Staff Paper

INTRODUCTION TO STATISTICS FOR AGRICULTURAL ECONOMICS FOR USING SPSS

Scott M. Swinton and Ricardo Labarta

Staff Paper 2003-13E September, 2003

Department of Agricultural Economics
MICHIGAN STATE UNIVERSITY
East Lansing, Michigan 48824

Introduction to Statistics for Agricultural Economists Using SPSS

Scott M. Swinton and Ricardo Labarta
swintons@msu.edu and labarta@msu.edu

Abstract

This document is a primer in statistics for applied economists using the SPSS statistical software. It is intended for use with a one-week training workshop designed to acquaint research professionals with basic statistical procedures for analyzing socio-economic survey data. The document introduces users to database creation and manipulation, exploratory univariate and bivariate statistics, hypothesis testing, and linear and logit regression. The text is supported with 19 text boxes that illustrate how procedures can be applied to a farm survey dataset.

34 pages

INTRODUCTION TO STATISTICS FOR AGRICULTURAL ECONOMISTS USING SPSS[1]

Scott M. Swinton and Ricardo A. Labarta[2]

Michigan State University
Department of Agricultural Economics

Staff Paper No. 03-13E

September 2003

Abstract

This document is a primer in statistics for applied economists using the SPSS statistical software. It is intended for use with a one-week training workshop designed to acquaint research professionals with basic statistical procedures for analyzing socio-economic survey data. The document introduces users to database creation and manipulation, exploratory univariate and bivariate statistics, hypothesis testing, and linear and logit regression. The text is supported with 19 text boxes that illustrate how procedures can be applied to a farm survey dataset.

[1] Based on a training workshop organized by Instituto Nicaraguense de Tecnología Agropecuaria and the Bean-Cowpea CRSP Project, Montelimar, Nicaragua. January 20-24, 2003. The authors thank Lesbia Rizo for permitting the use of an INTA database in the examples.

[2] Scott M. Swinton (swintons@msu.edu) is professor and Ricardo Labarta (labartar@msu.edu) is graduate research assistant in the Department of Agricultural Economics at Michigan State University, East Lansing, MI 48824-1039.

First day

Introduction to statistical analysis

Learning to use SPSS can be compared with learning to use a new kitchen appliance. This appliance would be worthless if it is not used to prepare food. But food preparation requires more than one appliance. It requires inputs (such as data to analyze), other tools (like spreadsheets) and especially, the knowledge of how to prepare food (in the analysis the research methods and the use of statistics).

The goal of this document is not simply to educate in the use of software like SPSS, but also to communicate ideas on how to plan a good analysis. In order to ensure full group participation during the workshop we will alternate among four areas of focus:

a) General information about SPSS

b) Principles of research design and statistical analysis

c) Applications to suitable data sets

d) Group research projects by participants

1. SPSS general presentation

Advantages of using SPSS. SPSS is a user friendly software, especially for managing and analyzing large databases. It can directly read files with spreadsheet and database formats such as DBF, WK1, and XLS. Another advantage over similar software is its diversity of output presentation formats through tables and graphs.

Databases in SPSS. A database consists of a general structure built over a base of variables and observations. It can be thought of as a matrix like a spreadsheet where each row contains one observation and each column contains data of a particular variable (across many observations) (Wolf 1990). When dealing with multi-level data it is preferable to build a database with a different observation for each unit of analysis (e.g., per household).

Files generated in SPSS. SPSS generates files with special extensions that differ according to the type of information each of them contain. The most common extensions are SAV for data files, SPS for command files, and SPO for output files

2. Database Creation in SPSS

2.1 Direct data entry. You can enter survey information directly into SPSS to generate a database with a SAV extension. First of all, you have to define variables that will be included in the database, providing each of them a name no longer than 8 characters. Adding descriptive labels is also recommended for describing variable characteristics.

2.2 Database import into SPSS. You can convert databases generated in different software into SAV files. Common formats like Excel, Access and FoxPro can be converted into a SPSS database, by opening the file while using SPSS. The file can then be saved as a SAV file.

