obstacles prevented widespread use of regular check-ups. They conducted an interview survey in Hanoi and Hung Yen, Vietnam, by traveling to “secondary schools, hospitals, companies, government agencies and randomly selected households in Hanoi” and

interviewing people in person for about 10–15 minutes. This dataset contains the raw data from that survey, totaling 2,068 valid responses.

Table 2 describes some of the categorical variables in the dataset. Over half the sample ( $n = 1,059$ , 51.21%) had had a GHE less than a year ago. One of the most common reasons given for hesitating to have a GHE was that they are waste of time; nearly 52% of participants who were reluctant to attend GHEs mentioned this as a reason. Amongst those who were prepared to attend GHEs, the main reason given

If they experienced symptoms of ill-health the majority of participants would choose to go to a clinic (43.04%). Most respondents (86.32%) believed that a GHE should cost less than 2 million VND, indicating that reasonable pricing is a big concern for people in relation to periodic GHEs

If a healthcare app indicated symptoms of disease then 39.41% of participants would be willing to have a GHE.

## Introduction

In addition, a number of people remain skeptical about the value of health examination (GHE) programmes, either finding them costly and without benefit<sup>7–9</sup> or questioning their quality.

The dataset includes three categories of variables about the participants. The first category are demographics such as BMI, age, education and sex. The second category quantifies their attitude towards health such as whether they can basic medical equipment, and how much time the respondent spends on sports and physical exercise. The last category quantifies their attitudes relating directly to the GHEs such as their perceived ability of examiner and the perceived attractiveness of information they received in check-ups.

## Methods

### Removal of outliers and missing data

After importing the data, we found there was 90 cases with missing data out of the 2068 participants. The variables with missing data included numeric variables only: height, weight and the ratings of GHEs e.g. perceived timeliness of the checkup. We first discarded height and weight since these two correspond

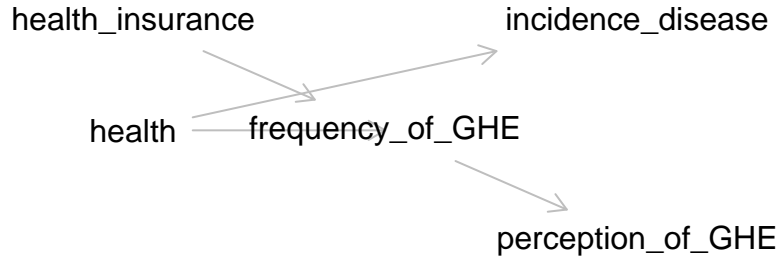


Figure 1: Causal diagram illustrates hypothesized relationships between checkups, disease and cost of public healthcare.

fully to BMI (see EDA). There seemed to be a possibility that the remainder missing cases were informative missing data, since the participants had answered everything else. Therefore we kept the other observations that had missing data.

We also looked for outliers in the data among the remaining variables. Only the Age and BMI directly inform us about the patient’s health so outliers in these variables are very important. Additionally, we removed outliers in the date to restrict participants from September to October 2016. This left 1941 responses.

## Exploratory Data Analysis

We may hypothesize that the respondent’s general health is inversely related to their incidence of disease. Furthermore their perceptions of GHEs may be approximately correlated with what value they have derived from the procedures in improving their health. Lastly the number of GHEs brings up the cost of healthcare.

### Univariate variable distributions

Data were collected in 2016 from participants from 13 to 83 years of age, on 31 separate dates in the year 2016. As seen in Figure 2 A-C, there are outlier values in the variables date, age and BMI. There are outlier dates in Figure 2 A which may be typos and fall outside of real dates e.g. ‘20169828’. Therefore we trimmed the few outlier dates which fall outside of September and October 2016. For age and BMI, we can see in Figure 2B-C that the data are upward skewed by a few outliers. We therefore removed data with age greater than 50 and BMI greater than 35.

