36-402 DA Exam 2

Chae Yeon Ryu (chaeyeor)

4/24/2020

Introduction

The Vietnamese Ministry of Health is interested in knowing how they can increase the number of people who get annual health exams. To be specific, they want to know the reasons that people are not willing to sign up for exams and how to fix such issue. They want to examine whether people's willingness is related to the cost of health exams, quality of information received, or the value of health service provided. This study utilizes the data from 2,068 interview surveys in Hanoi and Hung Yen, Vietnam, from secondary schools, hospitals, companies, government agencies and randomly selected household.**(1)** The goal of this analysis is to quantify the statistical relationship with these variables and see how the Vietnamese Ministry of Health can promote the annual health exams more effectively. **(2)**The results of the analysis provide strong evidence for a relationship between HadExam and job status, Waste of Time, Not Imp, Suit Frequency, and the Health Insurance after improving the model. It turns out that the affordability may be the reason for less likeliness to get check-up every 12 months as well as the health insurance, as observed in the difference between job status. People are at most twice more likely to get check-ups done when they have strong belief in the quality of information than those with the least belief. This result is expected to be the same for both people with and without health insurance. The quality of information is not expected to be the most significant factor; rather, it is suggested that the Ministry of Health focuses on advertising the value of check-ups.

Exploratory Data Analysis

The data used consist of 2,068 interview surveys with measured scores and values that are divided into three categories: Demographic Variables, Value and Quality of Medical Service, and Quality of Information. Demographic Variables include age, sex, jobstatus(jobstt), BMI, and Health Insurance(HealthIns). Value and Quality of Medical Service variables include WasteTime(Wsttime), WasteMoney(Wstmon), FaithInQuality(Lessbelqual), NotImp, Tangibles, Empathy, and SuitFreq. Quality of Information variables include SuffInfo, AttractInfo, ImpressInfo, and PopularInfo.

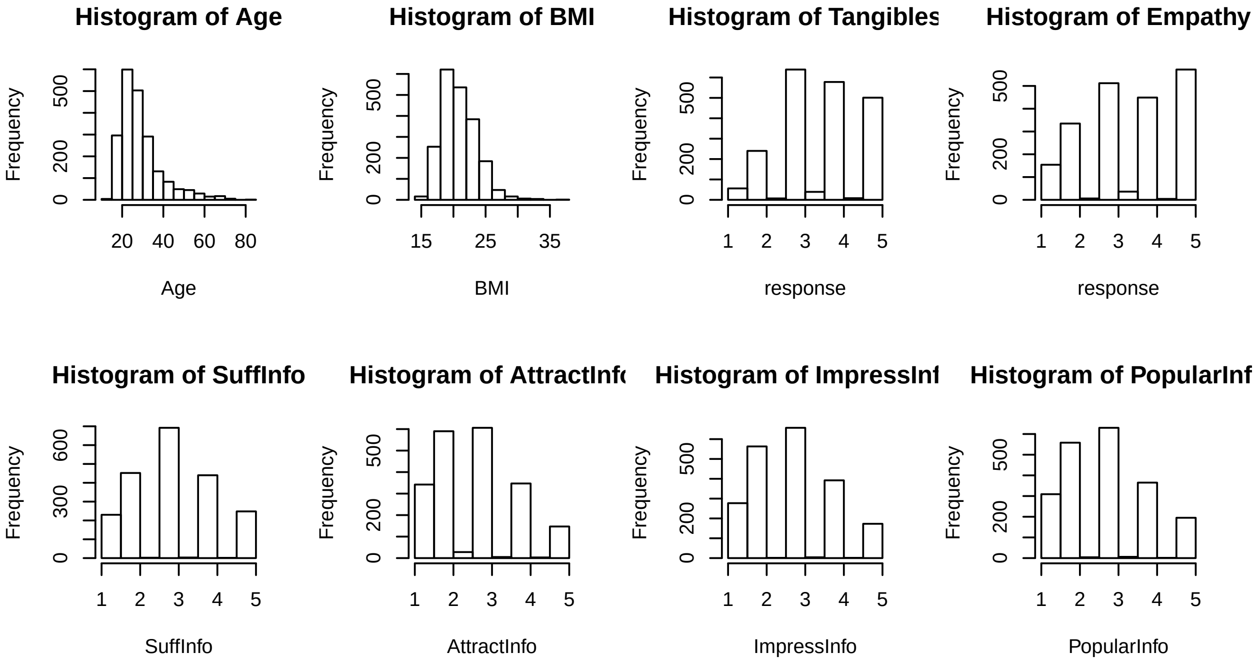

Figure 1: Histograms of Continuous Independent Variables.

