**IHSD 7440 - Homework #3 2023**

**Name\_\_Claire Hubley\_\_\_**

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)?

The unit of analysis is households, n=6439

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

285

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

proportional to size - Range=6-48

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

16

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

16

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, se: 0.0100 | 0.520, se: 0.012 | 0.520, Se: 0.019 | 0.520, se: 0.018 |
| Rural | 0.572, se: 0.008 | 0.556, se: 0.008 | 0.556, Se: 0.018 | 0.556, se: 0.018 |
| SES |  |  |  |  |
| Poorest | 0.506, se: 0015 | 0.481, se: 0.016 | 0.481, se: 0.028 | 0.481, se: 0.027 |
| Poorer | 0.529, se: 0.013 | 0.510, se: 0.015 | 0.510, se: 0.023 | 0.510, se: 0.022 |
| Middle | 0.582, se: 0.013 | 0.585, se: 0.014 | 0.585, se: 0.022 | 0.585, se: 0.020 |
| Richer | 0.560, se: 0.013 | 0.549, se: 0.014 | 0.549, se: 0.025 | 0.549, se: 0.023 |
| Richest | 0.623, se: 0.015 | 0.589, se: 0.017 | 0.589, se: 0.021 | 0.589, se: 0.020 |
| HH head education |  |  |  |  |
| None | 0. 421, se: 0.016 | 0.405, se: 0.017 | 0.405, se: 0.026 | 0.405, se: 0.025 |
| Primary | 0.546, se: 0.009 | 0.528, se:0.010 | 0.528, se:0.018 | 0.528, se: 0.017 |
| Secondary | 0.612, se: 0.011 | 0.599, se:0.013 | 0.599, se:0.018 | 0.599, se: 0.016 |
| Higher | 0.704, se: 0.020 | 0.681, se: 0.023 | 0.681, se: 0.022 | 0.681, se: 0.022 |
| **All Households** | **6.195/11=0.563**  **Se: 0.143/11=0.013** | **0.546 se: 0.014** | **0.546 0.022** | **0.546 se: 0.021** |

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

Compared to SRS, sampling with weights and SRS lowers the point estimates but increases the standards error. However, compared to SRS with weights, 2-stage cluster sampling with weights and with stratification does not change point estimates. Cluster sampling with just weights raises standard errors, but cluster sampling with stratification produces lower standard errors than 2-stange cluster sampling with just weights.

* + 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)?

The standard error increases.

* + 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?

(0.021/0.013) = 1.615

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

More rural houses surveyed had at least 1 ITN than urban households.

* + 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?

2-stage cluster sampling with weights and stratification because it uses the right sampling design in analysis and has the lowest standard errors by taking into account clusters and the different strata within the clusters.

**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)?

Children, 5,194

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 = 5,194 | (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 = 1128 | 0.375, se: 0.014 | 0.360, se: 0.015 | 0.360, se: 0.021 | 0.360, se: 0.020 |
| 1 = 1121 | 0.349, se: 0.014 | 0.333, se: 0.015 | 0.333, se: 0.020 | 0.333, se: 0.019 |
| 2 = 1013 | 0.286, se: 0.014 | 0.270, se: 0.015 | 0.270, se: 0.018 | 0.270, se: 0.017 |
| 3 = 963 | 0.244, se: 0.014 | 0.231, se: 0.015 | 0.231, se: 0.018 | 0.231, se: 0.017 |
| 4 = 969 | 0.207, se: 0.013 | 0.192, se: 0.013 | 0.192, se: 0.015 | 0.192, se: 0.014 |
| Residence |  |  |  |  |
| Urban = 1711 | 0.288, se: 0.011 | 0.264, se:0.013 | 0.264, se: 0.020 | 0.264, se: 0.019 |
| Rural = 3483 | 0.301, se: 0.008 | 0.290, se:0.008 | 0.290, se: 0.027 | 0.290, se: 0.015 |
| SES |  |  |  |  |
| Poorest = 1122 | 0.216, se: 0.012 | 0.199, se: 0.012 | 0.199, se: 0.019 | 0.199, se: 0.018 |
| Poorer = 1029 | 0.301, se: 0.014 | 0.287, se: 0.016 | 0.287, se: 0.024 | 0.287, se: 0.024 |
| Middle = 1162 | 0.325, se: 0.014 | 0.333, se: 0.015 | 0.333, se: 0.023 | 0.333, se: 0.022 |
| Richer = 1179 | 0.328, se: 0.014 | 0.325, se: 0.014 | 0.325, se: 0.023 | 0.325, se: 0.022 |
| Richest = 682 | 0.317, se: 0.018 | 0.270, se: 0.019 | 0.270, se: 0.028 | 0.270, se: 0.027 |
| Mother’s education |  |  |  |  |
| None = 700 | 0.230, se: 0.016 | 0.228, se: 0.017 | 0.228, se: 0.025 | 0.228, se: 0.025 |
| Primary = 3245 | 0.293, se: 0.008 | 0.281, se: 0.008 | 0.281, se: 0.016 | 0.281, se: 0.014 |
| Secondary = 1139 | 0.334, se: 0.014 | 0.308, se: 0.015 | 0.308, se: 0.021 | 0.308, se: 0.020 |
| Higher = 110 | 0.436, se: 0.048 | 0.414, se: 0.054 | 0.414, se: 0.057 | 0.414, se: 0.058 |
| **All Households** | **0.302, se: 0.015** | **0.287, se: 0.017** | **0.287, se: 0.023** | **0.287, 0.022** |

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

Including the sampling strata made the data more precise.

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?

(0.022/0.015) = 1.467

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

Younger children are more likely to have an ITN for use than older children.

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?

Out of all households surveyed in Zambia, only about 29% of children living in them had slept under a ITN the previous night after being surveyed. The data was gathered using a strategy taking into account the different factors associated with ITN use such as type of residence, the mother’s education level, the child’s age, and the family’s socio-economic status. While this sampling strategy has a higher margin of error than other designs, it ensures that the proportion measured does not over estimate the use of ITNs for certain factors or groups of people that might be smaller in nature than others.