BMI is an important indication of fitness and calculated as  $\text{weight (kg)} / [\text{height (m)}]^2$ . When we calculated the BMI from the height and weight we found that the quantities corresponded exactly, as seen in Figure 2 D. Therefore since BMI is more indicative of health, we discarded height and weight variables.

### Pairwise distributions

We checked the correlation among the numeric variables of the filtered dataset in Figure 3. We observed that there doesn’t appear to be collinearity among the numeric variables. Furthermore, age and BMI have a reasonable correlation. Additionally, we found that the ratings naturally grouped by correlation into responses concerning the quality of GHEs (Tangibles, perceived quality of tangible equipment and personnel to Empathy, perceived empathy of the staff) and about the type of information they receive during GHEs (SuffInfo, rating of the sufficiency, to Popular info, rating of the popularity of the information).

## Introduction

### Q1.

Overall, how do people rate the attractiveness, impressiveness, sufficiency, and popularity of information they receive in checkups? Give us some summaries of these variables, as well as variables like assurance, reliability, and empathy that tell us how well our doctors and nurses are doing, so we know how to improve.

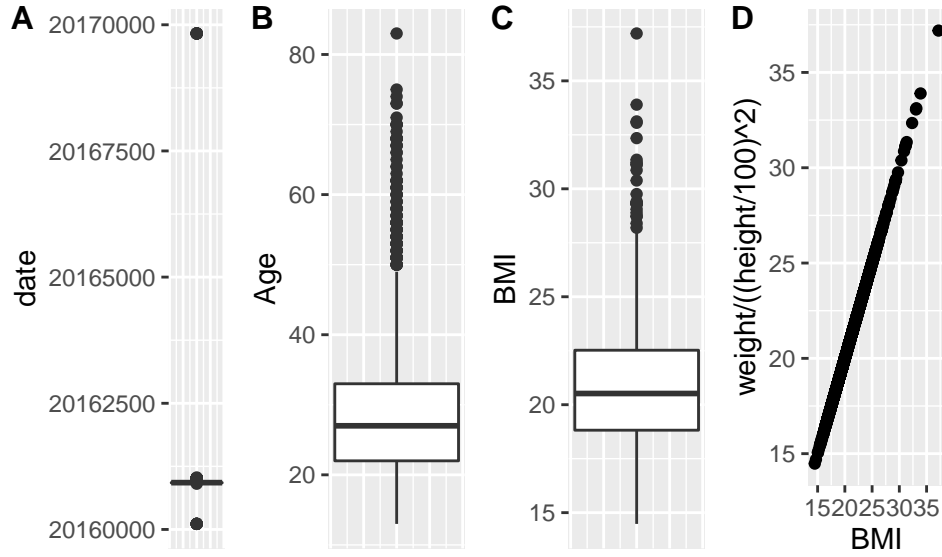


Figure 2: A-C) Distributions of important demographic variables in the raw data show outliers. We trimmed the outliers in the date, age and BMI. D) We can also see BMI is perfectly calculated from height and weight.

In order to summarize how respondents felt about checkups, we first looked at the distributions of the ratings they gave of the checkups. The distributions across the diagonal in Figure 3 show summaries of the main rating variables. Ratings of the type of information recieved during GHEs include their attractiveness (AttractInfo), impressiveness (ImpressInfo), sufficiency (SuffInfo), and popularity (PopularInfo) of information. These ratings are correlated. There are also ratings of the perception of quality of periodic GHE sessions which includes the tangibles (Tangibles), assurance (Assurance), reliability (Reliability), responsibility (Respon), and empathy (Empathy) variables.

We can see that the average rating about the type of information recieved in GHEs is significantly correlated and tends to score a 3 on the 5-point scale. As well, the

We were further interested in whether

## Q2.

**What factors make a person less likely to get check-up every twelve months? Find the most important factors that could help us design our advertising, and give us some measure of how important they are.**

We aimed to determine what factors make a person less likely to get check-up every twelve months by modelling the variable RecPerExam (the time since the respondent got an unprompted checkup). RecPerExam has four levels: less12 = less than 12 months, b1224 = between 12 and 24 months, g24 = over 24 months, unknown = respondent doesn't know. We first trimmed the cases that are 'unknown' since this is not an informative response, leaving 1467 cases. Since the response variable is composed of discrete events, we used a binomial glm.

