First Analysis Write-Up

Yiying Wu

Note: All codes are available here: https://github.com/wuyiying2018/food-insecurity

1. Introduction

2. Methodology

2.1 Data sources

The dataset I chose is the NHANES 2017-March 2020 dataset. This survey is conducted, and the data is collected by the National Center for Health Statistics (NCHS) from 2017 to March 2020. The NHANES 2019-2020 field operations were suspended in March 2020 due to the COVID-19 pandemic, resulting in incomplete data collection that is not nationally representative. To address this issue and ensure national representativeness, the partially collected 2019-2020 data were merged with the complete 2017-2018 dataset, creating a combined dataset representative of the U.S. civilian non-institutionalized population before the pandemic. The dataset contains demographic, socioeconomic, dietary, and health-related information.

The target population is adults over 20 years old (includes 20) in U.S.

2.2 Weighting the Survey Data

In NHANES dataset weights are adjusted for selection probability, non-response, and post- stratification to match U.S. population figures. To obtain a valid statistical inference, a domain analysis for the adult subpopulation (age ≥ 20) are conducted by using the subset function in the survey package in R.

2.3 Multiple Imputation

To handle missing data in our NHANES dataset, we utilize Multiple Imputation (MI) to generate multiple datasets by imputing missing values repeatedly. This method retains the advantages of single imputation, such as consistent analyses and data collector's knowledge, while also accurately reflecting uncertainty and accounting for imputation error. We implemented it using the MICE (Multivariate Imputation by Chained Equations) package in R, which allows for flexible and efficient imputation of missing values. This ensures that the imputed values are plausible and improves statistical efficiency.

2.4 Logistic Regression

Given the binary outcome: food security (Yes/No), we chose logistic regression as our statistical model. Suppose there are n covariates $X_i, i = 1, \dots n$, the model can be expressed as:

$$log(\frac{\pi}{1-\pi}) = X\beta = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

- π is the risk
- $\frac{\pi}{1-\pi}$ is the odds
- β_0 is the log odds for $X_i^{\prime}s = 0$

- β_i is the log odds ratio per unit change of X_i, holding all other covariates fixed

In the logistic regression analysis for survey data, I have integrated multiple imputation and complex survey design in NHANES. Each imputed dataset was then individually analyzed with logistic regression models, incorporating essential survey design elements like stratification and weights via the survey package. Lastly, these results are aggregated to derive final estimates using Rubin's multiple imputation combining rule (Rubin, 2018).

2.5 Machine Learning Method

plan to use the `surveyCV` package

some useful links:

https://github.com/ColbyStatSvyRsch/surveyCV

https://stats.stackexchange.com/questions/238141/two-worlds-collide-using-ml-for-complex-survey-data

https://cran.r-project.org/web/packages/surveyCV/index.html

3. Results

3.1 Descriptive Statistics

Table 1: Summary of Dataset by Food Security

Characteristic	**food insecurity**, N = 3,163[1]	**food security**, N = 5,380[1]
gender		
male	1,490 (47%)	2,639 (49%)
female	1,673 (53%)	2,741 (51%)
age	48 (33, 62)	54 (38, 67)
race		
Mexican American	507 (16%)	478 (8.9%)
Other Hispanic	446 (14%)	419 (7.8%)
Non-Hispanic White	778 (25%)	2,244 (42%)
Non-Hispanic Black	1,038 (33%)	1,224 (23%)
Non-Hispanic Asian	208 (6.6%)	798 (15%)
Other Race - Including Multi-Racial	186 (5.9%)	217 (4.0%)
not_born_in_us	997 (32%)	1,418 (26%)
Unknown	3	0
bmi	30 (25, 35)	28 (25, 33)
Unknown	216	471
education		
Less than 9th grade	402 (13%)	250 (4.6%)
9-11th grade	542 (17%)	414 (7.7%)
High school graduate/GED or equivalent	928 (29%)	1,132 (21%)
Some college or AA degree	993 (31%)	1,786 (33%)
College graduate or above	292 (9.2%)	1,796 (33%)
Unknown	6	2
marital_status		
Married/Living with Partner	1,547 (49%)	3,340 (62%)
Widowed/Divorced/Separated	842 (27%)	1,161 (22%)
Never married	769 (24%)	876 (16%)
Unknown	5	3
poverty	1.27 (0.78, 2.06)	3.28 (1.76, 5.00)
Unknown	328	387
hbp	1,252 (40%)	2,034 (38%)
Unknown	6	6
diabetes	547 (17%)	771 (14%)
Unknown	1	1
ckd	154 (4.9%)	196 (3.6%)
Unknown	5	9
insurance	2,383 (76%)	4,807 (89%)
Unknown	11	5

[1] n (%); Median (IQR)

3.2 Regression Analysis Results

coef names	**OR**	**CI low**	**CI up**
Intercept	0.0865675	0.0455321	0.1645858
gender			
male	ref	ref	ref
female	0.8882700	0.7938057	0.9939758
age	1.0242605	1.0165347	1.0320450
race			
Mexican American	ref	ref	ref
Other Hispanic	0.6944784	0.4787410	1.0074344
Non-Hispanic White	1.6626392	1.0989956	2.5153595
Non-Hispanic Black	1.1297815	0.7313368	1.7453055
Non-Hispanic Asian	1.7310878	1.1105557	2.6983472
Other Race - Including Multi-Racial	0.7473892	0.3991244	1.3995402
not born in U.S.	1.1155619	0.8022678	1.5512006
bmi	0.9850796	0.9752864	0.9949711
education			
Less than 9th grade	ref	ref	ref
9-11th grade	1.3439491	0.9764702	1.8497227
High school graduate/GED or equivalent	1.6825892	1.2053256	2.3488314
Some college or AA degree	2.1286519	1.6312931	2.7776485
College graduate or above	3.7148680	2.7390089	5.0384081
marital status			
Married/Living with Partner	ref	ref	ref
Widowed/Divorced/Separated	0.7185600	0.5343786	0.9662221
Never married	0.8816866	0.6728986	1.1552577
poverty	2.0249656	1.8245303	2.2474199
hbp	0.8545878	0.6848649	1.0663715
diabetes	0.7961340	0.5730458	1.1060710
ckd	0.8875632	0.5684104	1.3859149
insurance	1.2391050	0.9339267	1.6440060

3.3 Machine Learning Results

4. Conclusion

Reference

Rubin, D.B. (2018). Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC.