MODULE 6

ANALYZING MICRODATA WITH A GENDER ANGLE

TRAINING SYLLABUS

Curriculum on Gender Statistics Training

This product was developed under the guidance of the Subgroup on Gender Statistics Training, within the Asia-Pacific Network of Statistical Training Institutes.

Introduction

This syllabus has been designed to guide trainers on how to conduct related training. The syllabus can also be used by trainees who wish to know more about this topic and people who are generally interested in analyzing gender data from household surveys.

This syllabus is part of a wider module on this area of gender statistics. Other materials within this module might include exercises, sample datasets, PowerPoint presentations and example quizzes. Please refer to the additional set of materials for a comprehensive and effective learning experience.

Who is this module for?

- Statisticians and other experts that wish to analyze data generated from household surveys with a gender angle
- Policymakers and decision-makers who are looking to conduct their own data analysis to enhance their use of gender data for evidence-based decision-making
- Academics who wish to use to this module as teaching materials for gender analysis in classrooms
- Civil society organizations who wish to enhance their skills in analyzing gender data for advocacy or communication purposes
- Anyone who has plans to give trainings on analysis of survey data with a gender perspective

What do I need to know before going through this module?

This is an intermediate module on gender statistics mainly targeted to applied statisticians and gender policy analysts involved in gender data analyses. No advanced knowledge of statistics is necessary. However, it would be helpful for the trainee to have basic knowledge of statistical estimation, significance testing and regression modeling.

Learning objectives

The expected learning outcomes for this module include:

- Performing statistical analysis beyond descriptive statistics, such as regression analysis, to disentangle the complexity of the multidimensional aspects of gender statistics.
- Becoming familiar with the preparatory steps required to construct a dataset from household surveys for further statistical analysis. These preparatory steps involve understanding the data structure, sampling design, weighting scheme, treatment of missing data and merging of several datasets if variables of interest are presented in two or more questionnaires.
- Providing hands-on experience to analyze gender related microdata from household surveys using R and STATA, both of which are statistical software widely used for statistical modeling and data analysis. The trainee is exposed to statistical programming (e.g. simple data manipulation) and basic statistical analysis skills in R and STATA.

Note to trainer: This module is conducted through hands-on exercises by analyzing the Multiple Indicator Cluster Survey (MICS) microdata for illustration purposes. It also assumes that trainees do not have previous experience using R or STATA and applying regression models. Depending on the trainees' familiarity with R, STATA or other statistical software and knowledge of regression modeling, it is expected that training for this module can be delivered in 2 to 3 hours. This module is practically oriented with the aim of giving trainees some exposure to regression data analysis (logistic regression). Trainees should refer to the list of existing resources provided in this module for deeper understanding of theoretical assumptions on some of the statistical methodologies introduced in this module.

Table of Contents

Sources of data for gender analysis

Household surveys, such as Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), Labour Force Surveys (LFS), Household Income and Expenditure Surveys (HIES) and others, are important sources of gender data for measuring inequalities in women and men's lives. In comparison to censuses, they are conducted more frequently and are less costly. They are also a key data source for Sustainable Development Goals (SDGs) monitoring. Many of these household surveys contain a separate survey module for women, which collects key information needed for in-depth gender analysis. For example, in addition to basic socio demographic variables for women, DHS and MICS have survey modules that collect data on violence against women, thus providing practical alternatives for measuring intimate partner violence against women when financial resources are limited. Dedicated surveys, such as victimization surveys or specialized surveys on violence against women are the preferred method of data collection for prevalence of violence data. However, these dedicated surveys can be relatively expensive and require careful selection and focused training of the interviewers[1]. Thus, related modules attached to standardized surveys such as MICS and DHS provide a useful alternative to quantify these issues.

Censuses are also important sources of gender data. Although censuses provide key variables related to women and girls in entire geographical units of the country, most censuses are conducted only every 10 years. Administrative records (e.g. school records, police records, etc.) are becoming increasingly popular as sources of data for gender analysis. However, there are challenges associated with the quality and comprehensiveness of administrative records in some contexts.

For illustration purposes, this training module uses the Multiple Indicator Cluster Survey (MICS) to showcase key elements to be considered when analyzing gender issues with microdata.

Understanding MICS data

The first round of MICS was conducted around 1995 in more than 60 countries to provide internationally comparable data on women and children. UNICEF manages the global MICS programme and provides technical support and necessary trainings to countries to conduct MICS. The latest (6th) round of MICS data (hereinafter referred to as MICS6) questionnaires and summary reports are available at the MICS website[2].

MICS6 collects data using several questionnaires (household, individual women age 15–49 years, individual men age 15–49 years, children under five and children age 5–17 years). Each of the questionnaires are further disaggregated into several modules. For example, the questionnaire for individual women contains modules that attempt to capture information on the following areas: women's background, birth history, use of mass media and ICT, maternal and newborn health, contraception, marriage, sexual behavior, maternal mortality, tobacco and alcohol use, life satisfaction, victimization and attitudes towards domestic violence.

[1] see UNSD. 2016. https://unstats.un.org/unsd/demographic/standmeth/handbooks/05323 Integrating a Gender Perspective into Statistics Web Final.pdf

[2] https://mics.unicef.org/about

Analyzing factors that influence attitudes towards domestic violence

As mentioned above, MICS6 collects data on attitudes towards domestic violence. This question has been included in MICS since round four, acknowledging the seriousness and wide prevalence of violence against women by an intimate partner. The data shows that there is a wide variation on attitudes towards wife- beating across countries and within them. Further analysis at country level needs to be performed to understand the underlying factors that contribute to the formation of attitudes towards domestic violence. Because some academic studies suggest that less-educated women are more likely to tolerate attitudes on wife-beating than women with higher level of education (Uthman, Lawoko and Moradi, 2009[3]), this statement will be used as a means of illustrating how to use microdata to consider the accuracy of a hypothesis.

Thus, for the purpose of demonstrating each step involved in analyzing gender data from MICS6, this module attempts to investigate factors that influence the attitudes of both women and men towards domestic violence by analyzing MICS6 data of Mongolia. Country specific MICS6 datasets can be accessed for completed surveys at the official MICS website[4].

