|  |
| --- |
| Data Science with R II (JEM220) – Charles University Prague, FSV |
| Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients |
| by Anchana Khemphila & Veera Boonjing |

|  |
| --- |
| Analysis and improvements by Erik Nemcik & Paul Mainka  17.5.2020 |

Overview

In their study Anchana Khemphila and Veera Boonjing compare the performance of logistic regression, decision trees and artificial neural networks in classification problems. For this purpose, the authors employ a data set on the development of heart disease among individuals.

In order to evaluate the performance of the three different methods, the authors use measurements of area under the curve (AUC), sensitivity, specificity, accuracy and the error rate. Since artificial neural networks have the lowest error rate and the highest accuracy in their estimations, they conclude that artificial neural networks is the most suitable of the three classification techniques assessed for this data set.

# Data

The data set used by Boonjing and Khemphila includes 303 individuals, 13 explanatory variables and the dependent variable whether the patient has developed heart disease or not. The data set is quite balanced with 165 patients having developed a heart disease which corresponds to approximately 54% of all the patients. Of the 13 explanatory variables, 8 are factors and 5 are continuous variables.

Description of the eight factors:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Name | Meaning | Different Levels | Frequency of occurrence |
| 1 | Sex | Patient gender | Female | 96 (32%) |
| Male | 207 (68%) |
| 2 | Chest Pain Type |  | Angina | 143 (47%) |
| Abnormal | 50 (17%) |
| Nonanginal pain | 87 (29%) |
| Asymptotic | 23 (8%) |
| 3 | Fasting Blood Sugar | Is fasting blood sugar less than 120? | True | 45 (15%) |
| False | 258 (85%) |
| 4 | Resting ECG | The electrocardiogram (ECG) measures the heart’s electrical activity, and a resting ECG is administered when the patient is at rest. | Normal | 147 (49%) |
| Abnormal | 152 (50%) |
| Left ventricular hypertrophy | 4 (1%) |
| 5 | Induced Angina | Does the patient experience angina as a result of exercise? | True | 99 (33%) |
| False | 204 (67%) |
| 6 | Slope | Slope of the peak exercise ST segment. | Up | 21 (7%) |
| Flat | 140 (46%) |
| Down | 142 (47%) |
| 7 | Number Coloured Vessels | Number of major vessels coloured by fluoroscopy. | 0 | 175 (58%) |
| 1 | 65 (21%) |
| 2 | 38 (13%) |
| 3 | 20 (7%) |
| 4 | 5 (2%) |
| 8 | Thal | Thalassemias, blood disorders characterized by decreased hemoglobin production | Normal |  |
| Fixed defect |  |
| Reversable defect |  |

Description of the 5 continuous variables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # | Name | Meaning | Min | Median | Mean | Max |
| 9 | Age | Age in years | 29 | 55 | 54.37 | 77 |
| 10 | Blood Pressure | Resting blood pressure upon hospital admission | 94 | 130 | 131.6 | 200 |
| 11 | Cholesterol | Serum cholesterol | 126 | 240 | 246.3 | 564 |
| 12 | Maximum Heart Rate | Maximum heart rate achieved. | 71 | 153 | 149.6 | 202 |
| 13 | Ole Peak | ST depression induced by exercise relative to rest. | 0 | 0.80 | 1.04 | 6.20 |

# Theory/Model

The authors develop a simple supply and demand side model for health care.

The demand side is determined by the utility of patients which depends on the health status , preference for health care and healthcare utilization :

Hereby a higher represents worse health and a greater preferences for more aggressive treatment. The utility function is maximized by maximizing while minimizing , ergo maximizing by increasing . Note that people with preference for more intensive treatment will receive a disproportional utility gain from additional healthcare utilization compared to people with less preferences for aggressive care. Minimizing translates to that utilization of healthcare should match the required healthcare according to the health status[[1]](#footnote-1). The utility function is maximized at:

Thus, healthcare utilization is optimal when it matches health status plus preferences.

Physicians believes and their private costs shape the supply side of Finkelstein et al.’s model. Doctors maximize the observed utility of their patients according to what they believe is the appropriate treatment minus their private costs.

whereby:

represents more aggressive treatment methods here.

Finkelstein et al. apply this theoretical demand and supply model to the data by a regression. In addition, they provide an overview of what private costs and believes of doctors and the health status and preferences of patients reflect[[2]](#footnote-2).