We performed best subset variable selection among all of the variables available (except the participant id) in order to exclude variables without large effects in this model. First we picked how many terms we should have in the best predictive linear model by using 10-fold cross validation with best subsets regression and plotted the mean squared error in Figure 5. We used the bestglm package to perform the cross validation. This analysis shows validation error is lowest around 8 terms.

Next we fit the model to RecPerExam with the top XXX variables. We found the top

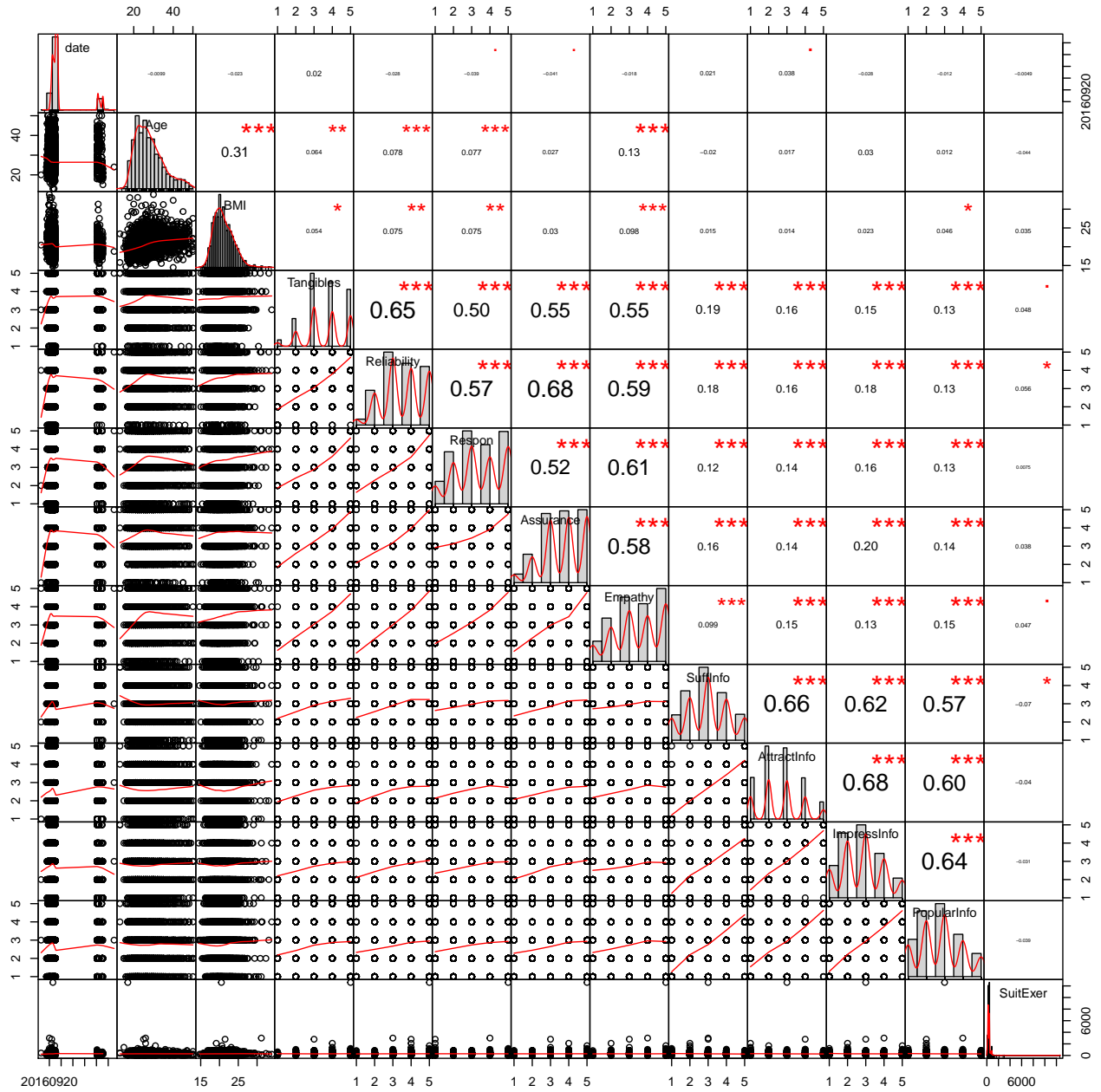


Figure 3: Pairwise correlations of numeric variables including their Pearson correlation coefficient.

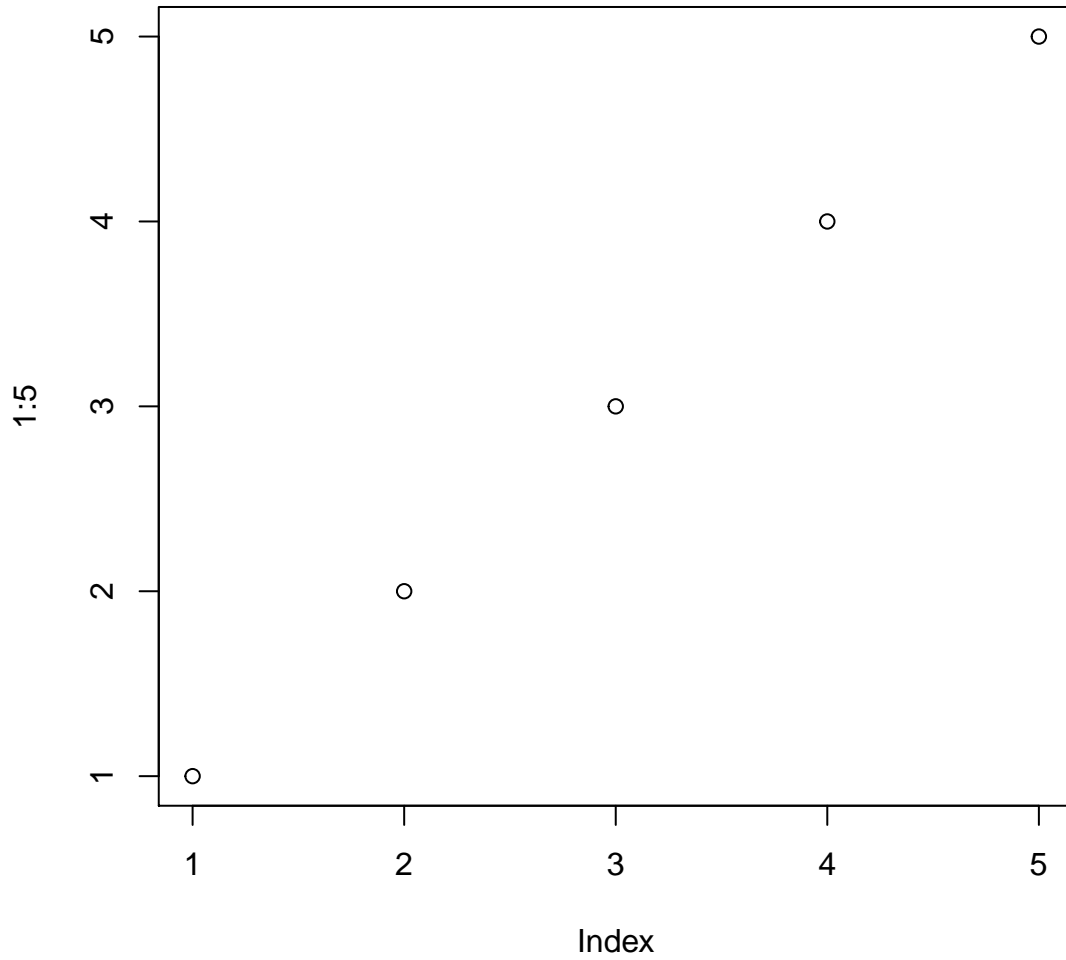


Figure 4: Diagnostic plots for the fit of model ‘views ~ posts + authors + offset(log(startdate))’. Multiple outliers are apparent. These outliers occurred in topics belonging to Miscellaneous sciences, Speculations, and Medicine.

