Math 242 Final project report

Ngoc Duong

11/9/2017

Socioeconomic factors and viral hepatitis B infection

1 Abstract

Hepatitis B virus (HBV) remains an important cause of acute and chronic liver disease globally and in the United States. Although the prevalence for HBV in the United States has been estimated to be quite low overall, many current literatures have pointed out population groups with high prevalence of HBV infection, namely foreign-born minorities, were underrepresented (Kim, 2009). This paper seeks to build an exploratory model to examine different socioeconomic factors that are significantly associated with the odds of viral hepatitis B infection and have through binary logistic regression. I found a model that has relatively good discriminatory ability (c-index = 0.7538) and concluded that there were strong relationships between certain key predictors and the odds of having hepatitis B among Americans.

Key words: hepatitis B, socioeconomic factors, binary logistic regression

2 Introduction

Hepatitis B is a viral infection that is the major cause of liver cancer and an estimated 257 million people infected worldwide (WHO, 2017). There are three main transmission routes of hepatitis B virus, which are through birth (from mother to child during birthing process), through blood (including blood transfusion and sharing needles), and through unprotected sex. In terms of disease distribution, while Asian Americans make up only 4% of the total population of the United States, they account for half of the country's chronic hepatitis B infection cases. Chronic hepatitis B infection leads to higher risk of cirrhosis and liver cancer in Asian Americans. As a matter of fact, within the US, Asian Americans experience the highest incidence of hepatitis-related liver cancer (Chen et al., 2015). That said, preventive measures taken among the community are still weak. Less than half of adult Asian American patients were reported to have received the hepatitis B vaccine, which suggested inadequacies in screening and vaccination for hepatitis B for Asian Americans (Chu et al., 2013). This indicates very large health disparities in the United States, which needs more attention and effort to eliminate.

In fact, while Asian Americans are more likely to have a college education than other racial groups, on average, Asian Americans also have a higher poverty rate than the national average, exposing them to risk for having limited access to healthcare and health information (Philbin et al., 2012). Among the main reasons, low English proficiency can put restrictions on social mobility and opportunities which leads to a poorer life quality. Additionally, although initially immigrants tend to enjoy better health than the U.S.-born population, this advantage often disappears with longer U.S. residence, attributable in part to adoption of less healthy behaviors such as a less balanced diet (Santa Clara County Public Health Department, 2011)

Considering the different aspects that might play into determinants of the high prevalence of hepatitis B in Asian Americans, it is important that a model can predict whether a person is more likely to be infected with hepatitis B given their individual and socioeconomic characteristics.

It is also worth noting that the final model suffers from issues that should be taken into consideration. These include limitations brought about by large amount of missing observations (which in this model directly affects the usability of the binary outcome variable), cut-off points for answers to certain survey questions, and the need to recode (categorical) variables that have too many levels, as well as the assumption made with missing observations.

3 Methods

The data were collected through the National Health and Examination Survey (NHANES), a program that aims to assess the health and nutritional status of adults and children in the United States through studies. The data collection methods involve a combination of interviews and physical examinations. Data were randomly selected from across the US, in a top-down process from groups of counties to county to group of households to individual households and finally individuals within each household.

The original data set included 46 variables with 6,744 observations. More directly relevant variables were then picked from each group (Demographics, Housing Characteristics, Income, etc.) to subset the original dataset into a new working dataset of 19 variables with 6,744 observations. This process also involved some re-coding of certain variables and elimination of variables that had at least 75% observations missing (e.g. MCQ 203 – ever told to have jaundice, DMDYRSUS – length of stay in the US, etc.) The final working dataset ended up having 3,546 observations across 19 variables.

Statistical Analysis

Descriptive statistics (particularly distribution histograms) were used to check for erroneous entries (Figure 1). Box plots were also created to examine the relationship between these variables and the binary output, as a preliminary check whether the variable should be included in the model (Figure 2). Log odds of having hepatitis B antibody was then computed using each individual variable of interest, adjusting for age and sex. Variables that are not statistically significantly associated with the outcome variable was then excluded before proceeding with the stepwise selection procedure (both forward and backward) to come up with a more parsimonious model. The model's discriminatory power was then examined by looking at the ROC curve/concordance index.

4 Results

Statistically significant predictors included in the final model are: gender, age, education level, race, country of birth, citizenship status, total number of household members, and whether individual was sufficiently vaccinated.

