**IHSD 7440 - Homework #3 2023**

**Name\_\_\_Michaela Proffitt**

Answer the questions to this assignment in the spaces below. **Turn in your completed assignment via the personal GitHub repository you created for this class by 3/31 at 5pm.**

Background

The Roll Back Malaria (RBM) initiative began in 1998 by WHO as an international effort to halve the 2000 levels of malaria morbidity and mortality by 2010 and to reduce this malaria burden by a further 50 percent by 2015. One of RBM’s core indicators is the proportion of households with at least one insecticide treated net (ITN). ITNs are a key tool in reduction of malaria transmission and subsequent reduction in child and adult morbidity and mortality. An ITN is defined as a mosquito net treated with a long-lasting insecticide or a mosquito net that has been dipped in insecticide within the past 12 months. Efforts to scale up ITN coverage are underway in most African countries. Nationally representative population-based surveys such as the DHS are the data collection methods preferred to measure RBM indicators including proportion of households with at least one ITN. More information on RBM can be found at <http://www.rbm.who.int>.

Assignment

In this assignment, we are interested in the following indicators at the household and child level:

Among households:

1. Proportion of households with at least 1 ITN

Among children:

1. Proportion of children under the age of 5 that used an ITN the previous night

We will look at overall estimates for these two indicators as well as by the following factors:

1. Residence: urban/rural (all indicators)
2. Household socioeconomic status: wealth quintile (all indicators)
3. Household head education (for the household ITN possession indicators)
4. Child’s age (for the ITN use analysis among children)
5. Mothers education (for the ITN use analysis among children)

We will be using subsets of the household and child-level dataset for Zambia 2007. Both are available for download on Canvas under Assignments, Homework #3, as well as in the IHSD 7440 HH Sampling GitHub repository.

You must submit 3 things for this assignment. **Please make sure to push all these files to your personal GitHub repository for this class.**

1. The attached tables and questions completed

2. An R Markdown file featuring your code used to complete the assignment

3. The html notebook output created when running your finalized R Markdown file

These data were collected using the standard DHS sampling protocol, which consists of a 2-stage cluster design with first stage selection of primary sampling units selected proportional to their size (PPS). All women of reproductive age were asked for information of their children. Independent samples were selected within survey domains at the Regional level. Within each survey domain, data were collected using a proportional stratification system to improve the precision of the estimates. For all household and child-level data, sample weights were created based relative strata sizes, and on the difference between estimated cluster size (M) and actual cluster size (B).

Please note variable **HV005 is the sampling weight**, **variable HV021, labeled as Primary Sampling Unit, provides the cluster number for the analysis**, and **variable HV022, labeled Sample stratum number, provides the strata number for the analysis**.

**Problem 1: HH-level analysis**

Using the ***2009\_Zambia\_HH\_2023.csv*** dataset located in the IHSD 7440 GitHub repository, please answer the following questions. You can use either R Studio or STATA to complete this exercise, but R Studio is recommended.

* + 1. What is your element in this analysis and how many are there (n)?

Element is the household; n= 6439

* + 1. How many clusters are there in this sample?

285

* + 1. How many households were selected in each cluster / PSU?

6-48

* + 1. How many survey domains are there in this dataset?

16

* + 1. How many strata are there in this dataset?

18

Now complete the following tables. The tables show the proportion of households that own at least 1 ITN. For column d, you need to analyze the data appropriately, taking into account the following: 1) the use of a 2-stage cluster design that results in correlated data at the cluster / PSU level; 2) adjustment for differences in the ultimate probability of selection through sampling weights; and 3) uses the strata information to improve the precision of your estimates.

**In each cell, include the proportion and the standard error (round to 3 decimal points)**

Table 1: Proportion of households that own at least 1 ITN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n = | (a) Assuming SRS | (b) Assuming SRS, with weights | (c) 2-stage cluster sampling with weights | (d) 2-stage cluster sampling with weights and stratification |
| Residence |  |  |  |  |
| Urban | 0.540 (0.010) | 0.520 (0.012) | 0.520 (0.018) | 0.556 (0.018) |
| Rural | 0.572 (0.008) | 0.556 (0.008) | 0.556 (0.019) | 0.520 (0.018) |
| SES |  |  |  |  |
| Poorest | 0.421 (0.015) | 0.480 (0.016) | 0.481 (0.028) | 0.481 (0.027) |
| Poorer | 0.560 (0.013) | 0.549 (0.014) | 0.549 (0.025) | 0.549 (0.023) |
| Middle | 0.582 (0.013) | 0.585 (0.014) | 0.585 (0.022) | 0.585 (0.020) |
| Richer | 0.529 (0.013) | 0.510 (0.015) | 0.510 (0.023) | 0.510 (0.022) |
| Richest | 0.624 (0.015) | 0.589 (0.017) | 0.589 (0.021) | 0.589 (0.020) |
| HH head education |  |  |  |  |
| None | 0.421 (0.016) | 0.405 (0.017) | 0.405 (0.026) | 0.405 (0.025) |
| Primary | 0.546 (0.009) | 0.528 (0.010) | 0.528 (0.018) | 0.528 (0.017) |
| Secondary | 0.612 (0.011) | 0.599 (0.013) | 0.599 (0.018) | 0.599 (0.016) |
| Higher | 0.704 (0.012) | 0.681 (0.023) | 0.681 (0.022) | 0.681 (0.022) |
| **All Households** | **0.560 (0.006)** | **0.542 (0.007)** | **0.543 (0.014)** | **0.543 (0.013)** |

* + 1. What is the effect of sample weights on point estimates and standard errors?