(1) Variables such as age, BMI, Tangibles, Empathy, and all the quality of information variables such as SuffInfo, AttractInfo, ImpressInfo, and PopularInfo are considered continuous independent variables while any other variables are considered categorical. Looking at Figure 1, which shows the histograms describing the distributions of each of these variables, we see that age and BMI have right-skewed distribution while Tangibles and Empathy are slightly left-skewed. Other quality of information variables display very approximately bell-curves. From the summary of the data, it is noticed that the mean age of the respondants is around 29, and the respondants were more females than males and mostly have stalbe jobs or students. The mean BMI is 20.85 which is considered normal, and most of respondants have health insurance. On average, respondants think that the information they receive in check- ups is of moderate quality, and most of respondents believe taht check-ups should be done every 6 months or 12 months. Lastly, respondents believe that the value of check-ups and the quality of the medical service are slightly above the moderate level on average. **(2)**HadExam, which is a response variable, takes 1 if the respondent had a check-up in the past 12 months, and 0 otherwise. Its median value is 1, while the mean value is 0.5121. Since it takes a binary, boolean output, HadExam follows a Bernoulli Disbribution.

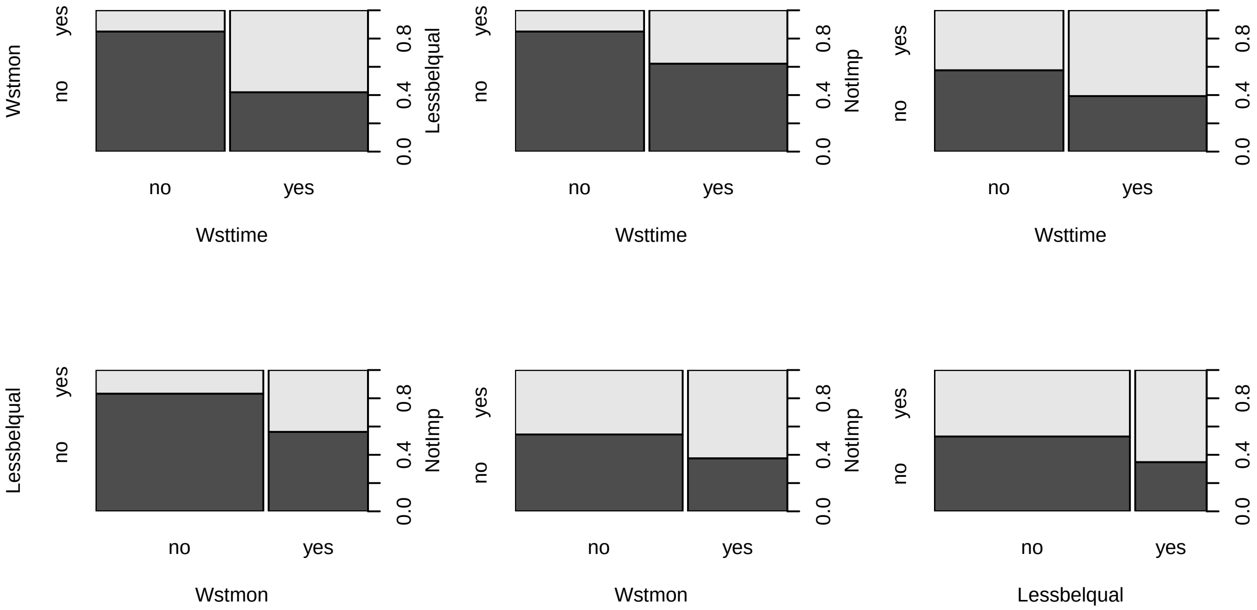

Figure 2: Bivariate relationships between Value and Quality of Medical Service Variables

(3) Figure 2 displays the bivariate relationships between the scores people rated for the Value and Quality of Medical Service. It is observed that a lot of people who believe that check-ups are a waste of time also think that check-ups are a waste of money and have little faith in the quality of medical service.

