Research Methods II

Session 1: Surveys, IO and SAM

Fernando Rios-Avila

Why are we here?

- If you are reading this, you are probably interested in the Microeconometrics path of Research methods II.
- This course assumes you are familiar with the basics of econometrics and statistics.
  - We will not cover questions from Research methods I. You probably know the answers to those questions already!
- This course will focus on tools that are commonly used in empirical research in economics:
  - We will emphasize the use of Household Surveys (with the issues it entails)
  - And focus on applied econometrics using cross-sectional data, for distributional analysis and policy simulation.
- Thus, we will do much more use of empirical tools and software than we did in Research methods I.
- We will have 7 Sessions, and 2 homeworks .

Surveys: What are they?

What is a Survey?

- A Survey is a source of data that aims to collect information from a population of interest, to underestand some characteristics, behaviors, or opinions of that population as a whole.
  - The population of interest can be individuals, households, firms, etc.
- They can be useful to identify and analyze policy questions.

- However, they are secondary data, and thus have limitations in terms of the questions that can be answered with them.
  - You cannot answer questions that require data that was not collected.
  - They can also be limited to Interviewee recall, or willingness to answer.
  - Or how accessible the population of interest is.

Example of Surveys

- Current Population Survey (CPS)
  - Monthly survey of 60,000 households in the US.
- American Community Survey (ACS)
  - Annual survey of 3.5 million households in the US.
- American Time Use Survey (ATUS)
  - Annual survey of 6,000 individuals in the US.
- Enterprise Surveys - WB (ES)
  - Survey of firms in developing countries. Different years of Collection

What makes a good survey?

- A Good survey is one that allows you to obtain estimates of statistics of interest for the population with "Tolerable" levels of Accuracy.
- To do this, you need to have a good sampling design (representation and "independence of the population") and a good questionnaire design (questions that are clear and easy to answer).
- A good survey needs to be representative of the population of interest. To do this appropriately, data will be collected based on a **frame** that will be used to select the sample.

Types of Data Seletion

Simple

- Each observation in the "frame" has the same probability of being selected in the sample.
- It may be difficult to implement in practice, because of cost and logistics. (distance)
- It could also have problems of representativeness for small groups. (rare events)

Clustered

- Using some criteria, location for example, the population is divided into clusters.
- For the sample selection, certain clusters are selected at random, and “some” observations within each selected.
- This is more feasible in practice, because takes advantage of the “clustering” of the population.
- However, one may need to account for possible “common shocks” that people within the same cluster may have.

Stratified

- Some times, statistics are required to accurately represent certain groups of the population. (by region, race, income level, etc)
- In such cases, data can be collected in a way that ensures that the sample has enough observations for each group of interest.
    - It would be as collecting multiple samples, one for each group of interest.
- Within each Strata, it would also be possible to use a simple or clustered sampling design.

What to be aware of?

- The sampling design is important to ensure that the sample is representative of the population of interest.
- However, there are limitations:
    - Not every-one selected will respond to the survey. (Is it random?)
    - Rarely one assumes equal probability of selection. (Different sampling weights)
    - Use of Stratatification and Clustering may require special treatment.
- Something else: Panel data
    - Because of attrition, it can be difficult to analyze if representativeity is required.
    - However, it can be useful to analyze dynamics.

Descriptive Statistics

- Once you have your data, you can start analyzing it by simply applying your survey weights.
    - Point estimates are straightforward to obtain.
- However, when considering the estimation of precision of the estimates (Variance and Standard Errors), there are two approaches that are important to consider:

- **Finite Population Approach**:
  - Associated with data description
  - Assumes that the population is finite, and thus Selection probabilities are not independent.
- **Superpopulation Approach**:
  - Associated with data modeling.
  - Population is infinite. Selection probabilties are independent.
- For practical purposes, the difference between the two approaches is not that important.
  - With large enough samples, he "finite sample correction" is negligible.

Basic Summary Statistics:

- n is sample size. N is population size.
- a_i an indicator of belonging to the sample. $\sum a_i = n$
- Assume y_i is the outcome of interest, Say income.

$$\texttt{mean: } \hat{\mu}_y = \bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i y_i = \frac{1}{n}\sum_{j=1}^{n} y_j$$

$$\texttt{Variance: } \hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{N} a_i (y_i - \hat{\bar{y}})^2$$

$$\texttt{V of mean: } \widehat{V}(\bar{y}) = \frac{\sigma_y^2}{n} * fpc$$

Where $fpc = \frac{N-n}{N-1}$ is the finite population correction.