Getting ready for statistical analysis

Before jumping into performing any statistical analysis, it is necessary to prepare the dataset so that it is ready to be read into R or STATA for further statistical analysis. In order to get the data ready for analysis, the response variable ('attitudes towards domestic violence') needs to be reconstructed. MICS6 has five questions to assess the attitudes towards domestic violence. Namely: "a husband is justified in hitting or beating his wife in the following situations (1) she goes out without telling him (2) she neglects the children (3) she argues with him (4) she refuses to have sex with him (5) she burns the food". The same questions were asked to both women and men. In order to turn this five questions into a single variable with dichotomous values (yes or no), the response variable ('attitudes towards domestic violence') is coded as 1 if at least one of the five questions have a 'yes' response and coded as 0 only if all five questions have 'No' as responses. This step is needed in order to perform the logistics regression explained in the later section of this module.

Furthermore, the following variables are considered in order to assess whether or not they are connected with attitudes towards domestic violence in Mongolia: wealth index, age, education level, place of residence (urban or rural) and marital status. These variables are scattered across several questionnaires, which requires merging of datasets in order to create a single dataset with all the desired variables prior to starting the analysis. When creating a single dataset, one also needs to carefully consider the missing observations. If these are occurring at random, removal is the best solution. However, if there is a pattern (e.g. the observations are missing for a particular reason or population group) removing the observations might produce biased estimates. In this case, advanced statistical techniques might be needed to impute the missing data, which is beyond the scope of this module. Preparing the data for statistical analysis might be tedious and challenging work, but this important step is too often overlooked. Fortunately, the dedicated experts at the Global MICS Team have already cleaned, checked and prepared the microdata, so when downloaded

[3] https://bmcinthealthhumrights.biomedcentral.com/articles/10.1186/1472-698X-9-14
[4] https://mics.unicef.org/surveys

from the MICS website, it is ready for use. Users, however, still need to perform data transformations, such as merging datasets and creating new variables, according to the objectives of the analysis. Complete R and STATA codes used for preparing the data on attitudes towards domestic violence can be found in the appendix.

Considering sampling design and sample weights

MICS6 takes a two-stage stratified cluster sampling approach for the selection of the survey sample. This type of sampling approach is very common in many household surveys. For MICS6, the data of Mongolia's 4,444 clusters (primary sampling units) are selected from the census enumeration areas (2,805 urban and 1,639 rural) in the first stage. For the second stage, a complete household listing is carried out in each selected cluster and a sample of 889,817 households (579,314 urban and 310,503 rural) is selected.

When performing statistical analysis to sample survey data, in most cases, the data must be weighted. This is because the overall probability of selection of each household is not constant. Simple weighting might be enough when doing simple tabulation on an indicator (e.g. sum of women or men). However, for any analysis that involves estimation of standard errors, confidence intervals or significance testing, it is important to consider the complex sample design parameters, such as the primary sampling units (PSUs), the stratification variable and the sampling weights. R has "survey" package that enables users to perform analysis for multistage stratified cluster samples. Similarly, STATA has a survey set command, "svy", that allows users to share with the software the characteristics of the sample prior to performing this analysis. In this module, all analyses are conducted utilizing R and STATA, and considering the sampling weights and sample design parameters. (see appendix for R and STATA codes)

Correlation between the variables

It's always a good practice to explore the general structure of the data before performing any advanced statistical analysis. Do 'place of residence (urban or rural)' and 'wealth' have any relationship? If there is a pattern between them, is it a positive or a negative relationship? Correlation is a useful numerical summary of the strength of a relationship between two variables ranging from -1 to 1. Using R/STATA, we can create a correlation plot for the explanatory variables. Figure 1 and 2 show the correlation plot of explanatory variables (wealth index, age, education level, place of residence (urban or rural) and marital status) for women and men respectively. The size of the dot represents the strength of the relationship (the bigger the size, the stronger the relationship). The blue and red color represents the direction of the relationship. The darker blue color indicates a stronger positive relationship and the darker red color indicates a stronger negative relationship.

In Figure 1, the variables displayed include:

- UBN.UBN1 – Urban area

- UBN.UBN0 – Rural area
- AGE – Age of respondent
- EDU.EDU2 – Respondent only completed secondary education or below
- EDU.EDU3 – Respondent completed vocational
- EDU.EDU4 – Respondent completed college, university
- windex5.windex51 – Wealth index quintile (poorest)
- windex5.windex52 – Wealth index quintile (second)
- windex5.windex53 – Wealth index quintile (middle)
- windex5.windex54 – Wealth index quintile (fourth)
- windex5.windex55 – Wealth index quintile (richest)
- MA.MA0 – Not in union
- MA.MA1 – Currently married or living with a partner

As Figure 1 and 2 show, for both women and men, education is correlated with different levels of wealth. As expected, the highest level of education (EDU.EDU4) is positively correlated with the top wealth quintile (windex.windex55) and negatively correlated with the bottom wealth quintile (windex.windex51).

For both women and men, the place of residence (urban or rural) is also correlated with wealth. Urban area (UBN.UBN1) is positively correlated with top wealth quintile (windex.windex55), whereas rural area (UBN.UBN0) is positively correlated with the bottom wealth quintile (windex.windex51).

As expected, age is positively correlated with marital status: AGE is positively correlated with 'currently married[5]' (MA.MA1) and negatively correlated with "not in union" (MA.MA0).

Both education and place of residence (urban/rural) are correlated with wealth. This creates redundancy in the information contained in explanatory variables. This redundancy needs to be addressed when building the logistic regression model in the later part of this module.

[5] Currently married includes "living with a partner"