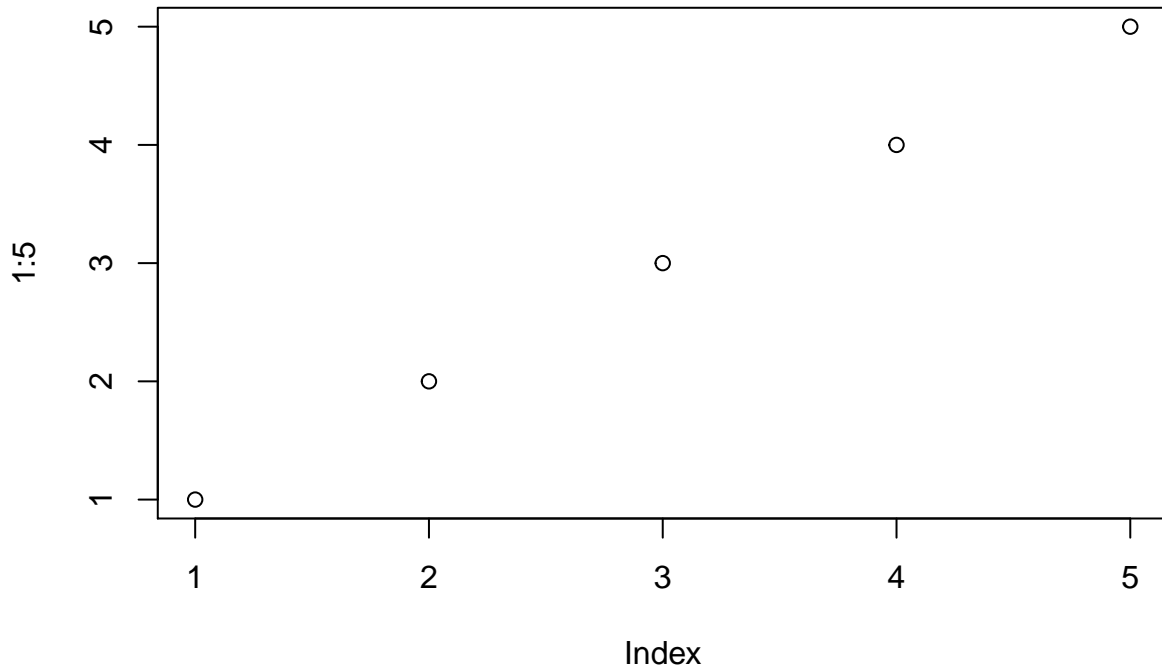


Figure 5: Cross validation error for each number of predictors

### Q3.

**Can we predict which people would be easiest to convince? That is, some people might be on the edge, and would get an exam with a little extra push; some people are very determined and would not get an exam no matter how hard you try. Using a classifier, can you find the patients who haven't gotten an exam but are most like other patients who have? Be sure to tell us how well your classifier works, so we know whether this is reliable.**

In order to predict which people would be easiest to convince to take a GHE every 12 months, we again build a model of the time since respondent last visited a doctor for a check-up (RecPerExam). One of the four responses is 'unknown', meaning the respondent doesn't know when they last visited a doctor for a check-up when not prompted by a specific illness. However, we can try to predict for the respondents who answered 'unknown' whether they would fall into one of the other three categories (less12 = less than 12 months, b1224 = between 12 and 24 months, g24 = over 24 months).