Accounting for Weights

- The previous formulas did not account for weights.
- Weights are factors used to "reweight/expand" the sample to make it representative of the population.
- They can are typically related to the inverse of the probability of selection.
- Simply said, a weight $w_i = \frac{1}{n * \pi_i}$ is a measure of how many observations in the population are represented by observation i in the sample.

But how do we account for weights in the formulas?

Summary Statistics with Weights

Population:

$$\hat{N} = \sum_{i=1}^{n} w_i$$

Normalized weights

$$v_i = \frac{nw_i}{\sum w_i} \rightarrow E(v_i) = 1$$

Mean:

$$\hat{\mu}_y = \bar{y} = \frac{1}{\hat{N}} \sum_{i=1}^{n} w_i y_i = \frac{1}{n} \sum v_i y_i$$

Variance with Weights

Variances are a bit more complicated. Normally you would consider:

$$Var(\bar{y}) = \frac{1}{n}\hat{\sigma}_y^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n} v_i (y_i - \bar{y})^2$$

However, with survey weights you need to consider something else:

$$Var(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} v_i^2 (y_i - \bar{y})^2$$

Which is similar to “robust” Standard errors in OLS.

What about Clusters and Strata?

How to account for weights in Stata?

Lets use data two Examples.

- First Labor Force Survey: Oaxaca
- Second National Health and Nutrition Examination Survey (NHANES)

```
 frause oaxaca, clear
** Create Weights <- This will be provided
sum wt, meanonly
replace wt = round(wt/r(min))
gen wage = exp(lnwage)
tab wt
```

```
<IPython.core.display.HTML object>
```

```
(Excerpt from the Swiss Labor Market Survey 1998)
(1,647 real changes made)
(213 missing values generated)

   sampling |
    weights |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        489       29.69       29.69
          2 |        924       56.10       85.79
          3 |        160        9.71       95.51
          4 |         64        3.89       99.39
          5 |          8        0.49       99.88
          6 |          2        0.12      100.00
------------+-----------------------------------
      Total |      1,647      100.00
```

Weights Distribution

```
tab wt
```

```
   sampling |
    weights |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        489       29.69       29.69
          2 |        924       56.10       85.79
          3 |        160        9.71       95.51
          4 |         64        3.89       99.39
          5 |          8        0.49       99.88
          6 |          2        0.12      100.00
------------+-----------------------------------
      Total |      1,647      100.00
```

Summary Statistics:

```
** unweighted
sum wage,d
** weighted
sum wage [aw=wt],d
```

```
                            wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.907204       1.661434
 5%      11.8007       1.873127
10%      17.4216       2.197802       Obs               1,434
25%     23.19902       2.442003       Sum of wgt.       1,434

50%     30.08896                      Mean           32.39167
                        Largest       Std. dev.      16.12498
75%      38.5662       137.3627
90%     49.95005       152.6252       Variance       260.0151
95%     58.71271       164.8352       Skewness       2.486954
99%     85.47008       192.3077       Kurtosis       18.23702

                            wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     3.453689       1.661434
 5%     7.370678       1.873127
10%     15.83933       2.197802       Obs               1,434
25%     21.72247       2.442003       Sum of wgt.       2,686

50%     29.48271                      Mean            31.5322
                        Largest       Std. dev.      16.32811
75%     38.46154       137.3627
90%     49.95005       152.6252       Variance       266.6071
95%     58.11497       164.8352       Skewness       2.081806
99%     85.47008       192.3077       Kurtosis       14.68647
```

Accounting for weights for summary Statistics

```
** unweighted
mean wage
** weighted
mean wage [pw=wt]
mean wage [pw=wt], over(female)
```

```
Mean estimation                          Number of obs = 1,434

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
        wage |   32.39167   .4258186      31.55637    33.22696
--------------------------------------------------------------

Mean estimation                          Number of obs = 1,434

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
        wage |    31.5322   .4765835      30.59733    32.46708
--------------------------------------------------------------

Mean estimation                           Number of obs = 1,434

---------------------------------------------------------------
              |       Mean   Std. err.     [95% conf. interval]
--------------+------------------------------------------------
c.wage@female |
            0 |   33.87649   .6152059      32.66969    35.08329
            1 |   28.76133    .722173       27.3447    30.17796
---------------------------------------------------------------
```