2.3 **Using transformed databases**. Before analysis, a transformed database should be checked to determine whether the structure and data can meet analytical objectives. Five steps are helpful in the verification process:

a) To determine the characteristics of the current database: number of variables, variable types, number of observations and the level of each observation (Wolf 1990),

b) To know whether the database includes all the relevant information in one file or in more than one file,

c) To verify whether all the information referring to a unit of analysis (e.g., farmer, household) is included in a single record,

d) To verify whether there exist redundant variables or whether all the variables have unique information,

e) To verify the appropriateness of the variable names and labels.

Example of non-SPSS files import and review. Nicaragua1

Import the Excel databases SISTEMAS.XLS and GENERAL.XLS. This example will also let you diagnose the structure type of the imported file

Analyzing the transformed database:

a) This database has information at farm level, crop level and plot level. There are more than 7000 observations for each variable.

b) There exists important information for the same farmer in different files (SISTEMAS.XLS and GENERAL.XLS)

c) There exists more than one record per farmer. For example, there is information on different crops from the same farmer in separate records. Also there are plots of the same crop and the same farmer in separate records. However, for analyzing farmer behavior, it is generally more convenient to have only one record per farmer case.

d) There are redundant variables with the same information (for example: variable codes and description). This feature causes duplication, increases file size and is impractical.

e) Variable names are long, which causes problems for older programs, and is not practical for running SPSS. It is better to assign shorter, clearer variable names with labels to provide fuller information.

3. Farm level database creation

3.1. **Use of pre-existing information in more than one database**

A database structure should be as flexible as possible. Depending on the objectives and the type of analysis planned, it may be easier to manage and analyze data in more than one database. In such a case, all databases should include common index variable that allows linking the existing information in more than one database.

When the information is entered directly, database structure is defined from the beginning. If the information was already entered in another software (Excel, Dbase, etc.), it is necessary to analyze whether to keep the current structure or create a new, more appropriate structure. If keeping the current structure, then the file may be directly imported into SPSS, as was previously explained. However, changing the file structure requires some manipulation within SPSS, depending on the modifications that are planned for the database. These modifications can include merging two or more files with complementary information, reducing the number of variables, reducing the number of observations, and others. Merging files is perhaps the procedure that requires the most work. The steps for merging two files in SPSS are described below:

a) Sort in ascending order the database you will merge, using the common index variable as the factor for this procedure. Commands required are: ***Data, Sort case, Sort by* COMMON VARIABLE*, Sort order, Ok.***

b) Proceed to merge files. Keep one of the files active and use the merge command with the common variable as the factor for merging: ***Data, Merge files, Add variables, File name, Key variable*** (PRODUCTOR), ***Ok***

c) Check for possible problems generated because of the original structure. Often merged databases fail to retain the same number of observations for the same common variables.

d) Adjust the new database as needed. If merged databases generate a new database with missing values due to differences in the original structure, it may be necessary to adjust the new database manually.

Example for merging files: Nicaragua 2
Merging files SISTEMAS.SAV and GENERAL.SAV

a) Sort SISTEMAS.SAV and GENERAL.SAV according to the variable PRODUCTOR: For running SPSS proceed with: ***Data, Sort files, Sort by*** (PRODUCTOR), ***Sort order ascending***, ***Ok***
b) Take SISTEMAS.SAV as the base file and merge the GENERAL.SAV file. The common variable between both files is PRODUCTOR: ***Data, Merge files, Add variables, file number*** (GENERAL.SAV), ***Common variable*** (PRODUCTOR), ***Ok***
c) Check for problems caused while merging the original databases. A useful suggestion is to discuss the history of the survey and data entry, in order to understand the original database organization.
d) One problem originates from the existence of many records for the same farmer in one of the databases. SPSS will assign the new information only to the first record for each farmer. This information must be copied manually into other records by using ***Copy*** and ***Paste*** Commands.

As an example, manually correct the variable SEXO. Copy the "value" of SEXO (F or M) that appears only in the first record of each farmer, into the remaining records of the same farmer. With this procedure, the variable SEXO will have a value for each of the records in the file.

