The Course Report for Generalized Linear Models

Addressing Bias in Data Collection: A Generalized Linear Model Approach in Healthcare Research

卢毓昕 鲁周语

2022~23 Spring

The Course Report for Generalized Linear Models

Addressing Bias in Data Collection: A Generalized Linear Model Approach in Healthcare Research

Abstract

The aim of this report is to investigate and address the issue of bias in data collection for generalized linear models (GLMs) in the context of healthcare and population health. Specifically, we explore the potential biases introduced by participant self-selection and its impact on the validity of research findings. We propose a methodology to quantify, correct, and account for these biases. The report also discusses data acquisition, preprocessing, model selection, and the implications of the findings for future discussion.

I. Introduction

The increasing availability of large-scale health datasets have revolutionized medical research and risk assessment. Dataset like the Nutrition Examination Survey (NHANES) has laid a solid foundation for related researches. However, the accuracy and generalizability of GLM results heavily depend on the representativeness of the collected data. Biases, in which occurs due to the subjectivity of data collection shall not be ignored because certain population segments are more likely to participate or be included in the dataset, leading to distorted conclusions and limited applicability of the findings. Even if NHANEs has already claimed its creditability through large population survey, issues such as participant self-selection bias is still spotted by the previous researches[1]. Participant self-selection bias can arise due to various factors, such as socioeconomic status, health-consciousness, and accessibility to healthcare facilities. For instance, insurance company data may primarily include individuals with higher income and better health management practices. Conversely, data collected from hospitals may predominantly consist of patients with severe medical conditions. Understanding and addressing these biases are crucial for unbiased inference and accurate risk assessment.

II. Methods and Models

2.1 Data Acquisition and Preprocessing

We collected data from NHANES questionnaire data collected in 2017-18 from: https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2017. The dataset includes variables related to insurance coverage, occupation, income, alcohol consumption, and health outcomes. It is collected by trained interviewers in the participant's home and the target population is non-institutionalized civilian resident population of the United States with 16,211 persons selected from 30 different survey locations. Of those selected, 9,254 completed the interview and 8,704 were examined.[2]

To simulate and discover whether the bias appears, we simulate the full NHANES dataset here as the real world situation while extracting subsets according to the attendance to health insurance(HIQ011) and the acceptance for health care(HUQ051).

The dataset underwent preprocessing for sure, including missing value removal and standardization, to align with the GLM frameworks. GLMs were employed to explore relationships between variables and health outcomes, adjusting for potential confounding effects. The variables in the following articles can be categorized to:

-Demographic data: participants' racial information, which was categorized into five dummy variables: Mexican, Hispanic, White, Black, and Asian. Age (RIDAGEYR) and gender (RIAGENDR) were also extracted.

-Body measures data: information on BMI (BMXBMI) and waist circumference (BMXWAIST).

-Blood pressure data: involved three measurements, and the average values of systolic (SBP) and diastolic (DBP) blood pressure were computed.

-Alcohol use data: we initially explored drinking frequency (ALQ121) and average alcohol consumption per drinking day (ALQ130), but they were not significant predictors. Considering the impact of excessive and frequent alcohol consumption on liver health, we focused on the variable ALQ290, representing the times of consuming 12 more drinks in a single day over the past 12 months. We excluded refused or unknown responses and recoded "Never in the last year" as 11 to indicate decreasing drinking frequency.

-Liver ultrasound dataset: providing the median E value (LUXSMED) as an indicator of liver stiffness, eliminating value>10 to insure linear regression.

-Health insurance data: including the variable HIQ011, where 1 denoted individuals covered by insurance.

-Healthcare data: involving the variable HUQ051, and all non-zero values (indicating at least one healthcare visit) were recoded as 1.

-Hepatitis dataset: defining the variable ILL, classifying individuals with Hepatitis B or C as 1 and healthy samples as 0 based on HEQ010 and HEQ030.

2.2 Model Selection and Analysis

We applied GLMs to explore the relationship between the selected variables and the health outcomes of interest. Covariates, including age , gender, BMI, SBP, SDP, Mexican, Black were included in the models to adjust for potential confounding effects. The confounding effects is investigated and optimized through stepwise method.We used regression analysis to examine the associations and calculate relevant statistical measures. The results were visualized using charts and tables to facilitate interpretation.

2.3 Code Environment and Source Code

All our work is done on R studio with R 4.2.3. You may find our source code in the file enclosed: (GLM final project.qmd).
It is also noteworthy that you may prepare your device with Quarto before you open our source code file, for more information about installation, please check: https://quarto.org.

