Homework 2

BS 728, Fall 2021

This homework will have you work through a data analysis of data collected from a complex sample. Here are some guidelines/instructions for this portion of the homework:

- Please hand in your log and output in the Dropbox on blackboard. Make sure that the log and output only have information relevant to the final results in your homework (this should not contain runs that did not work or were abandoned).
- In general, round to 2 decimal places (eg 1.46582 -> 1.47; 0.436765 -> 0.44). For p-values that are very small, it is fine to write p<0.001.
- If you are relatively new to using SAS, feel free to use the shell code on Blackboard to complete this assignment. If you are more experienced, please avoid using this code.
- You will be working with a reduced dataset from the original dataset since the original dataset is too large to run on SAS Studio. So, results are not going to be entirely accurate and true to what they would be with the complete dataset. (if you want to use the complete dataset, let me know-I can give you access to it).

The Demographic and Health Surveys (DHS) are large multi-stage surveys that are carried out globally. For this exercise, we are going to consider the DHS survey conducted in India in 2005-6. This survey contains information on 515,507 individuals from 109,041 sample households. Respondents provide information on a large number of demographic and health related areas (thus the name of the survey!).

You have access to a subset of this data for this exercise. (See data dictionary at the end of the assignment) We will focus on Tuberculosis (TB) and consider TB prevalence and correlates with TB in India. TB is currently the leading cause of death due to infectious disease globally and India is one of the countries with a particularly high TB burden. TB is primarily a respiratory infection, making smoking and cooking fuel, as well as ventilation of interest as potentially increasing the chance of disease in individuals.

1. Read the sampling design section of Volume II of the Country report associated with this study (page 4 of Appendix C or page 599 of the pdf report).
    a. Why were different sampling strategies employed for rural and urban areas? What is the difference between the sampling strategies?

    b. In one sentence, what was the sampling frame in rural areas (do not go in to detail of what was and was not included)?

    c. Why was stratification used in the rural sample? What was the first level of stratification done on?

2. If you are using SAS, you will need to first read in the formats file for this dataset. Open the file ‘formats_dhs.sas’ within SAS and run this file. Once you have done this, clear your log, so this does not clutter it. Now open a new editor window in sas to complete your homework (do not put your homework in the ‘formats_dhs.sas’ editor!!!)

3. Unweighted analysis of data. We will first analyze the data without accounting for the complex sampling structure.
    a. What is your overall estimate of TB prevalence?

    b. *Descriptive and bivariate analysis of the data*. Fill in the missing parts of Table 1a without accounting for the sampling scheme. For the first column provide appropriate summary statistics for each of the variables listed for the entire sampled population. For the next two rows, provide this information for those with TB (column 3) and those without TB (column 4). For the missing rows, do not include a %, just tabulate the number of individuals missing this information. The percents for the other entries should add to 100.
    c. *Bivariate analysis*. In the final column of Table 1a, put in the p-value for comparing those with and without TB by the variable in column using an appropriate statistical test. For instance to compare the proportion of males who are TB+ to the proportion of males who are TB- using a chi square test.
    d. *Multivariate analysis of the data*. Fill in the missing parts of Table 2a, using an appropriate multivariate regression model.

4. Weighted analysis of the data. Follow the directions for question 3, but in this problem account for the complex sampling design in your analysis.

    a. What is your overall estimate of TB prevalence?

    b. Descriptive statistics. Use weighted frequencies and percents.
    c. Bivariate analysis.
    d. Multivariate regression