Proceed similarly with other variables like MODATP, MUNICIPIO, REGION, EPOCA, NOMTIENE, ANOINGRESO and all the variables that have missing values after merging files. The final product will be the creation of a database that you can name BASE.SAV that will have the field-level structure of SISTEMAS.SAV linked to the farm household data in GENERAL.SAV

3.2 Recoding and generating new variables

Recoding lets variables to be redefined. For example, many statistical procedures work better with numeric rather than alphanumeric or chain variables. All recoded variables or variable types are generated from existing variables.

Recoding is an easy procedure that is widely used in SPSS. Like any SPSS variable, recoded variables should be given short and indicative names. Labels with descriptive information should also be added. The recoded variable generated can retain the original name or receive a new name. In the case of a binary variable, it is customary to link the name to the presence of the attribute, so that Yes=1 and No=0.

The required commands for recoding a variable are: ***Transform, Recode, Into different variable, Select variable, Name the output variable, Change, Old and new variables, Assign values*** (which new values correspond to the old ones?)

Examples for recoding variables. Nicaragua 3
Recode the variable SEXO (creation of a binary variable)

To generate a binary variable for the alphabetic variable SEXO, link this new variable with the gender of the household's head. A variable for "female household head" (JEFEFEM) can be created by assigning the value "1" if "yes" and the value "0" if not.

Because SEXO is alphabetic, when generating the variable JEFEFEM, the old values "F" and "M" should be replaced by "1" and "0" respectively.

3.3 Creation of categorical variables

Categorical variables are discrete variables (non-continuous) that state the category to which records belong (e.g., seed variety, region). These variables are useful for doing cross tabulation or descriptive statistics by category, as will be seen later. The generation of binary variables from categorical variables is easy to do and sometimes facilitates analysis.

Creating categorical variables uses the following commands: ***Transform, Automatic recode, Variable to change, Name variable to be transformed, Ok***

Example for generating a categorical variable: Nicaragua 4
Transformation of the alphabetical variable TENENCIA into the new categorical variable NTENEN
This procedure can produce the following values:

1 = Alquilada (Rented)
2 = Mediaria (Sharecropped)
3 = Prestada (Borrowed)
4 = Propia (Owned)
5 = R.A

The same procedure can be applied to the variables TPROPIA, POSTRERA, APANTE, ATP1, ATPMA, SEMMEJ, and OCCSA.

3.4 Database review and sub-base creation

After generating and recoding necessary variables, the new database should be reviewed in order to decide whether it is complete and whether all the information is needed. Often a database has more information than required for specific analysis, which slows the process. For example, when analyzing information about a specific crop, data about other crops is superfluous. In this case there are two options:

a) To keep the complete database and constrain the information that will be used in each analysis. This procedure will restrict the analysis to specified observations. The commands required are: ***Data, Select case, If condition (define a variable and the break point value of the restriction)***, ***Continue, Ok***.

b) To divide the sample in sub-databases according to the type of information that will be used. For example a sub-base containing information only about fields with a particular crop can be created to reduce the size of the database. This procedure requires sorting the database using key variables as factors (i.e. crop). Then proceed to eliminate any other observation that does not belong to the determined crop. Finally you can save the new file with a new name.

Second day

3.5. Cleaning data

Cleaning data improves its quality. Does the database contain observations with unexpected values? Such observations are known as outliers. SPSS provides several procedures to detect the existence of outliers and to correct them. As will be discussed in the next section, these procedures depend on the type of variable to be analyzed. Outlier values are not necessarily wrong. If they are true, they can be very informative. But if these values are erroneous, they should be corrected or deleted (see the ***explore*** command in page 14).

4. Introduction to descriptive statistics

4.1 Theoretical foundation

The key purpose of descriptive statistics is to draw inferences about a population by observing sample members of the population. The most informative descriptive statistics are measures of central tendency and dispersion in the data.