Sample wrights seem to have a minimal effect on point estimates and standard error, slightly decreasing the point estimate and a slight increase to the standard error

* + 1. What is the effect of the cluster sampling design (i.e. use of clusters at first stage) on the standard errors (i.e. when using the Huber-White Sandwich estimator in SAS or STATA- e.g. using a cluster command)?

Clustering increases the standard error.

* + 1. What is the design effect for this 2-stage cluster sampling design for the proportion of households with at least 1 ITN, with sample weights and stratification included in the analysis?

0.013/0.006= 2.167

* + 1. How does household residence – urban versus rural - affect the proportion of households with at least 1 ITN?

Using all sampling strategies, with the exception of 2-stage clustering with weights and stratification, rural populations have a higher proportion of households that own at least one ITN.

* + 1. Which of the four estimates (a, b, c, d in table 1 above) provides the least biased point estimates and standard errors of the ITN household possession estimates, and why?

Sampling design does not influence bias when sampling is done appropriately; standard error represents to precision, and precision and bias have no relationship.

**Problem 2: Individual-level analysis**

Using the ***2009\_Zambia\_child\_2023.csv*** dataset located in the IHSD 7440 GitHub Repository, please answer the following questions.

1. What is your element in this analysis and how many are there (n)?

Individual children n=5194

1. How many clusters are there in this sample?

285

Now complete the following tables. The tables show the proportion of children that slept under an ITN the previous night. For column d, you need to analyze the data appropriately, taking into account the following: 1) the use of a 2-stage cluster design that results in correlated data at the cluster / PSU level; 2) adjustment for differences in the ultimate probability of selection through sampling weights; and 3) uses the strata information to improve the precision of your estimates. Please note variable **V005 is the sampling weight**, **variable V021, labeled as Primary Sampling Unit, provides the cluster number** for the analysis, and **variable V022, labeled Sample stratum number**, provides the strata number for the analysis.

**In each cell, include the proportion and the standard error (round to 3 decimal points)**

Table 2: Proportion of children that slept under an ITN the previous night, among all households

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n = | (a) Assuming SRS | (b) Assuming SRS, with weights | (c) 2-stage cluster sampling with weights | (d) 2-stage cluster sampling with weights and stratification |
| Child age |  |  |  |  |
| 0 | 0.375 (0.014) | 0.360 (0.015) | 0.360 (0.021) | 0.360 (0.020) |
| 1 | 0.349 (0.014) | 0.334 (0.015) | 0.334 (0.020) | 0.334 (0.019) |
| 2 | 0.285 (0.014) | 0.270 (0.015) | 0.270 (0.018) | 0.270 (0.017) |
| 3 | 0.244(0.014) | 0.231 (0.015) | 0.231 (0.018\_ | 0.231 (0.017) |
| 4 | 0.207 (0.013) | 0.192 (0.013) | 0.192 (0.015) | 0.192 (0.014) |
| Residence |  |  |  |  |
| Urban | 0.288 (0.011) | 0.264 (0.013) | 0.264 (0.020) | 0.264 (0.019) |
| Rural | 0.301 (0.008) | 0.290 (0.008) | 0.290 (0.017) | 0.290 (0.015) |
| SES |  |  |  |  |
| Poorest | 0.216 (0.012) | 0.199 (0.012) | 0.199 (0.019) | 0.199 (0.018) |
| Poorer | 0.328(0.014) | 0.325 (0.014) | 0.325 (0.023) | 0.325 (0.022) |
| Middle | 0.325 (0.014) | 0.334 (0.015) | 0.334 (0.023) | 0.334 (0.022) |
| Richer | 0.301 (0.014) | 0.287 (0.016) | 0.287 (0.024) | 0.287 (0.024) |
| Richest | 0.317 (0.018) | 0.270 (0.019) | 0.270 (0.028) | 0.270 (0.027) |
| Mother’s education |  |  |  |  |
| None | 0.230 (0.016) | 0.228 (0.017) | 0.228 (0.025) | 0.228 (0.025) |
| Primary | 0.293 (0.008) | 0.281 (0.008) | 0.281 (0.016) | 0.281 (0.014) |
| Secondary | 0.334 (0.014) | 0.308 (0.015) | 0.308 (0.021) | 0.308 (0.020) |
| Higher | 0.437 (0.048) | 0.414 (0.054) | 0.414 (0.057) | 0.414 (0.058) |
| **All Households** | **0.296 (0.006)** | **0.282 (0.007)** | **0.282 (0.013)** | **0.282 (0.012)** |

1. How did the inclusion of the sampling strata in the analysis affect the precision of the estimates?

Clustering greatly increases standard error, decreasing the precision of the point estimate. However, including strata seems to minimally decrease this standard error.

1. What is the **design effect** for this 2-stage cluster sampling design for the proportion of children that slept under an ITN, with sample weights and stratification included in the analysis?

0.012/.006= 2

1. How does child age affect the use of ITNs among children?

The younger the child, the more likely they slept under an ITN the night previous.

1. How would you explain the level of children that slept under an ITN to the Ministry of Health- is it high or low, and how would you interpret the point estimate taking total survey error into account?

Overall, I would say the level of children that slept under an ITN was low, with point estimates, using all sampling strategies, below 30%. Weighting, clustering, and stratifying only lowered this point estimate. Standard error is relatively small, even at its highest at 0.013 using 2-stage cluster sampling with weights. This produces narrow confidence intervals, with 95% confidence intervals revealing 95% of the time, using any of the sampling strategies analyzed, the proportion of children who slept under an ITN the night previous in a given sample will be less than 31%. Non-sampling error cannot be measured, and potential sources of bias, such as non-response bias, should be considered when interpreting these estimates.