III. Result

Our data is presented with different visualizations in Fig1. With model(Fig 2,3,4) optimized through stepwise processing following the introduction of covariant following, we may have a glimpse of the truth.
In the two subgroups analyzed, the estimated β values for the primary predictor variable ALQ290 (0.034875, 0.037042) were higher compared to the estimated β value (0.028402) for the entire dataset. Moreover, when grouping the data based on healthcare utilization, the significance level of β was higher than that of the model trained on the entire dataset. Conversely, when grouping the data based on health insurance coverage, the significance level of β was lower than that of the model trained on the entire dataset.

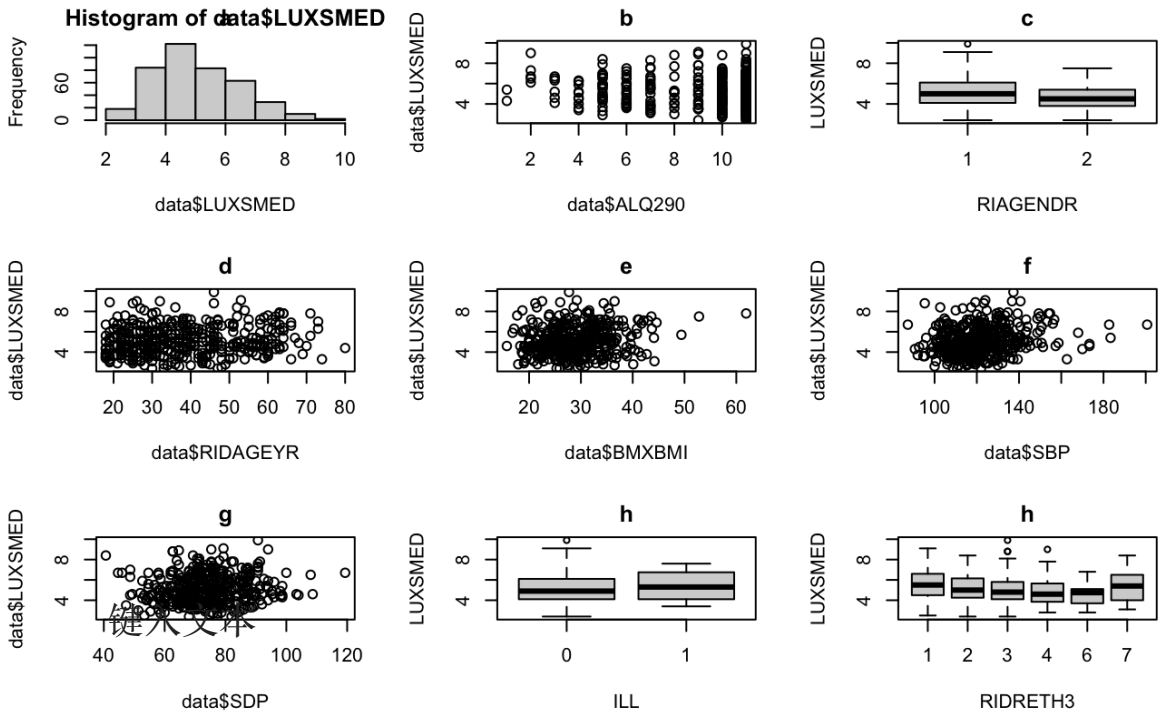

Figure 1: Overview of variable correlation. We plot all the covariates with the dependent variable to check their relationship. As shown in fig1, there is no obvious non-linear relation between any of the covariates and the dependent variable. Thus the method of multiple linear regression is somewhat appropriate.

These findings suggest that individuals with healthcare utilization or medical records may have a higher susceptibility and poorer ability to maintain physiological homeostasis in the face of environmental disruptions, such as heavy alcohol consumption, leading to a stronger (approximately 30%) correlation between liver stiffness and excessive drinking behavior.
These results highlight the potential influence of healthcare utilization and health insurance coverage on the relationship between ALQ290 (drinking frequency) and liver stiffness. It suggests that the observed associations may be more pronounced in populations with healthcare utilization or medical records, possibly due to underlying health conditions or greater vulnerability to alcohol-related liver pathologies. What we did in this research contribute to a better understanding of the complex interplay between alcohol consumption, liver health, and healthcare factors.

```
Call:
lm(formula = LUXSMED ~ ALQ290 + RIAGENDR + RIDAGEYR + BMXBMI +
    SBP + SDP + Mexican + Black, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0659 -0.9205 -0.0897  0.8778  4.7578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.333305   0.743017   4.486 9.48e-06 ***
ALQ290      -0.070084   0.028402  -2.468 0.014019 *
RIAGENDR    -0.270724   0.163366  -1.657 0.098267 .
RIDAGEYR     0.007948   0.005146   1.545 0.123215
BMXBMI       0.032769   0.011216   2.922 0.003678 **
SBP          0.020922   0.006021   3.475 0.000567 ***
SDP         -0.014294   0.007601  -1.880 0.060772 .
MexicanTRUE  0.418141   0.175644   2.381 0.017749 *
BlackTRUE   -0.340158   0.197576  -1.722 0.085901 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.337 on 402 degrees of freedom
Multiple R-squared:  0.1208,    Adjusted R-squared:  0.1033
F-statistic: 6.907 on 8 and 402 DF,  p-value: 1.571e-08
```