E[log odds of having hepatitis B antibody] = β_0 + β_1 × RIDAGEYR + β_2 × RIAGENDR + β_3 × RIDRETH3(Asian) + β_4 × DMDEDUC2(college education) + β_5 × DMDHHSIZ + β_6 × DMDCITZN (non-citizen) + β_7 × DMDBORN4 (born outside US) +β_8 × IMQ020 (sufficient vaccination)

The ROC curve was plotted and the concordance index was 0.7538, which indicates relatively good discriminatory ability. I also decide at the predictive threshold c = 0.15, the model has best discriminatory power with sensitivity at 0.26156 and specificity at 0.94333 (Figure 3).

5 Discussion

As can be seen from the results, many chosen significant variables are concerned more with the social side of the variable set. The exclusion of other income-based variables (that can also be confounders) might mean the effects of more purely economic and financial factors are not clear. This analysis also highlights the importance of maintaining high vaccination coverage in immunizing the community against hepatitis B infection.

There are some findings that are consistent with the current literature on the topic: Individuals with higher education, or those who are US citizens are more likely to have more protection against hepatitis B virus, adjusting for other variables. In addition, individuals who have more household members (conceivably equivalent to having a big family) are less likely to be immunized.

An interesting result that is inconsistent with literature: Asian Americans are more likely to have immunity against hepatitis B. Therefore, further investigation is needed with regards to this. In fact, since Asian Americans are more likely to have a college education than other racial groups (Philbin et al., 2012), an interaction term between these two variables might be helpful to tease out the pure effect of each variable.

Limitations

The original research question is to look at socioeconomic determinants in the prevalence of hepatitis B infection. However, there are around 90% missing observations for hepatitis B antigen (which is a more direct indicator of hepatitis B infection). Therefore, hepatitis B virus antibody was reluctantly used as a proxy for hepatitis B infection condition.

The raw datasets had many missing observations, and so variables that had at least 75% observations that are NA's, Refuse's, and Don't know's were deliberately left out although these variables might actually have a significant effect on the final predictive power of the model. Furthermore, the final working dataset including only complete cases means that the assumption that certain observations were randomly missing (even though they probably were not), which might be a source of bias and affect the reliability of the results.

Besides, some observations were coded as "x = x or more" so it was not feasible to accurately account for the full scope of all observations within the variable.

The exclusion of certain variables that were weakly statistically significant (p-value around 0.1) for the interest of being parsimonious might also rule out potentially interesting observations.

6 Conclusion

Although the final logistic regression model had comparatively good predictive ability using some significant socioeconomic factors, its weakness lies in exclusion of potential confounders and interaction terms. Stepwise selection using low AIC as a criterion and goal was employed to create the final model, which has relatively good discriminatory ability (c-index = 0.7538). Overall, there were statistically significant associations between socioeconomic factors and the risk of getting hepatitis B infection. The model also reveals the importance and reliability of maintaining high vaccination coverage in immunizing the community against hepatitis B infection. However, further investigation is needed to more accurately examine the effect of the selected predictors and potentially other variables that were not looked at in this analysis.

7 Reference

1. WHO, 2017. Hepatitis B Fact sheet. Available: https:www.who.int/mediacentre/factsheets/fs204/en/ (https:www.who.int/mediacentre/factsheets/fs204/en/)
2. Chen, SM, Jr., Dang, J., 2015. Hepatitis B among Asian Americans: Prevalence, progress, and prospects for control. World J Gastroenterol. 21(42): 11924–11930.
3. Chuang SC, Lee YC, Hashibe M, Dai M, Zheng T, Boffetta P., 2010. Interaction between cigarette smoking and hepatitis B and C virus infection on the risk of liver cancer: a meta-analysis. Cancer Epidemiology, Biomarkers & Prevention. DOI: 10.1158/1055-9965
4. Chu, D., Yang, J. D., Lok, A.,Tran, T., Martins, E. B., Fagan, E., Rousseau, F., Kim, R., 2013. Hepatitis B Screening and Vaccination Practices in Asian American Primary Care. Gut Liver. 7(4): 450–457.
5. Kim, R. W., 2009. Epidemiology of Hepatitis B in the United States. Hepatology. 49(5 Suppl): S28–S34.

6. Philbin, MM, Erby, AHL, Lee, S, Juon, H, 2012, Hepatitis B and Liver Cancer Among Three Asian American Sub-Groups: A Focus Group Inquiry, J Immigr Minor Health. 14(5): 858–868.
7. Status of Vietnamese Health, 2011. Santa Clara County, California Public Health Department, available: https://www.sccgov.org/sites/phd/hi/hd/Documents/VHA%20Full%20Report,%2011.pdf (https://www.sccgov.org/sites/phd/hi/hd/Documents/VHA%20Full%20Report,%2011.pdf)