4.1.1 Measures of central tendency

The mean is the main measure of central tendency. By definition, it depends on the probability of each case. The formula for a population mean is:

$$\mu_x = \sum_i x_i p(x_i)$$

Here x_i is the value of each observation and $p(x_i)$ is the associated probability. For sampled data, we assume that the probability of occurrence is the same for each observation, so the formula becomes

$$\bar{x} = \sum_i x_i / n$$

Other measures of central tendency are the median (the value of the halfway point in a sorted dataset) and the mode (the most frequently occurring value).

4.1.2 Measures of dispersion

The most common measures of dispersion are the variance, the standard deviation, and the coefficient of variation (CV). For a specific population with population mean μ_x and probability of occurrence p(x_i) for each observation, the population variance is defined as:

$$\sigma_x^2 = \sum_i (x_i - \mu_x)^2 \, p(x_i)$$

The sample variance is defined as:

$$s_x^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1}$$

The sample standard deviation offers a measure of dispersion in same units as the mean:

$$s_{\overline{x}} = \sqrt{s_{\overline{x}}^2}$$

Finally the coefficient of variation is calculated as the ratio of standard deviation to the mean:

$$cv_x = \frac{s_{\overline{x}}}{\overline{x}}$$

A coefficient of variation greater than 0.5 implies that the mean is not different from zero with 95% confidence if the data follow a normal probability distribution (see section 4.3.3 below)

4.2 Descriptive statistics in SPSS

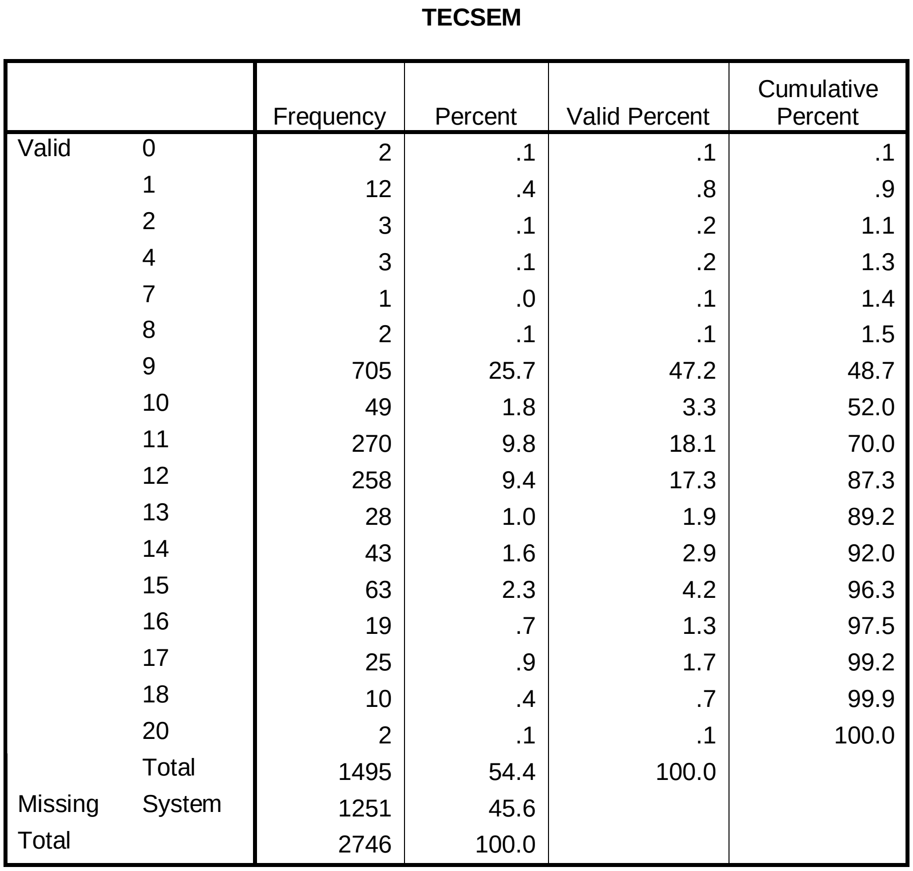
The objective of this section is to introduce ways to explore data using both graphical and statistical methods. Exploratory data analysis is an important first step before moving on to more formal methods.

