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)

```
##       wt         marstat obese
## 1  87.4         Married    No
## 3 116.8         Widowed   Yes
## 4  97.6         Married    No
## 5  86.7 Living Together    No
## 6  79.1         Married    No
## 7  89.6         Widowed    No
```

Table 2: Contingency Table

		Obesity	
		No	Yes
Marital Status	Married	2530	474
	Widowed	418	86
	Divorced	528	112
	Separated	158	35
	Never Married	863	160
	Living Together	388	66

Given the categories of marriage, we are interested in the number of obese people in each category. We have six categories in total. This forms a Multinomial random variable $\vec{X} = (X_1, ..., X_6)$ where each X_i is the number of obese people in category i. We get the estimates of the MLEs for the corresponding proportions:

```
## # A tibble: 6 x 4
## # Groups:   marstat [6]
##   marstat            No   Yes p_mle
##   <fct>           <int> <int> <dbl>
## 1 Married          2530   474 0.158
## 2 Widowed           418    86 0.171
## 3 Divorced          528   112 0.175
## 4 Separated         158    35 0.181
## 5 Never Married     863   160 0.156
## 6 Living Together   388    66 0.145
```

The covariance matrix of the MLE is:

$$Cov(\hat{p_1}, \hat{p_2}) = \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & ... & -\frac{n}{n_1 n_k} p_1 p_k \\ -\frac{n}{n_k n_1} p_1 p_k & ... & \frac{p_k(1-p_k)}{n_k} \end{bmatrix}$$

```
##                    Married   Widowed  Divorced Separated Never Married
## Married           0.000044 -0.000103 -0.000084 -0.000287     -0.000047
## Widowed          -0.000103  0.000281 -0.000539 -0.001851     -0.000301
## Divorced         -0.000084 -0.000539  0.000226 -0.001495     -0.000243
## Separated        -0.000287 -0.001851 -0.001495  0.000769     -0.000836
## Never Married    -0.000047 -0.000301 -0.000243 -0.000836      0.000129
## Living Together -0.000098 -0.000631 -0.000509 -0.001750     -0.000285
##                 Living Together
## Married               -0.000098
## Widowed                -0.000631
## Divorced               -0.000509
## Separated              -0.001750
## Never Married          -0.000285
## Living Together         0.000274
```

We are interested in testing the following hypothesis:

$$
\begin{gathered}
H_0 : p \\
H_1 : p2
\end{gathered}
$$

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).