The presented mechanism of the model of supply and demand side are plausible and consistent with the literature. Overall, the theoretical part is a strength of this paper. The reason is not because the model is particularly pioneering or sophisticated, but rather that the authors test the implications of the model directly on the data. Thus, they not only provide empirical results but also provide a theoretical basis for them.

# Identification Strategy

Since supply and demand for health care jointly determine the level of utilization, the key to differentiating between patients and location effects are patients who move during the observation period. Put simply, if the health care utilization of a person who has moved perfectly adapts to the average utilization of the destination, it can be concluded that health care utilization is determined only by the supply side, whereas if the utilization does not adapt at all after the move, it can be concluded that the patient effects determine everything.

Finkelstein et al. employ two main regressions:

1. Their baseline regression including movers and non-movers:

Where the log utilization per patient i at place j and time t is regressed on patient fixed effects , place fixed effects , year fixed effects , and dummies for age bins and fixed effects for movers . With some basic maths the variation of utilization () can then be additively decomposed into variation due to patient/demand side effects ( + ) and by place/ supply side effects ( + ).

1. An event study including only Movers

Where is the log utilization per patient at time t, is the single patient fixed effects (= ), is the average log utilization in the destination minus the average log utilization in the origin, the coefficient measures therefore the change in utilization due to the move. In this specification captures the place effect share directly as it weights the effect of the difference between the place of origin and the destination.

Using fixed effects for patients, places, years and for movers enables Finkelstein et al. to estimate observable and unobservable characteristics of individuals, years and places accurately and unbiased (at the expense of degrees of freedom). Furthermore, they perform plenty robustness checks (see chapter "Threats to internal validity") which indicate that the authors' identification strategy proved successful.

# Results

Finkelstein et al. arrive at the result that, on average, the patient effect and place effect each account for approximately 50% of the variation in log utilization of health care. This result is valid in both regression specifications, the baseline regression and the event study, and across their robustness checks. As the difference in average log utilization between origin and destination increases, i.e. with increasing importance of the place effect, the share of variation due to the place effect increases. For example, the place share explains 70% of the variation in log utilization between Minneapolis (the place with the lowest average utilization) and Miami (with the highest average utilization). In addition, they find that the patient share varies with the degree of decision-making involvement of the patient, for instance, the patient share is only 0.09 for diagnostic tests but 0.71 for emergency room visits. Moreover, Finkelstein et al. estimate that about 25% of the geographical variation in utilization can potentially be attributed to observable health of the patients. However, they cannot answer whether the remaining part of the patient component reflects preferences or unmeasured health.

The authors present their results in clear tables and graphs and also provide a detailed discussion and interpretation of their findings. The results hold over a variety of different specifications and generally seem to be coherent with intuition in size and direction. Furthermore, the results are generally consistent with the previous literature (see page 1686) in so far as they find that the supply side is an important driver of differences in utilization. However, Finkelstein et al. suggest that the demand side is somewhat more important than what the majority of the previous literature advocated.

# Threats to internal validity

Finkelstein et al. carry out a large number of robustness checks. These include:

1. Verifying that their results are not driven by different utilization trends among those moving which are systematically linked to their origin and destination.
2. Showing that patients and location effects are time constant.
3. Allowing for varying place effects for different quartiles of patient age.
4. Excluding all observations for patients who exit or enter the sample due to death or HMO status.
5. Using different market definitions and/or including only movers who cross state lines or census region boundaries.
6. Employing various other robustness checks with alternative definitions of movers, different dependent variables, excluding non-movers, dropping age and relative year as covariates, and excluding moves to Florida, Arizona, and California.

In summary, it can be said that Finkelstein et al. are well aware of the possible pitfalls of their regressions and have accordingly performed a multiplicity of robustness checks[[3]](#footnote-3). To my perception, they exercised due diligence and considered every aspect. The validity of their results can therefore be assessed as trustworthy and robust.

External validity

The authors conclude that no clear conclusions can be drawn from their results on welfare, since variation on the supply side does not necessarily have to be inefficient, since at least part of the supply can be attributed as endogenous responses to the health status and preferences of the population. Conversely, they argue that variation on the demand side may not be fully efficient because patient demand could include misinformation or behavioural bias. Instead, they point out that some other of their findings may have implications for possible policies, such as that doctors' treatment practices tend to have a direct rather than a gradual effect, while patients' preferences tend to adjust slowly, if at all.