4.2.1. Graphs. SPSS provides a large number of options for doing graphical analysis of data. The most widely used options are: ***Histograms, Error bars*** and ***Scatter plots***. All of these options offer visual data displays. The commands are located under the ***Graph*** menu.

4.2.2 Statistical diagnosis. These procedures offer numerical analysis. They are grouped under the commands ***Analyze, Descriptive statistics***. Descriptive statistics offer a large range of univariate and bivariate analysis for both categorical and continuous variables. The four procedures below are especially useful.

4.2.2.1. **Frequencies**. This procedure displays the frequency distribution of categorical variables. To run this procedure, follow the commands: ***Analyze, Descriptive statistics, Frequencies, Variable, Ok***.

Example of a frequency distribution for a categorical variable. Nicaragua 5
Example: Frequencies for the seed technology variable (TECSEM)



TECSEM

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	.1	.1	.1
	1	12	.4	.8	.9
	2	3	.1	.2	1.1
	4	3	.1	.2	1.3
	7	1	.0	.1	1.4
	8	2	.1	.1	1.5
	9	705	25.7	47.2	48.7
	10	49	1.8	3.3	52.0
	11	270	9.8	18.1	70.0
	12	258	9.4	17.3	87.3
	13	28	1.0	1.9	89.2
	14	43	1.6	2.9	92.0
	15	63	2.3	4.2	96.3
	16	19	.7	1.3	97.5
	17	25	.9	1.7	99.2
	18	10	.4	.7	99.9
	20	2	.1	.1	100.0
	Total	1495	54.4	100.0	
Missing	System	1251	45.6		
Total		2746	100.0		

Note the large number of cases with missing values (these are for technologies that do not involve seeds.

4.2.2.2. **Descriptives**. This procedure presents the main statistical measures of continuous variables: mean, standard deviation, minimum value and maximum value. The commands needed are: ***Analyze, Descriptive statistics, Descriptives, Variables, Ok.***

Example of descriptive statistics of continuous variables: Nicaragua 6
Example: Descriptive statistics for bean yields (RENDI) and farm size (AREA).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RENDI	2746	0	75	11.40	7.217
AREA	2746	0	23	1.55	1.533
Valid N (listwise)	2746				

4.2.2.3 **Cross tabulation (contingency tables)**. This command generates the joint frequency distribution of two categorical variables. Use the commands: ***Analyze, Descriptive statistics, Crosstabs, Row variable, Column variable, Ok.***

Example of cross tabulation: Nicaragua 7
Example: Cross tabulate regions by technical assistance type (ATP). Use variables REGION and NOMATP.