Figure 1. Histograms of predictors used to check for erroneous entries

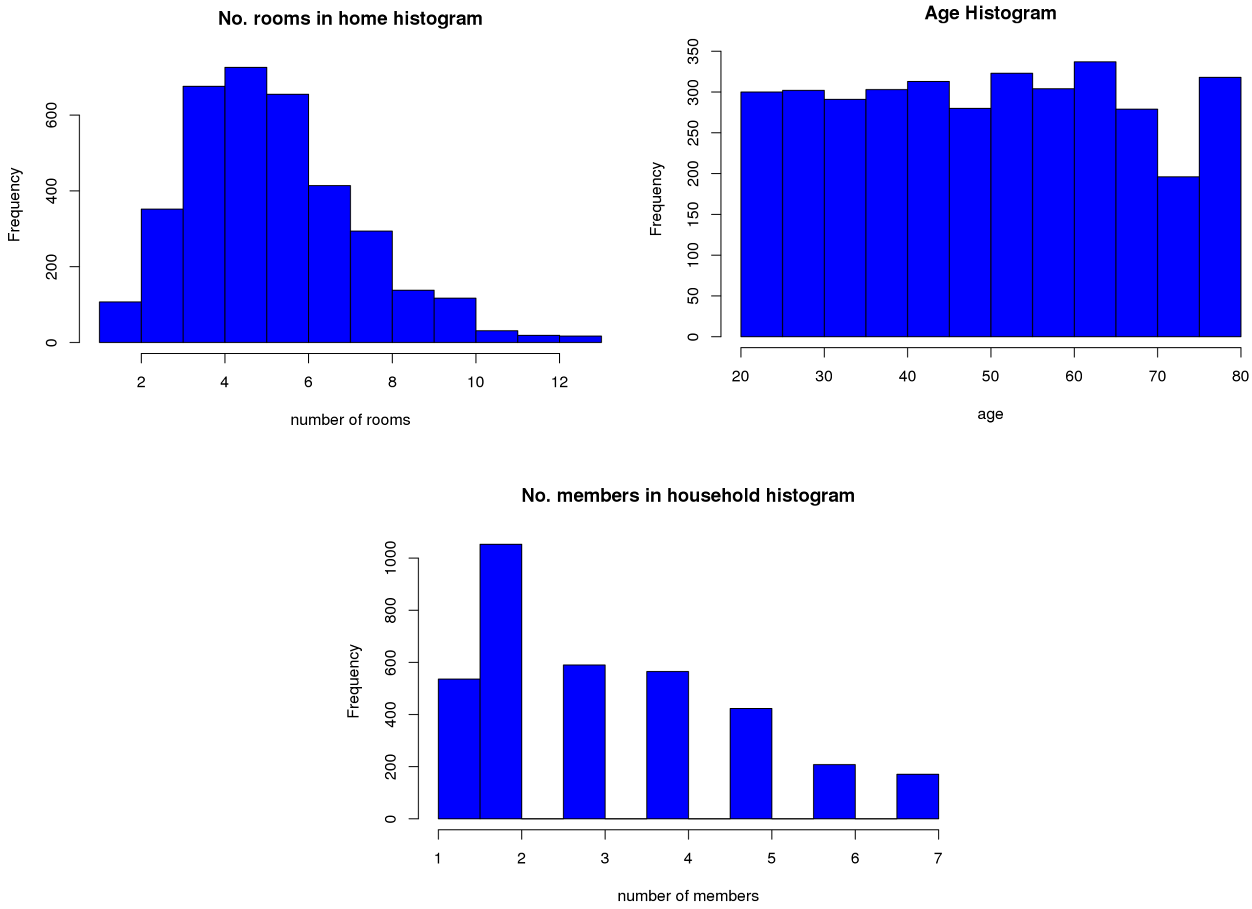
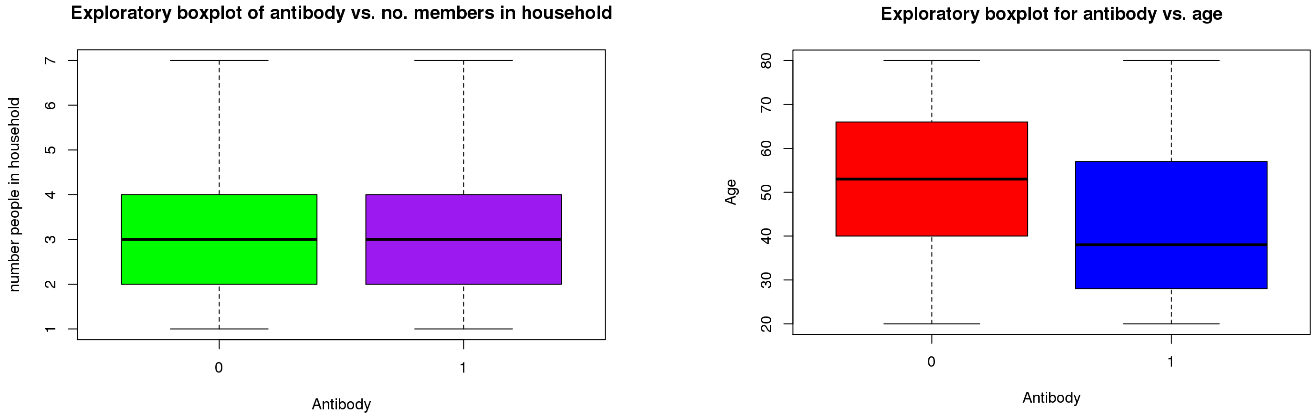