Tables and cross tables

```
** unweighted
tab educ female

tab educ female [w=wt]
```

```
           |   sex of respondent
  years of |      (1=female)
 education |         0          1 |     Total
-----------+----------------------+----------
         5 |         6         23 |        29
         9 |        47         93 |       140
      9.75 |         8         26 |        34
        10 |        10         42 |        52
      10.5 |       376        447 |       823
      11.5 |         7         14 |        21
        12 |       111         87 |       198
      12.5 |        56         80 |       136
        15 |        60         13 |        73
      17.5 |        78         63 |       141
-----------+----------------------+----------
     Total |       759        888 |     1,647

           |   sex of respondent
  years of |      (1=female)
 education |         0          1 |     Total
-----------+----------------------+----------
         5 |        17         45 |        62
         9 |       104        181 |       285
      9.75 |        14         52 |        66
        10 |        22         80 |       102
      10.5 |       725        833 |     1,558
      11.5 |        19         27 |        46
        12 |       201        148 |       349
      12.5 |       108        157 |       265
        15 |       111         21 |       132
      17.5 |       149        111 |       260
-----------+----------------------+----------
     Total |     1,470      1,655 |     3,125
```

Better approach: `svyset`

```
webuse nhanes2f, clear
svyset psuid /// Cluster
       [pweight=finalwgt], /// Survey Weight as Inverse of Prob of Selection
       strata(stratid)     // Strata Identifier
mean zinc
mean zinc [pw=finalwgt]
svy: mean zinc
```

```
Sampling weights: finalwgt
             VCE: linearized
     Single unit: missing
        Strata 1: stratid
 Sampling unit 1: psuid
           FPC 1: <zero>

Mean estimation                          Number of obs = 9,189

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
        zinc |   86.51518   .1510744      86.21904    86.81132
--------------------------------------------------------------

Mean estimation                          Number of obs = 9,189

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
        zinc |   87.18207   .1828747      86.82359    87.54054
--------------------------------------------------------------
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 31            Number of obs   =       9,189
Number of PSUs   = 62            Population size = 104,176,071
                                 Design df       =          31

--------------------------------------------------------------
             |             Linearized
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
        zinc |   87.18207   .4944827      86.17356    88.19057
--------------------------------------------------------------
```

Testing Significance across 2 groups

Consider two groups with the following characteristics:

Group	N	mean	Var
1	100	45	32.56
2	150	55	21.97

- Is the difference in means statistically significant?
- Test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$

$$t = \frac{\mu_2 - \mu_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{10}{\sqrt{\frac{32.56}{100} + \frac{21.97}{150}}} = \frac{10}{\sqrt{.47207}} = 14.55453$$

- If t is large enough, we can reject the null hypothesis.
- But this does not work if you have weights...

Testing Significance across 2 groups

- But you can use OLS to test the difference in means with weights!
  - Make sure you use “robust” standard errors, or “pw” option.

```
 frause oaxaca, clear
** Create Weights <- This will be provided
sum wt, meanonly
replace wt = round(wt/r(min))
gen wage = exp(lnwage)
reg wage i.female [pw=wt],
```

```
(Excerpt from the Swiss Labor Market Survey 1998)
(1,647 real changes made)
(213 missing values generated)
(sum of wgt is 2,686)

Linear regression                               Number of obs     =      1,434
                                                F(1, 1432)        =      29.05
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0244
                                                Root MSE          =     16.133

------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.female |  -5.115156   .9490209    -5.39   0.000    -6.976776   -3.253536
       _cons |   33.87649   .6154206    55.05   0.000     32.66927    35.08371
------------------------------------------------------------------------------
```

- you can also use `svy: regress` for complex designs.

see here for additional examples

IO-Tables

IO-Tables

- IO tables stand for Input-Output tables. They are a way to represent the production structure of an economy.
- They provide a Static representation of the Economy
- Each Row represents the production of a sector, and each column represents the use of that production by other sectors as Inputs.
- The information it contains represent a snapshot of the economy at a given point in time.
- It can be used to simulate changes in production, labor demand, and total production in the economy, under specific assumptions (production function)

- Consider a Economy with K=3 sectors: Agriculture, Manufacturing and Services, with a Final consumer agent (households)
- Each sector (i) produces a good (X_i), which is sold to other sectors or the final consumer.
- X_{ij} is the quantity of goods sector i sells to sector j, and y_i the final consumption by households.
- X_{ji} is also the quantity of goods sector i uses from sector j as inputs.
- L_i is the amount of labor used by sector i.