(3) Furthermore, Figure 3 suggests that people who give a high score in one of the quality of information variables tend to give higher scores in other quality of information variables.

(3) Looking at Figure 4, we see that people who believe in the values of check-ups are more likely to have had check-up done in the past 12 months. In addition, there is a high proportion of people who think that check-ups should be done every 18 months or less often than every 18 months and also did not get a check-up done in the past 12 months. People with stable jobs have a higher proportion of getting check-ups done while students are the least likely to get it done. Lastly, there seems to be a no clear relationship between HadExam and the quality of information since the scores are about the same between the groups of people who had the check-up and did not have the check-up. **(4)**Hence, it is expected that the quality of information scores won't help much in understanding the variation in HadExam, while the

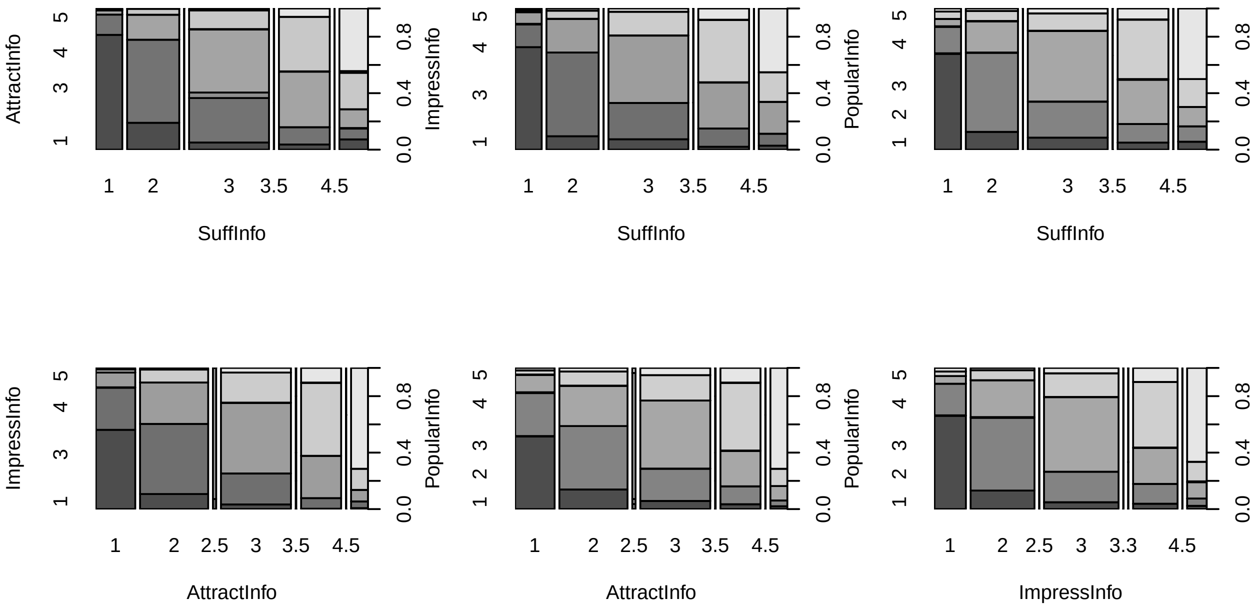

Figure 3: Bivariate relationships between Quality of Information Variables

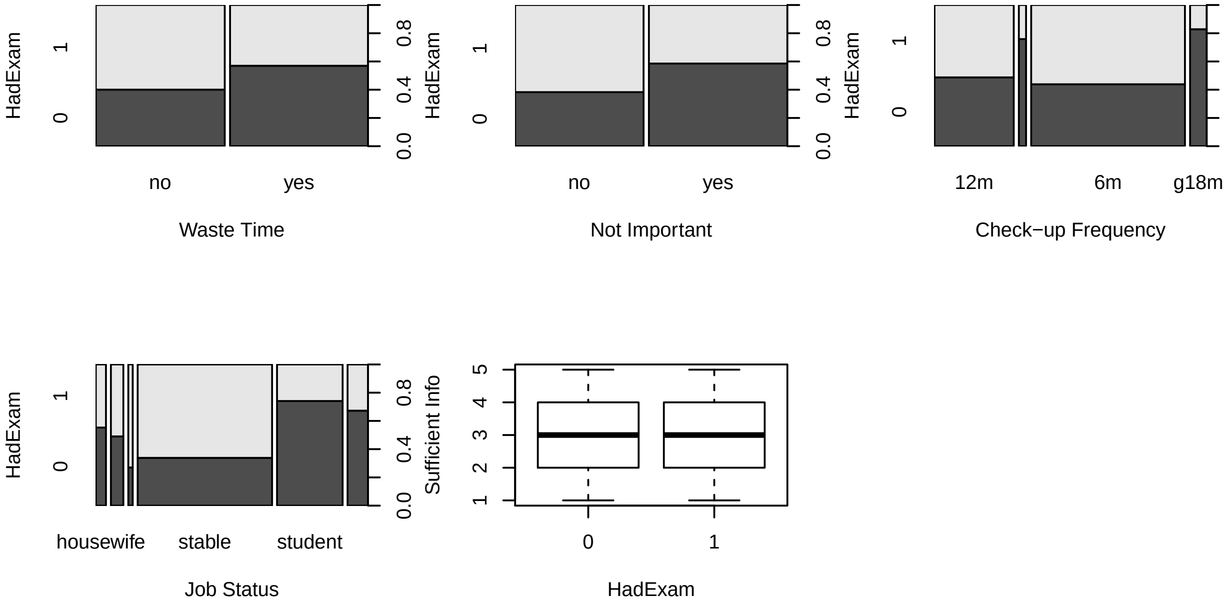

Figure 4: Bivariate relationships between HadExam And Some Independent Variables

variables of Value and Quality of Medical Service are expected to. As we found that some of those variables might be repetitive information, it is expected that only few of them will be used the explain the relationship between the response variable and the Value and Quality of Medical