MAT5317 Categorical Assignment 1

Teng Li(7373086)
Zhize Lu(300075114)
Chutong Zhang(300311325)

Introduction

We were given the data set of The National Health and Nutrition Examination Survey (NHANES). The survey program has been conducted as a series of surveys designed to assess the health and nutritional status of adults and children in the United States since the 1960s, according to CDC (2023). It combines in-person face-to-face interviews and physical examinations of participants for data collection.

The survey data wasn't a simple random sample, however. According to CDC's National Health and Nutrition Examination Survey: Plan and Operations, 1999–2010 (G et al. 2013), the sampling strategy consists of several stages: 1. Selection of counties as primary sampling units (PSU). 2. selection of segments within PSUs that constitute blocks of households. 3. Selection of specific households within segments. 4. Selection of individuals within a household.

We aim to study the relationship between the weight variable and the other health related variables of the data.

Method

We began our study by doing an exploratory analysis among the variables through various tables and charts. We then performed several hypothesis tests on some of the variables. Lastly we did a linear regression model fit to the response variable "weight" with other variables and confounders.

Table 1: Data Variable Definition

Variables	Type	Example	Number.Unique	MissingPct	Comment
id	integer	1, 2, 3	6482	0%	Identification Code (1 - 6482)
gender	factor	Male, Female	2	0%	Gender (1: Male, 2: Female)
age	integer	34, 16, 60	65	0%	Age (Years)
marstat	factor	Married, NA, Widowed	6	9.7%	Marital Status (1: Married, 2: Widowed, 3: Divorced, 4: Separated, 5: Never Married, 6: Living Together)
samplewt	numeric	80100.544, 13953.078, 20090.339	2499	0%	Statistical Weight (4084.478 - 153810.3)
psu	integer	1, 2	2	0%	Pseudo-PSU (1, 2)
strata	integer	9, 10, 1	15	0%	Pseudo-Stratum (1 - 15)
tchol	integer	135, 192, 202	251	6.09%	Total Cholesterol (mg/dL)
hdl	integer	50, 60, 45	112	6.09%	HDL-Cholesterol (mg/dL)
sysbp	integer	114, 112, 154	61	8.53%	Systolic Blood Pressure (mm Hg)
dbp	integer	88, 62, 70	40	9.16%	Diastolic Blood Pressure (mm Hg)
wt	numeric	87.400002, 72.300003, 116.8	957	0.57%	Weight (kg)
ht	numeric	164.7, 181.3, 166	527	0.57%	Standing Height (cm)
bmi	numeric	32.22, 22, 42.39	2276	0.57%	Body mass Index (Kg/m^2)
vigwrk	factor	No, Yes, NA	2	0.02%	Vigorous Work Activity (1: Yes, 2: No)
modwrk	factor	No, Yes, NA	2	0.02%	Moderate Work Activity (1: Yes, 2: No)
wlkbik	factor	No, Yes, NA	2	0.02%	Walk or Bicycle (1: Yes, 2: No)
vigrecexr	factor	No, Yes, NA	2	0.02%	Vigorous Recreational Activities (1: Yes, 2: No)
modrecexr	factor	No, Yes, NA	2	0.03%	Moderate Recreational Activities (1: Yes, 2: No)
sedmin	integer	480, 240, 720	37	1.22%	Minutes of Sedentary Activity per Week (0 - 840)
obese	factor	No, Yes, NA	2	0.57%	BMI>35 (1: No, 2: Yes)

We first test the independence between obesity and marital status. We form the following contingency table:

```
## # A tibble: 6 x 3
## # Groups:   marstat [6]
##   marstat             No   Yes
##   <fct>            <int> <int>
## 1 Married           2530   474
## 2 Widowed            418    86
## 3 Divorced           528   112
## 4 Separated          158    35
## 5 Never Married      863   160
## 6 Living Together    388    66
```

Let X be the categorical random variable for Marital Status and Y be the one for Obesity. Define the count random variable $N_{ij} := \sum_{i=1}^{I} \sum_{j=1}^{J} \mathbb{1}(X = i, Y = j)$, then the joint random variables $[N_{11}, ..., N_{IJ}]$ has a Multinomial distribution $\vec{p} = [p_{11}, ..., p_{IJ}]$. Our hypothesis test is therefore:

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j} \ \forall i, j$$
$$H_1 : p_{ij} \neq p_{i.} \cdot p_{.j} \ \forall i, j$$

We use the chi-squared test to conclude that there is not enough evidence to reject the null hypothesis with a p-value equal to 0.6893764. In other words, we cannot conclude that there is a relationship between obesity and marital status.

Conclusion

References

2023. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

G, Zipf, Chiappa M, Porter KS, et al. 2013. "National Health and Nutrition Examination Survey: Plan and Operations, 1999–2010." *National Center for Health Statistics* 1 (56).