- In a simple Economy, (value) Total labor demand is equal to total Household income. $L_a + L_m + L_s = L = y_a + y_m + y_s$

This simple Economy can be represented as follows:

$$\begin{aligned} X_{11} + X_{12} + X_{13} + Y_1 &= X_1 \\ X_{21} + X_{22} + X_{23} + Y_2 &= X_2 \\ X_{31} + X_{32} + X_{33} + Y_3 &= X_3 \\ L_1 + L_2 + L_3 + 0 &= L \end{aligned}$$

	S1	S2	S3	HH	
Supply:					
S1	X_{11}	X_{12}	X_{13}	Y_1	X_1
S2	X_{21}	X_{22}	X_{23}	Y_2	X_2
S3	X_{31}	X_{32}	X_{33}	Y_3	X_3
HH	L_1	L_2	L_3	0	L

From the consumption/inputs Side, we could also write the equations as:

$$\begin{aligned} X_{11} + X_{21} \quad &+ X_{31} + L_1 &= X_1 \\ X_{12} + X_{22} \quad &+ X_{32} + L_2 &= X_2 \\ X_{13} + X_{23} \quad &+ X_{33} + L_3 &= X_3 \\ Y_1 + Y_2 \quad &+ Y_3 &= Y \end{aligned}$$

And from here we can get the technical coefficients:

$$\begin{aligned} a_{11}X_1 + a_{21}X_1 \quad &+ a_{31}X_1 + \lambda_1 X_1 &= X_1 \\ a_{12}X_2 + a_{22}X_2 \quad &+ a_{32}X_2 + \lambda_2 X_2 &= X_2 \\ a_{13}X_3 + a_{23}X_3 \quad &+ a_{33}X_3 + \lambda_3 X_3 &= X_3 \\ \delta_1 Y + \delta_2 Y \quad &+ \delta_3 Y &= Y \end{aligned}$$

Where $a_{ij} = \frac{X_{ij}}{X_j}$ and $\lambda_i = \frac{L_i}{X_i}$

$$a_{1i} + a_{2i} + a_{3i} + \lambda_i = 1 \ \& \ \delta_1 + \delta_2 + \delta_3 = 1$$

With this, we can write the IO table

$$\begin{aligned}
a_{11}X_1 + a_{12}X_2 \quad & +a_{13}X_3 + \delta_1 Y \quad && = X_1 \\
a_{21}X_1 + a_{22}X_2 \quad & +a_{23}X_3 + \delta_2 Y \quad && = X_2 \\
a_{31}X_1 + a_{32}X_2 \quad & +a_{33}X_3 + \delta_3 Y \quad && = X_3 \\
\lambda_1 X_1 + \lambda_2 X_2 \quad & +\lambda_3 X_3 && = L
\end{aligned}$$

Or into Matrix Form

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ \lambda_1 & \lambda_2 & \lambda_3 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ 0 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ L \end{pmatrix}$$

Solve for Production sectors:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = I \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} - A \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = (I - A) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Finally:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = (I - A)^{-1} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \rightarrow \begin{pmatrix} \Delta X_1 \\ \Delta X_2 \\ \Delta X_3 \end{pmatrix} = (I - A)^{-1} \begin{pmatrix} \Delta Y_1 \\ \Delta Y_2 \\ \Delta Y_3 \end{pmatrix}$$

Example

- Consider the following IO table for a simple economy with 3 sectors and a final consumer.

Sector	Agriculture	Manufacturing	Services	Final Consumer
Agriculture	102	103	153	129
Manufacturing	133	124	77	99
Services	71	92	51	165
Households	181	114	98	0

S1: What is the total production of each sector?

- $X_1 = 102 + 103 + 153 + 129 = 487$
- $X_2 = 133 + 124 + 77 + 99 = 433$
- $X_3 = 71 + 92 + 51 + 165 = 379$

S2: What are the technical coefficients?

Sector	Agriculture	Manufacturing	Services	Final Consumer
Agriculture	$a_{11} = 0.209$	0.238	0.404	0.328
Manufacturing	0.273	0.286	0.203	0.252
Services	0.146	0.212	0.135	0.420
Households	0.372	0.263	0.259	0.000

This captures a snapshot of an economy. And could use to simulate changes in production and labor demand.