Figure 2. The results of the linear regression model examining the relationship between LUXSMED and predictor variables, ALQ290, RIAGENDR, RIDAGEYR, BMXBMI, SBP , SDP, Mexican, and Black in real world population.

```
Call:
lm(formula = LUXSMED ~ ALQ290 + RIAGENDR + RIDAGEYR + BMXBMI +
    SBP + SDP + Mexican + Black, data = data_hi)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9967 -0.9047 -0.1047  0.8135  4.8224

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.672860   0.875226   4.196 3.62e-05 ***
ALQ290      -0.116540   0.034875  -3.342 0.000944 ***
RIAGENDR    -0.233028   0.178047  -1.309 0.191654
RIDAGEYR     0.002432   0.005860   0.415 0.678492
BMXBMI       0.038125   0.012678   3.007 0.002871 **
SBP          0.027619   0.007148   3.864 0.000138 ***
SDP         -0.022513   0.008949  -2.516 0.012424 *
MexicanTRUE  0.300631   0.216032   1.392 0.165124
BlackTRUE   -0.526991   0.238741  -2.207 0.028082 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.318 on 286 degrees of freedom
Multiple R-squared:  0.1543,    Adjusted R-squared:  0.1307
F-statistic: 6.524 on 8 and 286 DF,  p-value: 8.166e-08
```

Figure 3. The results of the linear regression model examining the relationship between LUXSMED and predictor variables, ALQ290, RIAGENDR, RIDAGEYR, BMXBMI, SBP , SDP, Mexican, and Black in healthcare accepted groups

```
Call:
lm(formula = LUXSMED ~ ALQ290 + RIAGENDR + RIDAGEYR + BMXBMI +
    SBP + SDP + Mexican + Black, data = data_hi)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7509 -0.9173 -0.1190  0.8807  4.6857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.129249   0.938104   3.336 0.000962 ***
ALQ290      -0.070574   0.037042  -1.905 0.057741 .
RIAGENDR    -0.254446   0.193132  -1.317 0.188723
RIDAGEYR     0.006116   0.005999   1.019 0.308819
BMXBMI       0.034914   0.013681   2.552 0.011222 *
SBP          0.021563   0.007125   3.026 0.002699 **
SDP         -0.012065   0.008874  -1.360 0.175009
MexicanTRUE  0.418238   0.235082   1.779 0.076271 .
BlackTRUE   -0.392170   0.244787  -1.602 0.110229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 289 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.09577
F-statistic: 4.932 on 8 and 289 DF,  p-value: 9.863e-06
```

Figure 4. The results of the linear regression model examining the relationship between LUXSMED and predictor variables, ALQ290, RIAGENDR, RIDAGEYR, BMXBMI, SBP , SDP, Mexican, and Black in health insurance population.

IV. Discussion

Our Result shows the potential difference that may occur in different populations while investigating through GLMs. It is noteworthy that the biased collection for data may result in a change of significance much better than people normally believed. The model shows a future potential and responsibility for more researches into the participant self-selection bias. Incorporating appropriate techniques such as propensity score matching or importance sampling weights may help correct for biases in the analysis while we still need more evidence in their effectiveness.

V. Reference and Acknowledgement

1. FORTHOFER, R. N. (1983). "INVESTIGATION OF NONRESPONSE BIAS IN NHANES II1." American Journal of Epidemiology 117(4): 507-515.
2. NHANES 2017-2018 Overview https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2017#print
3. Boursier, J., et al. (2022). "Quality criteria for the measurement of liver stiffness." Clinics and Research in Hepatology and Gastroenterology 46(1): 101761.
4. Chapman, T., et al. (2017). "Ultrasound Elastography of the Liver: What the Clinician Needs to Know." Journal of Ultrasound in Medicine 36(7): 1293-1304.

5. **We ‘d like to express our most sincere gratitude to Miss Wu, who inspired , guided, and motivated us so much in her Generalized Linear Model class.**

VI. Supplementary Data

Well, there ’s some unnoticed figure in our research:
So here ’s the list:

Gender	Count
Male	314
Female	97

Race	Count
Mexican American	79
Other Hispanic	40
None-Hispanic White	179
None-Hispanic Black	60
None-Hispanic Asian	26
Other Race	27

ILL	Count
Healthy	403
Hepatitis B or C	8