Figure 1. Correlation plot of explanatory variables for women

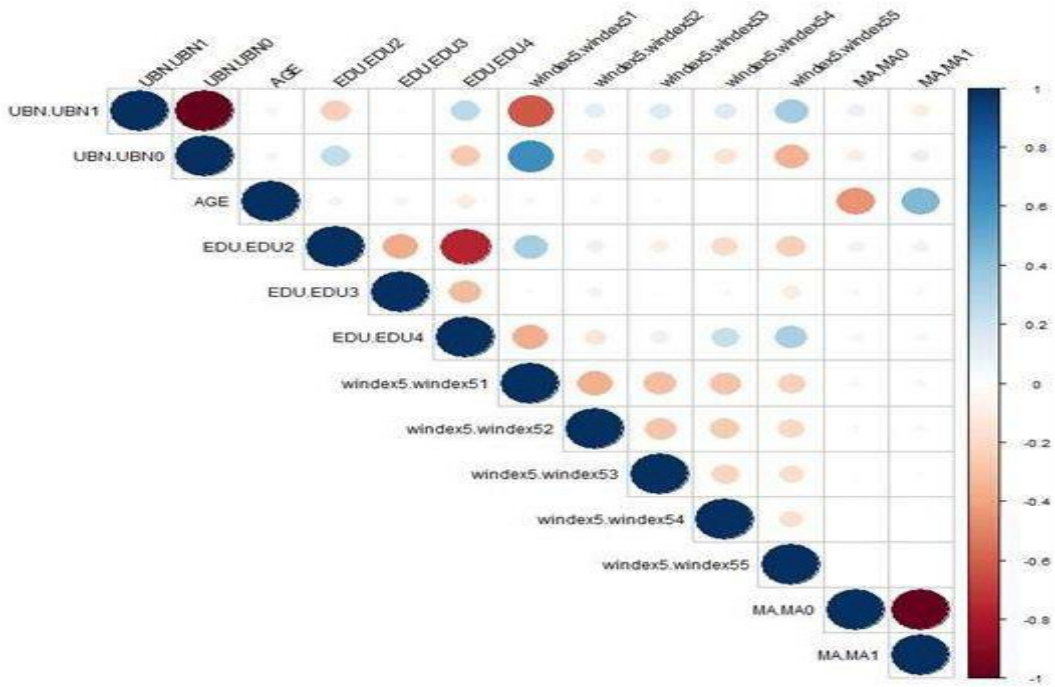
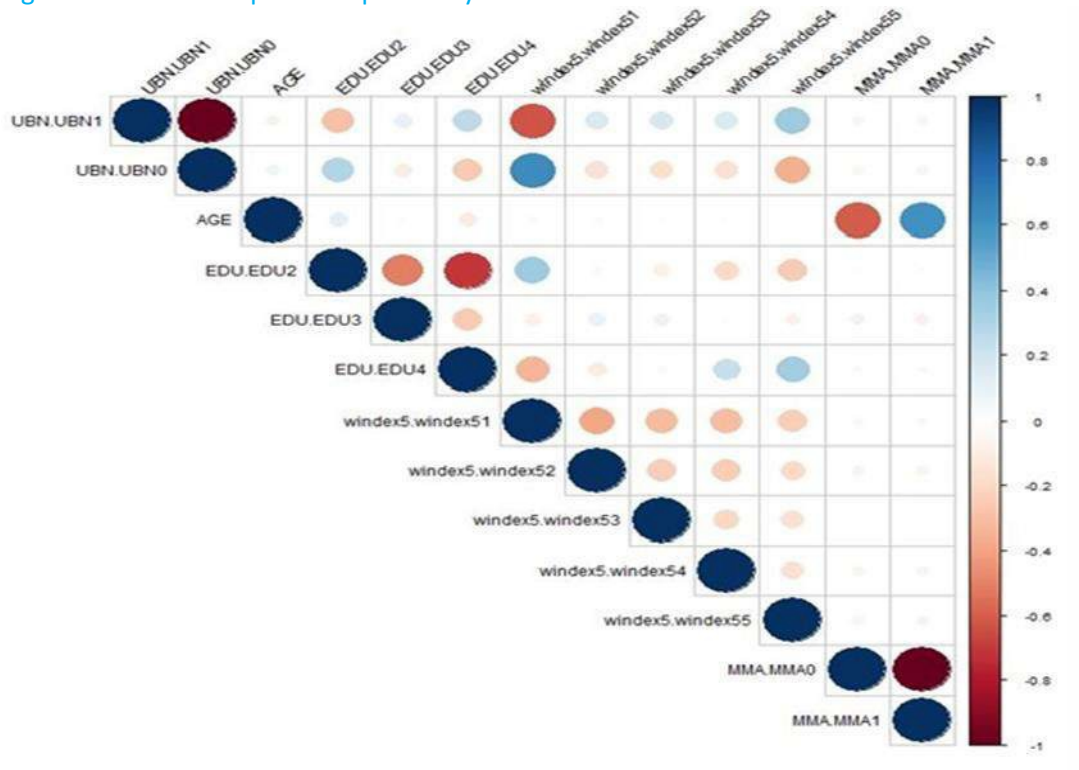


Figure 2. Correlation plot of explanatory variables for men



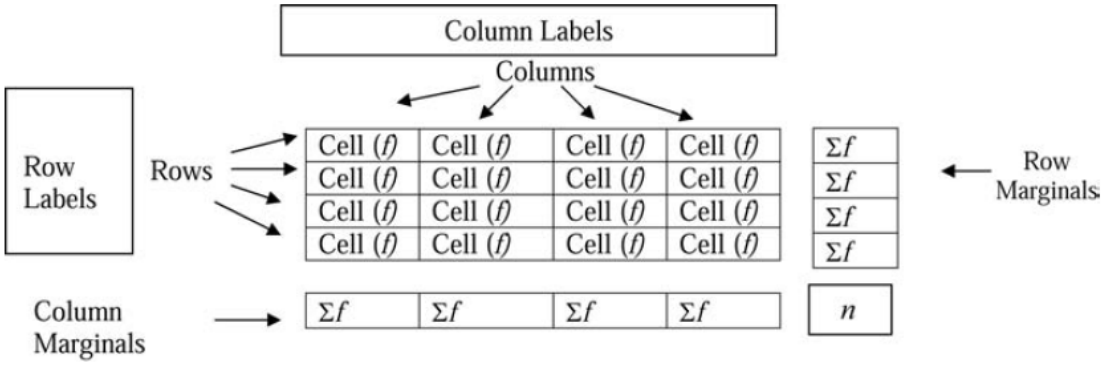
Cross tabulations

Cross tabulations are tables that group variables to assess the relationships between them (e.g. education and wealth in two-way tables, or education, wealth and sex in three-way tables). These tables that display information for more than one variable are called cross-tabulations or, for short, crosstabs. Such cross-tabulations enable disaggregation of data by variables of interest.

When two (or more) variables are used together for descriptive purposes, the procedures used are referred to as measures of association. It is often necessary to decide, in bivariate applications, which of the two is the independent and which is the dependent variable.