Service. Last but not least, HadExam is expected to vary depending on the job status as we discovered. It is possible that health exams are not affordable to students while people with stable jobs are provided an exam by the company or the insurance. Hence, job status will be an factor explaining people's response to the annual health exams.

Initial Modeling and Diagnostics

(1) The first linear model has factors related to demographics and the value and quality of medical service as its independent variables. These variables are Age, Sex, Jobstt, BMI, Wsttime, Wstmon, Lessbelqual, NotImp, Tangibles, Empathy, and SuitFreq. **(2)**Using a stepwise selection procedure, we came up with a second linear model with the lowest AIC value, which was 2467. As it was expecetd in the EDA, some of the repetitive variables in the value and quality of medical service are removed, and it turns out that demographic variables except the job status are not useful in understanding our response variable, HadExam. Thus, model 2 has Jobstt, Wsttime, NotImp and SuitFreq as its explanatory variables. **(3)**Then, we came up with a third model that adds health insurance and the quality of information variables to model 2. It also includes interactions between health insurance and the quality of information variables in order to check if the quality of information variables have different associations between patients with and without health insurance.

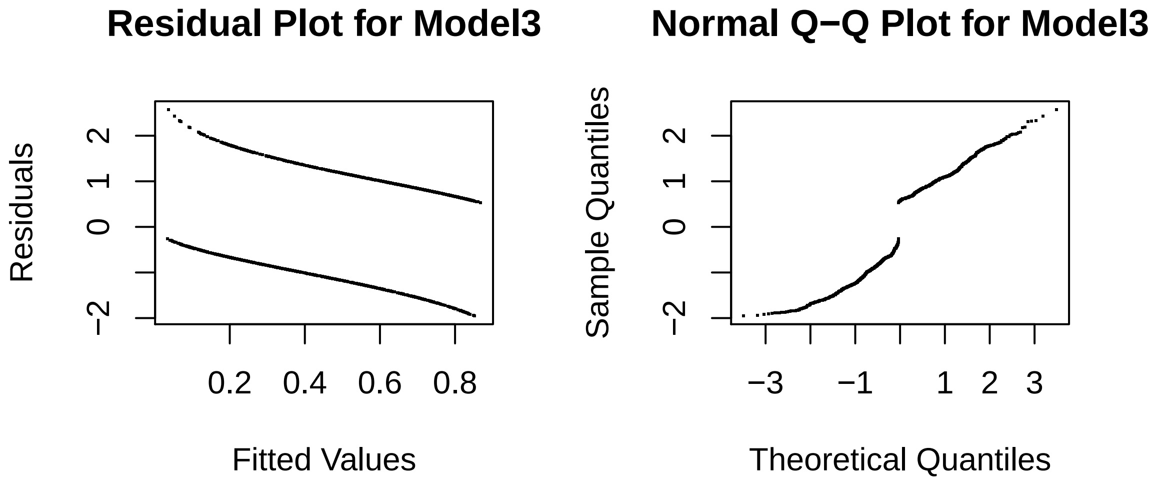

Figure 5: Diagnostic residual plot for model3.