Figure 2. Boxplots between predictors and binary outcome variable



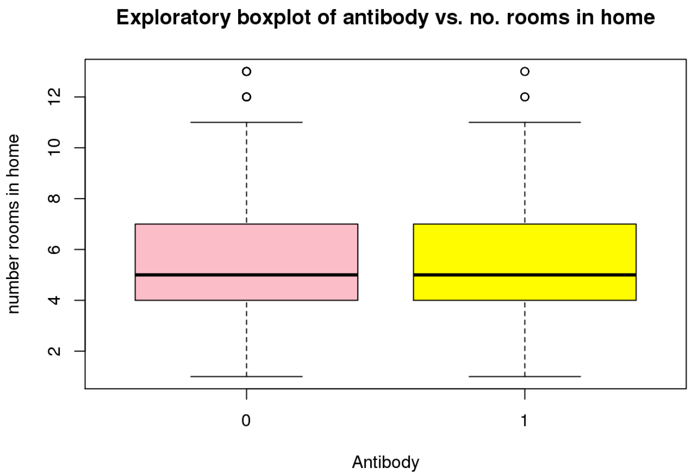
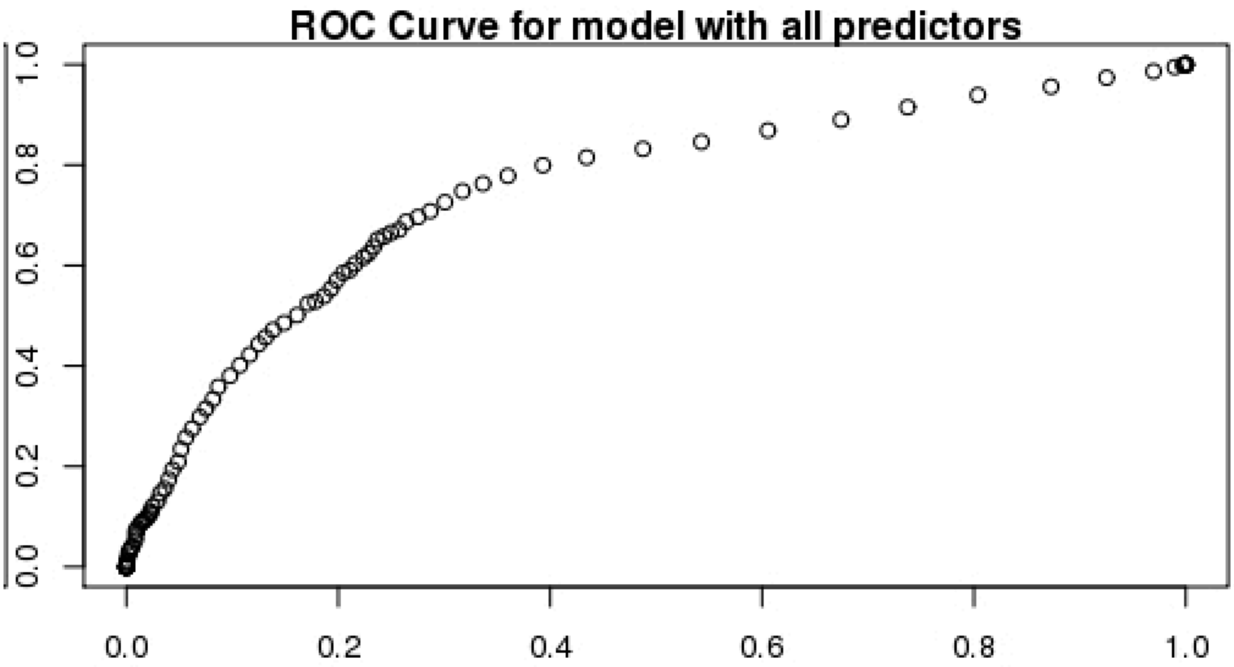


Figure 3. ROC curve with all predictors


ROC Curve for model with all predictors