The reason why the independent and dependent variables must be identified lies in the structure of a crosstab table. As indicated in Table 1, a crosstab, like any other table, consists of rows and columns. At the intersection of each row and column is a cell. And in each cell we find a frequency. The most widely used convention is to use the horizontal direction for the independent variable and to have each column represent an attribute of that variable. Similarly, the vertical direction is used for the dependent variable, with each row representing one of its attributes. All crosstabs include an extra row, column, and cell, referred to as marginal because they are on the margins of the table. These are used to show the totals.

Table 1. The structure of the Crosstab



One way table

A one-way table (also called a frequency table) shows frequency (absolute) counts or relative frequencies of a single categorical variable.

When a one-way table shows relative frequencies (i.e., proportions), the sum of the column should equal to one. If the relative frequencies are multiplied by 100, then it becomes percentages. The sum of the column values/ percentages should equal to 100. Table 2 below shows a one-way table with count frequencies, relative frequencies, and percentages of travel choices of 10 clients.

Table 2: One-way table with count frequency

Place of preference for travel	Frequency (absolute count)	Proportion	Percentage
USA	5	0.5 (5/10=0.5)	50 (5/10*100=50%)
Europe	3	0.3 (3/10=0.3)	30 (3/10*100=30%)
Asia	2	0.2 (2/10=0.2)	20 (2/10*100=20%)
Total	10	1 (0.5+0.3+0.2)	100 (50%+20%+30%)

Interpretation: 50% of the clients prefer to travel to the USA, whereas 30% of clients want to travel to Europe and 20% of clients want to travel to Asia.

A two-way table examines the relationship between two categorical variables. Entries in the cells can be displayed as frequency counts or as relative frequencies (just like a one-way table). Below, Table 3 shows the favorite leisure activities for 50 adults – 20 men and 30 women.

Table 3: Two-way table

Sex	Favorite leisure Activity			Total (Row)
	Dance	Sports	TV	
Women	16	6	8	30 *(16+6+8)*
Men	2	10	8	20

Total (Column)	18 *(16+2)*	16	16	50

Entries in the Total (Column) and Total (Row) are called marginal frequencies. They should have the same sum. In this table, the sum is 50. Entries in the body of the table are called joint frequencies.

If we looked only at the marginal frequencies in the Total (Column), we might conclude that the three activities had roughly equal appeal. Yet, the joint frequencies show a strong preference for dance among women; and little interest in dance among men.

Two-Way Relative Frequency Tables

Two-way tables can show relative frequencies for the whole table, for rows, or for columns. The following tables relative frequencies (proportions) for the whole table (Table 4), for table rows (Table 5), and for table columns (Table 6).

Table 4: Two-way table frequency (proportion) for the whole table

Sex	Activity			Total (Row)
	Dance	Sports	TV	
Women	0.32 *(16/50)*	0.12 *(6/50)*	0.16 *(8/50)*	0.60 *(0.32+0.12+0.16)*
Men	0.04 *(2/50)*	0.20 *(10/50)*	0.16 *(8/50)*	0.40
Total (Column)	0.36 *(0.32+0.04)*	0.32	0.32	1.00

Note: The denominator used is 50 as it is the total count of observations in the table

For interpretation, we can say 36% of the respondents prefer dancing, 32% of them prefer doing sports and 32% of them prefer watching TV.

Table 5: Two-way table frequency for the whole table (row)

Sex	Activity			Total (Row)
	Dance	Sports	TV	
Women	0.53 *(16/30)*	0.20 *(6/30)*	0.27 *(8/30)*	1.00*(0.53+0.20+0.27*)
Men	0.10*(2/20)*	0.50 *(10/20)*	0.40 *(8/20)*	1.00*(0.10+0.50+0.40)*

Note: The table above shows the row percentages. They were derived by dividing the count for each cell (joint frequency) by the total number of observations in each row. The row percentages therefore showed the preferred activities (column categories) among women and men (observations in row) as percentages.

For interpretation, we can say the most preferred activity among women is dancing (53%), whereas sports are the least preferred activity among women (20%). However, for men, the most preferred activity is doing sports, whereas dancing is the last preferred activity (10%).

Table 6: Two-way table frequency for the whole table (column)

Sex	Activity		
	Dance	Sports	TV
Women	0.89 (16/18)	0.38 (6/16)	0.50 (8/16)
Men	0.11 (2/18)	0.63 (10/16)	0.50 (8/16)
Total (Column)	1.00	1.00	1.00

Note: The table above shows the column percentages, which were computed by dividing the count for each cell (joint frequency) by the total number of observations in each column. The column percentages showed the distribution (as percentages) of observations in each column (women and men) among those in the column (preferred activity).

For interpretation, more women (89%) than men (11%) prefer dancing. For doing sports, more men prefer (63%) it than women (38%). As for TV, equal proportions of women and men prefer this activity (50% respectively).

Each type of relative frequency table makes a different contribution to understanding the relationship between gender and preferences for leisure activities. For example, the "Relative Frequency for Rows" table most clearly shows the probability that each sex will prefer a particular leisure activity. It is easy to see that the probability that a man will prefer dance is 10%; the probability that a woman will prefer dance is 53%; the probability that a man will prefer sports is 50%; and so on.

Appropriate sample weights need to be considered when creating cross tabulations from survey data. These weights help adjust for disproportionate sampling and non-response. In other words, weights help restore the representativeness of the sample.

Different weights are calculated depending on the different units of analysis. Therefore, it is important to apply the correct weights to the unit of analysis. For example, if the unit of analysis is household, then household weight should be used. If the unit of analysis is women, then women's weight should be used. The DHS domestic violence survey module has its own weight, which is different from the women's weight[6]. The unit of analysis of the MICS6 'attitudes towards domestic violence' data is women, and accordingly women's weight is used. Detailed information on sample weights should be included in the final report of household surveys so that analysts can apply the correct weights in their statistical analysis.

As mentioned before, in our analysis, the variable "attitudes towards domestic violence (DV)" is coded as 1 if at least one of the five questions, which are designed to measure attitudes towards domestic violence, have a 'yes' response and coded as 0 only if all five questions have 'no' as

[6] The DHS domestic violence questionnaire module selects only one woman per household. On the other hand, the woman's questionnaire surveys women aged 15 to 49 in sampled households. Thus, they have different units of analysis and sample weights.

responses. Applying the sample design and sample weights, the cross-tabulations of attitudes towards domestic violence (DV) and multiple variables of interest in the dataset are performed to examine the relationship between the variables. Both STATA and R output are provided for this exercise.