5. What appears to be associated with TB disease? How might the sample size impact your conclusions?

	Overall (N=X)	TB (N=X)	No TB (N=X)	p-value
Male (N, %)	25253 (50.5)	537 (48.8)	24716 (50.6)	0.25
Female (N, %)	24747 (49.5)	564 (51.2)	24183 (49.4)	
Age (mean, 95% CI)	26.8 (26.6, 27.0)	26.0 (24.9, 27.2)	26.8 (26.6, 27.0)	0.18
Age group				X
0-4	X	116 (10.5)	5130 (10.5)	
5-14	X	272 (24.7)	10953 (22.4)	
15-39	20849 (41.7)	442 (40.2)	X	
≥40	12675 (25.4)	271 (24.6)	12404 (25.4)	
Missing	5			
BMI (mean, 95% CI)	X	X	X	X
BMI category				0.001
Malnourished	1410 (12.4)	49 (18.3)	1361 (12.3)	
Underweight	1916 (16.8)	49 (18.3)	1867 (16.8)	
Healthy	6316 (55.5)	147 (54.9)	6169 (55.5)	
Overweight	1317 (11.6)	21 (7.8)	1296 (11.7)	
Obese	416 (3.7)	2 (0.8)	414 (3.7)	
Missing	38625			
Windows in home (N, %)	X	681 (61.9)	36358 (74.4)	X
Missing	30			
Cook with wood fuel (N, %)	22690 (45.4)	X	X	X

Table 1a. Analysis not accounting for sampling design.

	Overall (N= X)	TB (N= X)	No TB (N= X)	p-value
Male (N, %)	X	5596 (47.2)	250884 (49.9)	X
Female (N, %)	255853 (49.9)	6249 (52.8)	249604 (50.1)	
Age (mean, 95% CI)	26.2 (25.9, 26.4)	X	X)	X
Age group				0.58
0-4	58365 (11.4)	1515 (12.9)	56850 (11.4)	
5-14	121878 (23.8)	2893 (24.4)	118985 (23.8)	
15-39	205376 (40.1)	4485 (37.9)	200892 (40.1)	
≥40	126663 (24.7)	2952 (24.9)	123711 (24.7)	
Missing	5			
BMI (mean, 95% CI)	20.5 (20.4, 20.6)	X	X	X
BMI category				0.02
Malnourished	17296 (15.1)	X	16649 (14.8)	
Underweight	22560 (19.7)	X	21899 (19.6)	
Healthy	60822 (53.0)	1359 (47.5)	59463 (53.1)	
Overweight	10763 (9.4)	169 (5.9)	X	

Obese	3320 (2.9)	24 (0.8)	X	
Missing	38625			
Windows in home (N, %)	343541 (67.1)	6330 (53.4)	337210 (67.4)	X
Missing	30			
Cook with wood fuel (N, %)	253289 (49.4)	64053 (1.2)	2611066 (48.0)	X

Table 1b. Analysis accounting for sample design.

	OR	95% CI	p-value
Female vs Male	1.07	X	X
Age			
0-4	0.97	0.78, 1.21	X
5-14	1.07	0.91, 1.27	X
15-39	0.99	0.85, 1.15	X
≥40	REF		
Windows	X	0.52, 0.67	X
Wood Fuel	X	1.09, 1.39	X

Table 2a. Regression not accounting for study design

	OR	95% CI	p-value
Female vs Male	1.12	X	X
Age			
0-4	1.06	0.80, 1.41	0.68
5-14	0.97	0.78, 1.22	0.81
15-39	0.93	0.77, 1.22	0.49
≥40	REF		
Windows	X	0.45, 0.68	X
Wood Fuel	0.95	X	X

Table 2b. Regression accounting for study design

Data Dictionary

Variable name	Description	Coding
Gender	Gender of the individual	M, F
TB	Indicator of whether individual has TB disease	0: no disease 1: has TB disease
Age		
Age_cat		0: 0-4 1: 5-14 2: 15-39 3: ≥40
BMI	BMI quantitative value	
BMI_cat	BMI in categories established by WHO	Malnourished (BMI<17) Underweight (17 ≤ BMI < 18.5) Healthy (18.5 ≤ BMI < 25) Overweight (25 ≤ BMI < 30) Obese (BMI ≥ 30)
Wood_fuel	Indicator of whether wood fuel is used in cooking	0: no wood fuel 1: use wood fuel
Smoker	Indicator of whether the respondent is a smoker	0: not a smoker 1: smoker
Windows	Indicator of whether the home has windows	0: no windows 1: windows
Wgtdhs	Weighting variable for each individual	
Sho21	Cluster variable for sampling design	
hv022	Stratum variable for sampling design	