S3. How much would production change if the final consumer demand for Agriculture increases in 20%?

$$\Delta X = (I - A)^{-}1 \begin{pmatrix} 0.2 * 129 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 45.54 \\ 21.08 \\ 12.84 \end{pmatrix}$$

S4: How much would labor demand change if the final consumer demand for Agriculture increases in 20%?

$$\Delta L_i = \lambda_i * \Delta X_i$$

$\Delta L_1 = 0.372 * 45.54 = 16.95; \Delta L_2 = 0.263 * 21.08 = 5.544; \Delta L_3 = 0.259 * 12.84 = 3.326$

Example - `Stata`

`mata`: Input Data

```
mata: x  = (102, 103, 153 \ 133, 124, 77 \ 71, 92, 51)
mata: y  = (129 \ 99 \ 165)
mata: hh = ( 181 , 114 , 98)
```

Estimate Total Production

```
mata: tp = colsum(x):+hh ; tp
```

```
         1     2     3
    +-------------------+
  1 |  487   433   379  |
    +-------------------+
```

Estimate Technical Coefficients

```
// Technical Coefficients
mata:ai = x:/tp ; ai
```

```
                 1             2             3
    +-------------------------------------------+
  1 |  .2094455852   .2378752887   .4036939314  |
  2 |   .273100616   .2863741339   .2031662269  |
  3 |  .1457905544   .2124711316   .1345646438  |
    +-------------------------------------------+
```

Estimate Change in Demand: 20% increase in Agriculture

```
// Technical Coefficients
mata:dy = y :* (.2 \ 0 \ 0); dy
```

```
          1
    +--------+
  1 |  25.8  |
  2 |     0  |
  3 |     0  |
    +--------+
```

Estimate Change in Production

```
// Change in Production
mata:dx = qrinv(I(3)-ai)*dy; dx
```

```
                 1
    +----------------+
  1 |  45.54130481   |
  2 |  21.08637814   |
  3 |  12.84872246   |
    +----------------+
```

Estimate Change in Labor Demand

```
mata:dl = (hh:/tp)':*dx; dl
```

```
                 1
    +----------------+
  1 |   16.9260291   |
  2 |  5.551609949   |
  3 |  3.322360954   |
    +----------------+
```

SAM: Social Accounting Matrix

SAM: Social Accounting Matrix

- SAM can be thought as an upgraded version of IO-tables.
- They are a way to organize information about the production structure of an economy, but also the distribution of resources.
  - This it will not only register production of goods and services, but also transfers of resources between sectors and agents.
- You can also use it as basis for a plausible model of the economy.
  - Prediction of changes in production, income and distribution.

Example

Table 5: Closed Economy No GoV SAM

	Production	Consumption	Accumulation	Totals
Production		C	I	$C+I$
Consumption	Y			Y
Accumulation		S		S
Totals	Y	$C+S$	I	

- Goods/services are Transfered from left to Top-right
- Monetary Transfers are from Top-right to left

Example 2

Table 6: Open Economy with GoV SAM

	S1	S2	S3	S4	S5	Totals
S1: Prod		C	G	I	E	$C+G+I+E$
S2: HH	Y					Y
S3: Gov		Tx				Tx
S4: K acc		S_h	S_g		S_f	$S_h+S_g+S_f$
S5: RofW	M					M
Totals	$Y+M$	$C+S_h+Tx$	$G+S_g$	I	$E+S_f$	

Here $S1$ and $S2$ is what we had in the IO table. Thus, we could further expand the SAM to include more sectors and agents.

Example 3

Table 7: MoreDetailed SAM

	S1	S2	S3	S4	S5	S6	S7	S8
S1: Act		Gds						
S2: Commod	$IntGds$				C	G	I	E
S3: Factors	VA							FE
S4: Enter			$Prof$			ETr		

	S1	S2	S3	S4	S5	S6	S7	S8
S5: HH			$Wage$	$DProf$		Tr		REM
S6: Gov	ITx		FTx	ETx	DTx			$Tarf$
S7: K acc				$RetY$	S_h	S_g		$KTrM$
S8: RofW		M	FM		$REMA$	TrA	$KtrA$	

Other Extensions

- SAM also allow you to do further extensions to include more agents (heterogenous)
    - Green-Industry
    - Informal Sector
    - Households by income level
    - etc.

Thats all folks!

What to get from today?

- How to use weights in Stata to account for survey design, and how to obtain summary statistics.
- How to test differences in means across groups.
- How to use IO tables to simulate changes in production and labor demand.
- Understand how SAM can be used to represent an Economy