The results of the cross-tabulation between DV and education show that the proportion of accepting attitudes towards domestic violence is lowest (8.6%) for the women with the highest educational level (EDU=4) and highest (18.4%) for women with the lowest educational level (EDU=2).

Figure 3: Cross-tabulation between attitudes towards DV and Education, R output

```
colPercents(svytable(~DV+EDU, wm5design,round=TRUE))
```

```
##             EDU
## DV              2      3      4
##   0          81.6   82.7   91.4
##   1          18.4   17.3    8.6
##   Total     100.0  100.0  100.0
##   Count    4338.0 1185.0 5144.0
```

Figure 4: Cross-tabulation between attitudes towards DV and Education, STATA output

```
. svy: tabulate dv EDU, format(%14.1f) column percent
(running tabulate on estimation sample)

Number of strata   =        23                  Number of obs      =     10593
Number of PSUs     =       580                  Population size    = 10668.125
                                                Design df          =       557
```

dv	EDU 2	3	4	Total
0	81.6	82.7	91.4	86.5
1	18.4	17.3	8.6	13.5
Total	100.0	100.0	100.0	100.0

The cross-tabulation of DV against different levels (quintile) of wealth, shown below, suggests that the proportion of women accepting attitudes towards domestic violence steadily decreases (23.5% to 6.7%) as we move from the bottom (windex5=1) to top (windex5=5) wealth quintile.

Figure 5: Cross-tabulation between attitudes towards DV and wealth, R output

```
colPercents(svytable(~DV+windex5, wm5design,round=TRUE))
```

```
##             windex5
## DV               1      2      3      4      5
##    0          76.5   84.4   85.6   90.8   93.3
##    1          23.5   15.6   14.4    9.2    6.7
##    Total     100.0  100.0  100.0  100.0  100.0
##    Count    1929.0 1970.0 2215.0 2218.0 2336.0
```

Figure 6: Cross-tabulation between attitudes towards DV and wealth, STATA output

```
. svy: tabulate dv windex, format(%14.1f) column percent
(running tabulate on estimation sample)

Number of strata   =        23                  Number of obs      =     10593
Number of PSUs     =       580                  Population size    = 10668.125
                                                Design df          =       557

           |                    windex
        dv |         1          2          3          4          5      Total
-----------+------------------------------------------------------------------
         0 |      76.5       84.4       85.6       90.8       93.3       86.5
         1 |      23.5       15.6       14.4        9.2        6.7       13.5
           |
     Total |     100.0      100.0      100.0      100.0      100.0      100.0
-----------------------------------------------------------------------------
```

The proportion of women living in urban areas (UBN=1) with accepting attitudes towards domestic violence (10.9%) is lower than their counterparts residing in rural (UBN=0) areas (19.6%).

Figure 7: Cross-tabulation between attitudes towards DV and area of residence (urban), R output

```
colPercents(svytable(~DV+UBN, wm5design,round=TRUE))
```

```
##             UBN
## DV               1      0
##    0          89.1   80.4
##    1          10.9   19.6
##    Total     100.0  100.0
##    Count    7454.0 3214.0
```

Figure 8: Cross-tabulation between attitudes towards DV and area of residence (urban), STATA output

```
. svy: tabulate dv residence, format(%14.1f) column percent
(running tabulate on estimation sample)

Number of strata   =        23          Number of obs      =     10593
Number of PSUs     =       580          Population size    =  10668.125
                                        Design df          =       557
```

dv	residence 0	1	Total
0	80.4	89.1	86.5
1	19.6	10.9	13.5
Total	100.0	100.0	100.0

The proportion of women not in a union (MA=0) with accepting attitudes towards domestic violence (15.8%) is slightly higher than for women in a union (MA=1, 12.6%).

Figure 9: Cross-tabulation between attitudes towards DV and marital status, R output

```
colPercents(svytable(~DV+MA, wm5design,round=TRUE))
```

```
##            MA
## DV            0      1
##    0       84.2   87.4
##    1       15.8   12.6
##    Total  100.0  100.0
##    Count 3016.0 7652.0
```

Figure 10: Cross-tabulation between attitudes towards DV and marital status, STATA output

```
. svy: tabulate dv ma, format(%14.1f) column percent
(running tabulate on estimation sample)

Number of strata   =        23                  Number of obs      =     10593
Number of PSUs     =       580                  Population size    = 10668.125
                                                Design df          =       557
```

dv	ma 0	ma 1	Total
0	84.2	87.4	86.5
1	15.8	12.6	13.5
Total	100.0	100.0	100.0

As shown below, most of the women in the bottom wealth quintile (windex5=1) live in rural (UBN=0) areas (95.7%) whereas that same figure is only 0.1% for the women in the top wealth quintile (windex5=5).

Figure 11: Cross-tabulation between wealth quintile and area of residence (urban), R output

```
colPercents(svytable(~UBN+windex5, wm5design,round=TRUE))

##              windex5
##  UBN             1      2      3      4      5
##    1           4.3   71.7   80.9   82.6   99.9
##    0          95.7   28.3   19.1   17.4    0.1
##        Total 100.0  100.0  100.0  100.0  100.0
##        Count 1929.0 1970.0 2215.0 2218.0 2337.0
```

Figure 12: Cross-tabulation between wealth quintile and area of residence (urban), STATA output

```
. svy: tabulate residence wealth, format(%14.1f) column percent
(running tabulate on estimation sample)

Number of strata   =        23                  Number of obs      =     10593
Number of PSUs     =       580                  Population size    = 10668.125
                                                Design df          =       557

           |                    wealth
 residence |         1         2         3         4         5     Total
-----------+------------------------------------------------------------
         0 |      95.7      28.3      19.0      17.4       0.1      30.1
         1 |       4.3      71.7      81.0      82.6      99.9      69.9
-----------+------------------------------------------------------------
     Total |     100.0     100.0     100.0     100.0     100.0     100.0
```

Lastly, women with the highest educational level (EDU=4) are the most likely (37.7%) to live in households in the top wealth quintile (windex5=5). The same tables can also be produced for men and similar patterns are observed.