(4) For the goodness of fit of model 3, several tests and analyses were conducted. With the anova test, we assume that factors are randomly sampled, independent and come from a normally distributed population. We also assume that variances are unknown but equal. First, if we look at the residual plots for model 3 from Figure 5, we see that there are two lines from people who took the exam and people who didn't. We see that there's a decreasing trend in both, indicating that a linear regression may not be sufficient enough to explain the relationship between the independent and dependent variables and that the variation may not be constant. On the other hand, looking at the normal Q-Q plot, we see that a lot of points are clustered on the 45 degree line. Conducting a deviance test between model 2 and model 3, we find that the p-value is 3.471e-06, which is very close to 0. This implies that model 3 improves significantly on model 2. On the other hand, Boostrap p-value is obtained as 0.732, which implies insignificance. Based on the analysis of histogram of the bootstrap test statistics, it was found that the distribution is not close to chi-squared distribution, which was what we assumed for deviation testing. Hence, the nominal asymptotic chi-squared distribution of the deviance test statistic under the null hypothesis might be incorrect. The global goodness-of-fit gives a p-value which is 6.685647e-08, implying that model3 does not appear to fit well and that the saturated model may be correct. To be specific, GLM with the covariates included is not sufficient to explian the data. This is also based on the assumption that Dev ~ chi-squared(n-q) approximately.

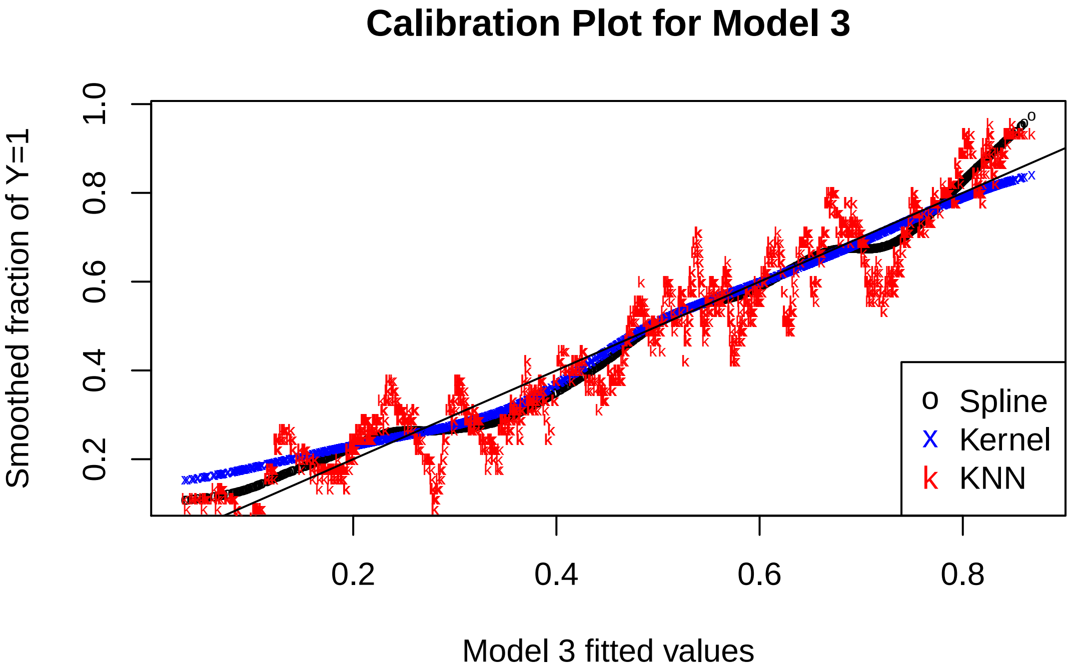

Figure 6: Calibration Plot for Model 3.