We first set aside the cases for which RecPerExam is 'unknown'. Working with the remaining 1467 cases, we built a XXXXX model to predict RecPerExam. Since here we are interested in predicting RecPerExam, rather than interpretation, we decided to include all of the potential predictor variables (other than id). We randomly shuffled the rows and split the data into a training and test set. The performance of the test set as seen in Figure XXXXXXXX

Next, we used our classifier to predict the status of the unknown cases. We were interested in characteristics of the patients who haven't gotten an exam in the last 12 months, but are most like other patients who have.

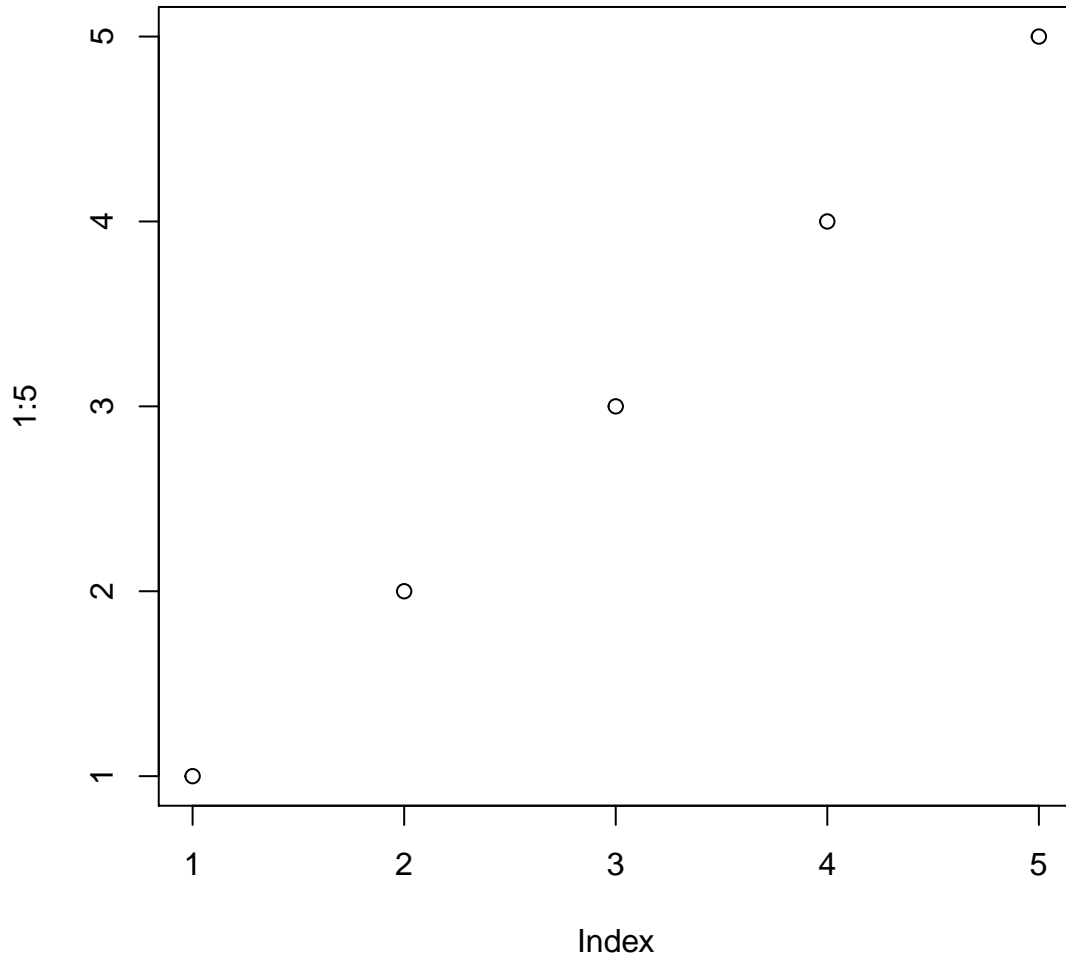
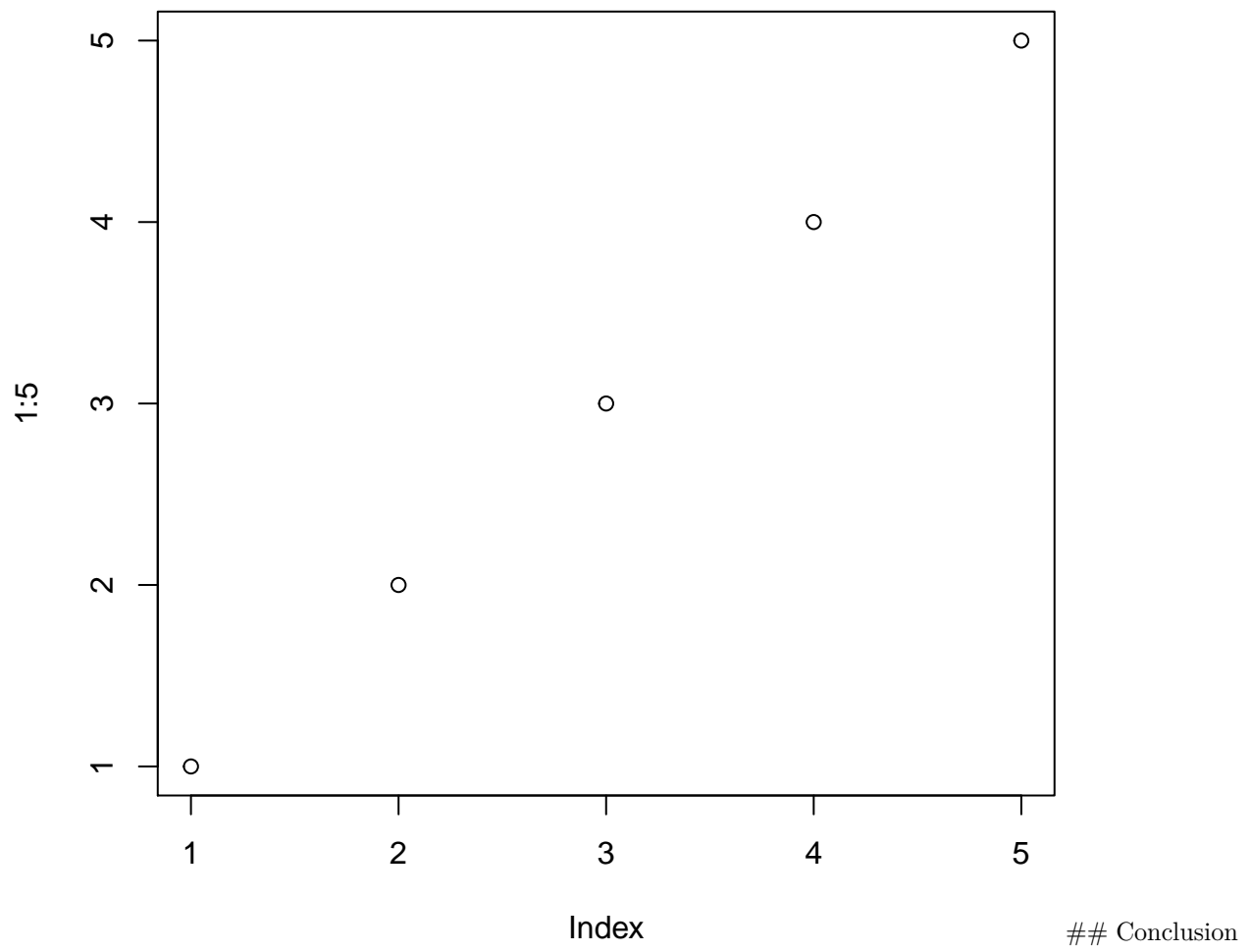


Figure 6: Diagnostic plots for the fit of model ‘views ~ posts + authors + offset(log(startdate))’. Multiple outliers are apparent. These outliers occurred in topics belonging to Miscellaneous sciences, Speculations, and Medicine.



Bibliography