In general, Finkelstein et al. are aware that their results are hardly generalizable and that there are almost no policy recommendations from their study to deduce. There are two main causes for this: The first is that they do not answer the question whether supply-side variation is in fact ineffective and the second reason is that their data set is based on a very specific population group (as I explained in more detail in the chapter "Data"). The fact that Finkelstein et al. are conscious of the issue is certainly good, but unfortunately it does not solve the fundamental problem. The resulting low external validity is definitely a weakness of this study.

Improvements to the Study

In summary, Finkelstein et al. have done an excellent work with regard to what they have done. They provided a simple and effective way for identification and to distinguish supply and demand effects, they understood the limitations of their approach and data, and they performed a considerable number of robustness checks. However, what remains a significant shortcoming is that their results cannot be fully generalized and that they are unable to draw conclusions about the welfare implications of the variation in utilization originating from the supply and demand side. Harshly speaking, the most thorough work is of little use if no proper conclusions can be drawn. Accordingly, I have two main suggestions for improvement:

1. Ideally, a data set that enables more general conclusions to be drawn should be employed. In the best case, this data set would correspond to a cross-section of the population, i.e. both demographically and from an insurance environment perspective. It would also be interesting to conduct a comparable study using data from another country, since the USA, with its highly privatized health care system, is quite unique compared to most other (developed) countries that tend to opt for various forms of universal health care. The prerequisite for this is of course the availability of an adequate data set.
2. Finkelstein et al. should have attempted to provide an answer to the question of whether the supply effects on health care utilization are effective or not. Although this was not the primary focus of this study, their motivation is based on the idea that more aggressive treatments (which are more expensive) may be ineffective. Their dataset includes information on the diagnosis and health status of patients and the treatment method. Since the data is available over several years, it would have been possible to investigate the success of different treatments. There are certainly better experimental designs and data for such a study, but Finkelstein et al. could have made a contribution to the question on whose answer this study here is based on.

In addition, I would like to make one more criticism, although it has less relevance. The readability and comprehensibility of the study is generally quite poor, especially in the first half (from the "I. Introduction" to "IV. Main Results: Patient versus Place"). The reason is that the authors incorporate a wealth of information in their text, but the text is not properly arranged thematically. For example, Finkelstein et al. present on pages 1687 and 1688 that patient preferences and health status as well as private costs and treatment preferences of doctors influence health care utilization, then describe their empirical models and the additive decomposition of variance before returning to preferences and etc. and discuss how they translate to reality. There are other examples of this "fragmentation" of topics, giving the text a somewhat erratic quality. As a result, the reader has to assemble the information for one topic from different sections of the text, which is cumbersome given the vast amount of information. Naturally, this is a purely subjective criticism, as there are varying preferences for writing styles. However, I personally had to read the study more than once to fully understand it. While the content should be the priority for a scientific study, it certainly benefits greatly from a good form. In the end, a flawless paper should be both factually correct and comprehensible to the reader, and in the best case also to readers who are not experts in the particular field.

# References

A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krackow, 2010, pp. 193-198, doi: 10.1109/CISIM.2010.5643666

1. In Erik’s and my presentation, I commented that the definition of the utility function is inappropriate, as could be minimized by increasing (ergo worse health). I abandon this representation here, because I realized my mistake, that is maximized by, which is conditional on and . So only can be adjusted according to the health status and not the health status to the utilization . [↑](#footnote-ref-1)
2. Doctors’ perceptions of marginal benefits: heterogeneous beliefs about appropriate or effective treatment

   Doctors private costs: reflect factors such as skill, training, or experience, liability concerns, affected by organizational features such as available physical capital, the prevalence of non-profit hospitals, nonmonetary career incentives, insurer constraints, peer effects

   Patients health driven by demographics such as age, behavioural factors such as diet, exercise, or smoking and genetic predispositions to disease

   Patient preferences: way patients trade off the disutility of the pain, suffering, or inconvenience of treatment against the value of improved health, as well as ethical or religious beliefs about the value of prolonging life [↑](#footnote-ref-2)
3. Note that the number of robustness checks far exceeds 6 (the enumeration above is an enumeration of categories of robustness checks). [↑](#footnote-ref-3)