Figure 13: Cross-tabulation between educational level and wealth quintile, R output

```
colPercents(svytable(~windex5+EDU, wm5design,round=TRUE))

##             EDU
##  windex5      2      3      4
##        1   33.6   22.1    4.1
##        2   25.4   26.7   10.8
##        3   20.5   22.6   20.6
##        4   13.6   21.1   26.8
##        5    7.1    7.5   37.7
##    Total  100.2  100.0  100.0
##    Count 4339.0 1185.0 5145.0
```

Figure 14: Cross-tabulation between educational level and wealth quintile, STATA output

```
. svy: tabulate wealth education, format(%14.1f) column percent
(running tabulate on estimation sample)

Number of strata   =        23                  Number of obs      =     10593
Number of PSUs     =       580                  Population size    = 10668.125
                                                Design df          =       557
```

wealth	education 2	3	4	Total
1	33.6	22.1	4.1	18.1
2	25.4	26.7	10.8	18.5
3	20.5	22.6	20.6	20.8
4	13.6	21.0	26.8	20.8
5	7.1	7.5	37.7	21.9
Total	100.0	100.0	100.0	100.0

Overview of Logistic regression analysis

Not one single factor but multiple factors influence the attitudes of both women and men towards domestic violence. Regression analysis is a powerful statistical method that can be used to clarify the effects of multiple factors that have influence on the attitudes of women and men towards domestic violence in Mongolia. Does education level have any impact on attitudes towards domestic violence? How about the place of residence (urban or rural)? How about the level of wealth? The usual descriptive statistics are limited to determine which factors matter most and which factors can be ignored in this situation. Regression analysis provides answers to these questions by separating the impact of each of these variables on the attitudes towards domestic violence. While controlling for the effect of all other explanatory variables being considered (*ceteris paribus*), it enables us to measure the effect of each separate explanatory variable (also known as independent variable), such as the wealth index (quintiles)[7], age, education level, urban/rural and marital status on the response variable (also known as dependent variable), 'attitudes towards domestic violence' (yes or no)[8].

Many types of regression analysis can be used, depending on the types of dependent variables. Categorical scales are very common in household surveys, such as MICS6. The response and explanatory variables considered as an illustration in this module are all categorical – except for age, which is a

[7] The wealth index is a composite measure of a household's cumulative living standard, which is equivalent to the one used in the DHS programme. See https://www.dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm

[8] The questionnaire also had a 'Don't Know (DK)' category, but this category was merged with 'Yes' for simplicity of analysis.

continuous variable. A standard linear regression model is not suitable for analyzing response variables with categorical scales. For this module, logistic regression is used because our response variable is a categorical variable that has a measurement scale consisting of a set of categories (yes or no) instead of a continuous response variable.

Logistic regression can be used to predict the value of a dependent variable, which is the probability of success (π), ranging between 0 and 1, that a given outcome will occur. The general logistic regression model form is shown (1) below

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta \quad (1)$$

In model (1) above, for a probability of success (π), the odds of success (π) are defined as:

$$odds = \frac{\pi}{1-\pi}$$

Forr example, if π=0.75, then the odds of success equal 0.75/(1-0.75) =3 which suggests that success is three times as likely as failure. Model (1) contains so-called "log odds" on the left-hand side of the equal sign, which are hard to interpret. By taking exponential[9] on both sides of model (1), the odds of success are shown in model (2):

$$\frac{\pi(x)}{1-\pi(x)} = e^{(\alpha+\beta x)} = e^{\alpha}(e^{\beta x}) \quad (2)$$

Model (2) allows for interpretation for coefficients β much more easily. For every 1 unit increase in x, the odds multiply by an amount of e^{β} . For illustration purposes, if the logistic regression line is fitted to the data and the fitted coefficient value β yields 0.497, then the estimated odds of a successful outcome multiply by $e^{0.497}$ = 1.64 for each unit increase in x, which suggests a 64% increase in odds. This exponential relationship is explained further in the next section.

Interpreting logistic regression output

A logistic regression model is fitted to MICS6 data of Mongolia using the "glm" (generalized linear model) function in R and the "logit" function in STATA. As stated before, the variable "attitudes towards domestic violence (yes or no)" is regressed against urban/ rural (UBN), level of education (EDU), age (AGE), marital status (MA) and level of wealth (windex5). Urban is coded 1 and rural is coded 0. Education is coded 2, 3 and 4 where 4 is the highest level of education corresponding to university. Age is the only continuous variable in the model. Marital status is coded 0 (not in union) and 1 (married or living with partners). Wealth quintile (windex5) is coded 1 (poorest) to 5 (richest) in increasing level of wealth. The regression output (Figure 15) shows the coefficients β, their standard errors, the t-statistic and the associated p-values. In general, coefficients β with p-values less than a specified significance level are considered

[9] The odds ratio varies from (0, ∞), which does not match the range (-∞, ∞) of independent variables on the right- hand side of the model (1) shown above. The general logistics regression model takes a logarithm on the odds ratio (left-hand side) to match the range on the right-hand side of the model (1). This log transformation allows fitting a sigmoid function (S-shaped curve) to the data. The logistic regression outputs of the coefficient βs in STATA and R are simply log odds ratios. It is easier to interpret the coefficient β in the 'odds ratio' terms rather than 'log odds ratio'. Using the inverse property of the log function, an odds ratio can be obtained by exponentiating the log odds ratio.

statistically significant. Wealth, marital status and education (highest level) are significant at a 5% significance level. Place of residence (urban/rural) and age are not significant.

Figure 15: Logistic regression for variables potentially influencing attitudes towards domestic violence, R output

```
## Call:
## svyglm(formula = DV ~ UBN + EDU + AGE + MA + windex5, design = wm5design, ##   family =
quasibinomial())
##
## Survey design:
## svydesign(id = ~PSU, strata = ~stratum, weights = ~wmweight, ##      data = wm5)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.778252                0.169953  -4.579 5.78e-06 ***
## UBN1                -0.118419        0.113559  -1.043 0.297506
## EDU3                 0.051980        0.114296         0.455 0.649446
## EDU4                -0.485919        0.122541  -3.965 8.30e-05 ***
## AGE                 -0.003747        0.005122  -0.731 0.464825
## MA1                 -0.299349        0.118083  -2.535 0.011519 *
## windex52            -0.390079        0.114836  -3.397 0.000731 ***
## windex53            -0.379013        0.126776  -2.990 0.002919 **
## windex54            -0.817311        0.157031  -5.205 2.75e-07 ***
## windex55            -1.027721        0.223102  -4.607 5.10e-06 *** ## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 16: Logistic regression for variables potentially influencing attitudes towards domestic violence, STATA output