(5) On the other hand, we can see that model 3 is well-calibrated, meaning that its predicted probabilities match the observed proportions of outcomes in the data. If we look at Figure 5, we can see that the estimated probabilities are close to the fractions of Y = 1 cases with a spline, a kernel regression, and K-nearest neighbors. Comparing the average distances between the smooth curves and the y = x lines in the plot, we get that the values are 0.05283279, 0.01605231, and 0.02809748 for three smoothings, which are pretty close to 0. Hence, model 3 is well-calibrated. However, it is possible that removing variables that are not helpful in explaning the response variable, such as health insurance and quality of information, may be useful for the model to fit better.

Model Inference and Results

(1) The interaction terms in model 3 controls for the different associations between patients with and without health insurance. Looking at the summary of model 3, it's observed that the health insurance and the quality of information are not the significant factors, as well as the interaction terms between health insurance and the quality of information. This means that health insurance and the quality of information don't help explain the variation in our response variable, HadExam. It also indicates no relationships between the health insurance and the quality of information. Thus, the Ministry of Health should not consider a marketing campaign focusing on the quality of information patinets receive in check-ups as well as thier relationships with the health insurance factor based on model 3.

In order to see the interaction terms are significant, F-test is conducted to compare the orginal model3 with the adjusted model3 without interaction terms. **(2)** The null hypothesis states that the adjusted model fits the data as well as the original model3 with interactions. The alternative hypothesis states that the original model3 taking into account the interactions fits better. In order to do F-test, we assume that the data are normally distributed and that samples are independent from one another. Furthermore, we assume the homogeneity of variance. **(2)** The F-statistic is about 0.3205, indicating that there is no significant difference between the means of the samples being compared (variation among group means is also expected to be insignificant). Hence, we can't reject the null hypothesis that the reduced model is as good as the original model 3. This means that the interaction terms between health insurance and the quality of information variables don't help explain the variation in the response variable, HadExam. The p-value from deviance, which was 0.8644558, also supports that there is no evidence of interaction. Thus, Ministry of Health doesn't need to consider the relationship between the health insurance and the quality of information scores when advertising the check-ups. After removing the interaction terms, health insurance factor became significant.

(3) Meanwhile, the ratio between the odds of having a checkup for people with the most belief in the quality of information and the odds for those with the least belief in the quality of information turns out to be about 1.39. Since it's greater than 1, strong belief in the quality of information is considered a risk factor for whether people get check-ups done or not. This means that people with the strongest belief in the quality of information are with higher odds of getting check-ups done in 12 months than those with the least belief. As the health insurance was not a significant factor in the model, it is believed to have similar effects for all cases. **(4)**The 95% confidence intervals for the odds ratio is, with the round up to two decimal places, [0.93, 2.07]. This implies that people with the most belief in the quality of information get the check-ups done in 12 months 1 to 2 times more than those people with the least belief in it. It it reasonable given that people who strongly believe in the benefit of health check are more likely to get it done.

Conclusions

(1) The results of the analysis provide strong evidence for a relationship between HadExam and job status, Waste of Time, Not Imp, Suit Frequency, and the Health Insurance after improving the model. Based on the EDA and the model analysis, the Ministry of Health should appeal the affordability of health care to the students becasue they seem to be hesistant getting the check-ups done. **(2)**It may be due to the fact that students don't have an insurance provided by companies and the cost of check-ups may appear expensive for them. It's also evident in high proportion of people with stable jobs getting check-ups done. **(1)**In addition, they should consider advertising the value of check-ups because people with low trust in the values of check-ups are less liekly to get check-ups. As the odd ratio suggests, people are at most twice more likely to get check-ups done when they have strong belief in the quality of information than those with the least belief. On the other hand, there are some liminations in this analysis. **(3)**First, it is still unknown as to why exactly job status matters. We assumed some of the possibilities (i.e. students with no insurance and seemingly expensive costs of health check-ups), but there may be other factors affecting the difference. Since these factors are not taken into account, it is suggested to account for other possible factors to decide on the extent to which job status matters. Some of the factors may also be confounding. Furthermore, we don't know whether the explanatory variables are causing the HadExam or the fact that people have/have not had check-ups make them score the survey questions differently. It may be useful to further examine such a causal inference. Lastly, some of the analyses conducted are based on the assumptions that may not be satisfied; it is suggested that we further examine whether our model is correct (i.e. we have a right distribution for the noise and a right shape for the regression function for parametric bootstrapping) and use the proper methods.