REGION * NOMATP Crosstabulation

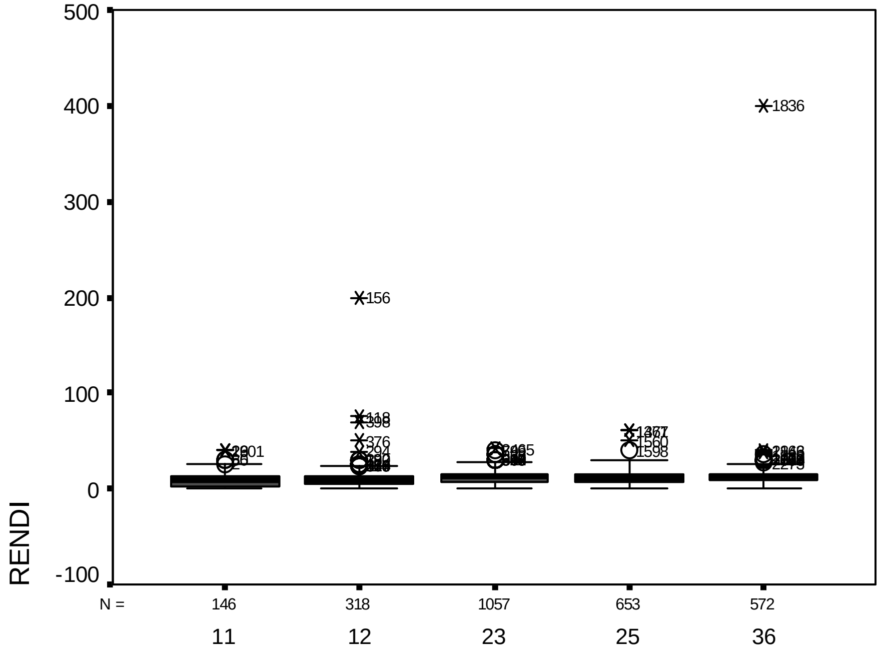
Count

		NOMATP			
		ATP1	ATP2	ATPM	Total
REGION	11	62	29	55	146
	12	163	0	155	318
	23	549	58	450	1057
	25	277	96	280	653
	36	269	71	232	572
Total		1320	254	1172	2746

4.2.2.4 **Explore**. The ***explore*** command displays the empirical probability distribution of a continuous variable. One useful sub-procedure is the stem and leaf plot. Use the commands: ***Analyze, Descriptive statistics, Explore, Dependent list, Factor list, Ok***.

Example: Explore a continuous variable by a categorical variable. Stem and leaf plot analysis: Nicaragua 8

Explore bean yields by region. Dependent list: RENDI and factor list: REGION.



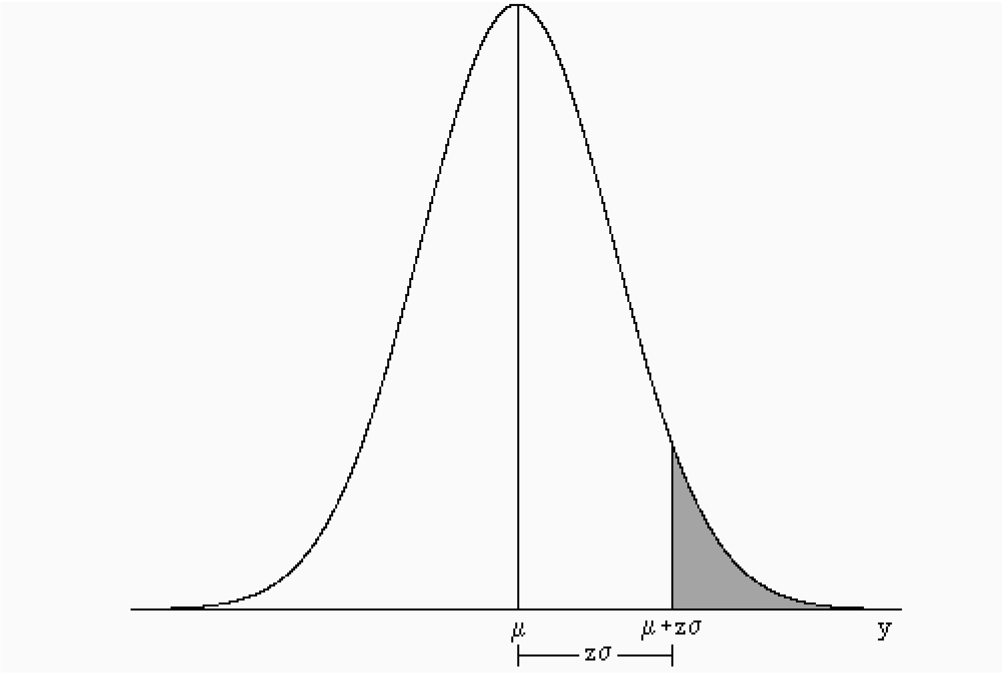
As this picture shows, stem and leaf plot analysis can be very useful for identifying outlier cases.

4.3 Statistical Inference

The purpose of statistical inference is to infer or find out characteristics of a population from characteristics that can be found in a sample of that population. This procedure is used with continuous variables (non-categorical). Statistical inference is subject to error because the available information corresponds only to a portion of the population.