```
. svy: logit dv i.UBN age i.EDU i.windex i.MA
(running logit on estimation sample)

Survey: Logistic regression
```

Number of strata	= 23	Number of obs	= 10593
Number of PSUs	= 580	Population size	= 10668.125
		Design df	= 557
		F(9, 549)	= 18.33
		Prob > F	= 0.0000

dv	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.UBN	-.1184185	.1135595	-1.04	0.297	-.3414757	.1046387
age	-.0037467	.0051224	-0.73	0.465	-.0138082	.0063148
EDU						
3	.0519797	.1142954	0.45	0.649	-.1725231	.2764824
4	-.4859194	.1225401	-3.97	0.000	-.7266166	-.2452223
windex						
2	-.3900794	.1148355	-3.40	0.001	-.6156429	-.1645158
3	-.3790135	.1267761	-2.99	0.003	-.6280311	-.1299959
4	-.8173107	.1570317	-5.20	0.000	-1.125757	-.508864
5	-1.027721	.2230996	-4.61	0.000	-1.465941	-.5895016
1.MA	-.2993491	.1180811	-2.54	0.012	-.5312878	-.0674104
_cons	-.7782522	.1699508	-4.58	0.000	-1.112075	-.4444294

As suggested in previous correlation analysis and cross-tabulation results, the variables place of residence (urban/rural) and wealth are correlated. In addition, there is a high correlation between age and marital status. Including explanatory variables that contain redundant information could inflate the standard error of coefficients β, which undermines the statistical significance of an explanatory variable in the

model (also known as multicollinearity[10]). To address this problem, the logistic regression is fitted without place of residence (urban/rural) and age. After dropping place of residence (urban/rural) and age, the other variables remain significant in the model. The logistic regression coefficients give the change in the log odds of the outcome for a one-unit increase in the explanatory variable. The following interpretation can be made for the coefficients β:

- Having the highest educational level (EDU4) decreases the log odds of accepting attitudes towards domestic violence (DV), versus the reference group with secondary school education (EDU2), by 0.47641, holding the wealth and marital status variables constant. (or the odds of DV decrease by 1-exp(-0.47641)=37.9%)
- Having marital status of 'married/living with partners' (MA1) decreases the log odds of accepting attitudes towards domestic violence (DV), versus the reference group with marital status of 'not in union' (MA0), by 0.32699, holding the education and wealth variables constant. (or the odds of DV decrease by 1-exp(-0.32699)=27.9%)
- Belonging to the top wealth quintile (windex55) decreases the log odds of accepting attitudes towards domestic violence (DV), versus the reference group with the bottom wealth quintile (windex51), by 1.14565, holding the education and marital status variables constant. (or the odds of DV decrease by 1-exp(-1.14565)=68.2%)

Figure 17: Logistic regression output including only significant variables R output

```
## Call:
## svyglm(formula = DV ~ EDU + MA + windex5, design = wm5design, ##          family =
quasibinomial())
##
## Survey design:
## svydesign(id = ~PSU, strata = ~stratum, weights = ~wmweight, ##     data = wm5)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.88757              0.10460  -8.486  < 2e-16 ***
## EDU3              0.04981         0.11466        0.434   0.66412
## EDU4             -0.47641         0.11709  -4.069 5.42e-05 ***
## MA1              -0.32699         0.10640  -3.073   0.00222 **
## windex52         -0.46848         0.09983  -4.693 3.41e-06 ***
## windex53         -0.47082         0.11393  -4.133 4.15e-05 ***
## windex54         -0.91308         0.14202  -6.429 2.79e-10 ***
## windex55         -1.14565         0.19064  -6.010 3.39e-09 *** ## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[10] 1997. The problem of multicollinearity. In: Understanding Regression Analysis. Springer, Boston, MA. Plenum Press, NY. Available from: https://link.springer.com/chapter/10.1007%2F978-0-585-25657-3_37

Figure 18: Logistic regression output including only significant variables, STATA output

```
. svy: logit dv i.EDU i.windex i.MA
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        23                  Number of obs      =     10593
Number of PSUs     =       580                  Population size    =  10668.125
                                                Design df          =       557
                                                F(   7,    551)    =     23.45
                                                Prob > F           =    0.0000
```

dv	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
EDU						
3	.0498135	.1146556	0.43	0.664	-.1753968	.2750238
4	-.4764057	.1170933	-4.07	0.000	-.7064042	-.2464072
windex						
2	-.4684792	.099833	-4.69	0.000	-.6645744	-.2723839
3	-.4708205	.1139307	-4.13	0.000	-.6946069	-.2470342
4	-.9130773	.1420207	-6.43	0.000	-1.192039	-.6341157
5	-1.145651	.1906384	-6.01	0.000	-1.520109	-.7711928
1.MA	-.3269907	.1063936	-3.07	0.002	-.5359724	-.118009
_cons	-.8875714	.1045972	-8.49	0.000	-1.093025	-.6821181

This concludes Module 1. For exercises, presentations, list of resources/links, and example tests, please refer to the separate files attached to this module. Note that the purpose of this module is to illustrate how to perform survey data analysis using R or STATA. It does not cover all aspects of microdata analysis that researchers are expected to perform. For example, it does not cover data cleaning, verification of assumptions and model diagnostics. Trainers and trainees are expected to cover this background knowledge on their own.

KEY TAKEAWAYS

- Before performing advanced statistical analysis, it is important to clean, merge and identify sampling design (PSU, Strata, sample weights) to be considered for analysis.

- Household surveys are comprised of mostly categorical questions and the use of standard linear regression analysis is inappropriate if the response variable is categorical. For categorical response variables, logistic regression is used.

- Correlation and cross-tabulations should be performed before logistic regression to assess the distribution and strength of relationships between explanatory variables.

- Interpretation of coefficients in the logistic regression output is easier by transforming the log odds into odds by taking the exponential of the coefficients. This value suggests the percentage increase or decrease of odds of an outcome happening, relative to a reference group.

Annex 1: R code used in this module

```
library(foreign) library(plyr)
library(dplyr) library(tidyr)
library(tidyverse) library(survey)
library(car) library(RcmdrMisc)
library(dummies) library(corrplot)

wm=read.spss(file="wm.sav",to.data.frame = TRUE)
hh=read.spss(file="hh.sav",to.data.frame = TRUE)
hl=read.spss(file="hl.sav",to.data.frame = TRUE)
mn=read.spss(file="mn.sav",to.data.frame = TRUE)

# Merging rows in hl2 and hh2 tables that have same HH1 and HH2 but HL2=HL1
hl2=select(hl,HH1,HH2,HL1,HH6,HL6,ED5A)
hl2$HL6=as.numeric(hh1$HL6)

# selecting independent variables from women and hh tables and merging
wm2=select(wm,HH1,HH2,LN,MA1,DV1A,DV1B,DV1C,DV1D,DV1E,windex5,wmweight,PSU,st ratum)
wm2=rename(wm2,HL1=LN) wm1=right_join(hh1,wm2)
wm1=rename(wm1,AGE=HL6,UBN=HH6,EDU=ED5A)
wm1$UBN=revalue(wm1$UBN, c("Urban"="1", "Rural"="0")) wm1$UBN <-
relevel(wm1$UBN, ref = "0")

# create variable DV to merge DK and Yes and making it to only 2 level wm1=mutate(wm1,
MA=ifelse(MA1=="NO, NOT IN UNION","0","1"),DV=ifelse(DV1A=="N O" & DV1B=="NO" & DV1C=="NO"
& DV1D=="NO" & DV1E=="NO","0","1"))

wm1$EDU=mapvalues(wm1$EDU, from = c("ECE","SECONDARY SCHOOL","VOCATIONAL TRAI NING
CENTERS, TECHNICUM","UNIVERSITY, INSTITUTE, COLLEGE"), to = c("2", "2"," 3","4"))

wm1$windex5=mapvalues(wm1$windex5, from = c("0","Poorest","Second","Middle"," Fourth","Richest"),
to = c("1","1","2","3","4","5"))

wm5=drop_na(wm1) wm5$EDU=droplevels(wm5$EDU)
wm5design<-svydesign(id=~PSU,strata=~stratum, weights = ~wmweight, data=wm5)
colPercents(svytable(~DV+EDU, wm5design,round=TRUE)) colPercents(svytable(~DV+windex5,
wm5design,round=TRUE)) colPercents(svytable(~windex5+DV, wm5design,round=TRUE))
colPercents(svytable(~UBN+DV, wm5design,round=TRUE))
```

```r
colPercents(svytable(~MA+DV, wm5design,round=TRUE))
colPercents(svytable(~UBN+windex5, wm5design,round=TRUE))
colPercents(svytable(~windex5+EDU, wm5design,round=TRUE))

wm5.clean= subset(wm5, select = c(4,5,6,13,17))

wm5.clean$EDU<-dummy(wm5.clean$EDU)
wm5.clean$windex5<-dummy(wm5.clean$windex5)
wm5.clean$MA<-dummy(wm5.clean$MA)
wm5.clean$UBN<-dummy(wm5.clean$UBN)
wcor=cor(wm5.clean)

jpeg("wcor.jpg", width = 850, height = 850)
corrplot(wcor, type = "upper", order = "original",
         tl.col = "black", tl.srt = 45)
dev.off()
wm.sglm1<-svyglm(DV~UBN+EDU+AGE+MA+windex5,
design=wm5design,family=quasibino mial())
summary(wm.sglm1)
wm.sglm2<-svyglm(DV~EDU+MA+windex5,
design=wm5design,family=quasibinomial())
summary(wm.sglm2)
```

Annex 2: STATA code used in this module

```
//Reading data
import delimited C:\Users\un153\Downloads\Rresources\hlmics.csv, clear

//Selecting only variables needed from household listing keep hh1
hh2 hl1 hh6 hl6 ed5a
save"C:\Users\un153\Downloads\Rresources\hlfinal.dta", replace

//Merging women data with household listing
import delimited C:\Users\un153\Downloads\Rresources\wmmics.csv, clear rename ln hl1
keep hh1 hh2 hl1 ma1 dv1a dv1b dv1c dv1d dv1e windex5 wmweight psu stratum merge 1:1 hh1
hh2 hl1 using "C:\Users\un153\Downloads\Rresources\hlfinal.dta" keep if _merge==3
save wmfinal.dta, replace

//Generating variables needed for further analysis generate UBN=.
replace UBN=1 if hh6=="Urban" replace
UBN=0 if hh6=="Rural"

generate MA=.
replace MA=0 if ma1=="NO, NOT IN UNION" replace
MA=1 if ma1=="YES, CURRENTLY MARRIED"
replace MA=1 if ma1=="YES, LIVING WITH A PARTNER"

generate dv=1
replace dv=0 if (dv1a=="NO" & dv1b=="NO" & dv1c=="NO" & dv1d=="NO" & dv1e=="N O")


generate EDU=.
replace EDU=2 if ed5a=="ECE"
replace EDU=2 if ed5a=="SECONDARY SCHOOL"
replace EDU=3 if ed5a=="VOCATIONAL TRAINING CENTERS, TECHNICUM"
replace EDU=4 if ed5a=="UNIVERSITY, INSTITUTE, COLLEGE"

generate windex=.
replace windex=1 if windex5=="0" replace
windex=1 if windex5=="Poorest" replace
windex=2 if windex5=="Second" replace
windex=3 if windex5=="Middle" replace
windex=4 if windex5=="Fourth" replace
windex=5 if windex5=="Richest"

generate age= real(hl6)

drop if missing(windex)|missing(EDU)|missing(MA)|missing(UBN)|missing(age)
```

```
//correlation
tab UBN, generate(UBN)
tab EDU, generate(EDU)
rename windex5 wealth
tab windex, generate(windex)
tab MA, generate(MA)
correlate UBN1 UBN2 age EDU1 EDU2 EDU3 windex1 windex2 windex3 windex4 windex
5 MA1 MA2


//Setting sampling and survey design parameters
svyset psu [pw = wmweight], strata(stratum) singleunit(centered)

//Sums the whole values in the column
total(wmweight)

//Cross tabulations
svy: tabulate dv EDU, format(%14.1f) column percent
svy: tabulate dv windex, format(%14.1f) column percent
svy: tabulate dv UBN, format(%14.1f) column percent
svy: tabulate dv MA, format(%14.1f) column percent
svy: tabulate UBN windex, format(%14.1f) column percent
svy: tabulate windex EDU, format(%14.1f) column percent

// Logistic regression
svy: logit dv i.UBN age i.EDU i.windex i.MA
svy: logit dv i.EDU i.windex i.MA
```