4.3.1 Statistical error types

There are two statistical errors than can be produced in a statistical inference process. Type I error refers to the probability of rejecting a hypothesis when it is true. The associated probability is known as significance level and it is denoted by α. The next graph describes the value of α for a normal distribution with a population mean μ and standard deviation σ.



The type II error, called the "power of test", refers to the probability of accepting a false hypothesis. Given the test structure of statistical hypothesis tests, type II error is mainly associated with the probability of failing in reject a hypothesis when it is false.

4.3.2 Confidence intervals

A confidence interval gives the probability of not making a type I error regarding the mean value. Its formal notation is defined as:

$$P(\mu - z_{\alpha/2}\sigma_{\bar{x}} \le \bar{x} \le \mu + z_{\alpha/2}\sigma_{\bar{x}}) = 1 - \alpha$$

In standardizing the normal distribution base on a mean equal to zero, the formula becomes:

$$P(-z_{\alpha/2} \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le z_{\alpha/2}) = 1 - \alpha$$

4.3.3 The t-test

The t-test is a tool to evaluate statistical validity of a population estimator when using sample data. For example if you expect that the population mean has a value of “c”, your t-test will be formulated in order to determine whether the sample mean is statistically different from the value “c”. In formal terms this test is defined as:

$$t = \frac{\bar{x} - c}{s_{\bar{x}}}$$

The symmetric statistical distribution that is used to analyze this test is called Student’s t. It has special characteristics:

- 67% of the distribution is located within one standard deviation from the mean,
- 95% of the distribution is located within two standard deviations from the mean,
- The significance level is the probability that a t value would be greater than the value of the t-test,
- In larger samples (greater than 30 observations), the Student’s t distribution approaches the normal distribution.

4.3.4 Covariance

The covariance reveals how much two variables vary jointly. The population covariance between variables x_1 and x_2 is defined as:

$$\sigma_{12} = \sum_i (x_{1i} - \mu_1)(x_{2i} - \mu_2)\, p(x_1, x_2)$$

The sample covariance is defined as:

$$s_{12} = \frac{\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n-1}$$

4.3.5 Comparison between two means

The hypothesis that two means are equal can be tested by evaluating whether the difference between the two means is zero. This t-test is defined as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{(\bar{x}_1 - \bar{x}_2)}}$$

The standard deviation required for this test differs from the standard deviation of either of the two distributions (it is lower). The standard deviation of a difference is the square root of the variance of a difference, which is defined as:

$$Var(x_1 - x_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

If the samples are independent, the covariance is zero. This makes the variance of a difference depend on whether both samples share the same population variance or whether they come from independent populations.

The commands required for developing these tests in SPSS are: ***Analyze, Compare means***, then you have choices among ***Means, Sample t-test, Independent samples t-test, Paired sample t-test*** and ***ANOVA***

Example for calculating sample means: Nicaragua 9a
Find mean bean yields for female and male household heads using the variables RENDI and JEFFEM

Report

RENDI

Jefe Femenino	Mean	N	Std. Deviation
0	11.60	2302	7.271
1	10.34	444	6.845
Total	11.40	2746	7.217

Example for an individual t-test: Nicaragua 9b
Is the mean bean yield statistically different from 10 quintales per manzana?

One-Sample Test

	Test Value = 10					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
RENDI	10.134	2745	.000	1.40	1.13	1.67

Results show that the null hypothesis is rejected. There is statistical evidence that the mean bean yield is different from 10 quintales per manzana.

Example for evaluating the difference between two means: Nicaragua 9c
Difference between mean yields of male and female headed households.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RENDI	Equal variances assumed	.063	.802	3.361	2744	.001	1.26	.373	.523	1.987
	Equal variances not assumed			3.502	650.869	.000	1.26	.358	.551	1.959

These test results reject the null hypothesis that the means are equal. The difference of 1.26 quintales per manzana between male headed household and female headed household is significantly different from zero, whether or not the two categories of households are assumed to share a common variance.

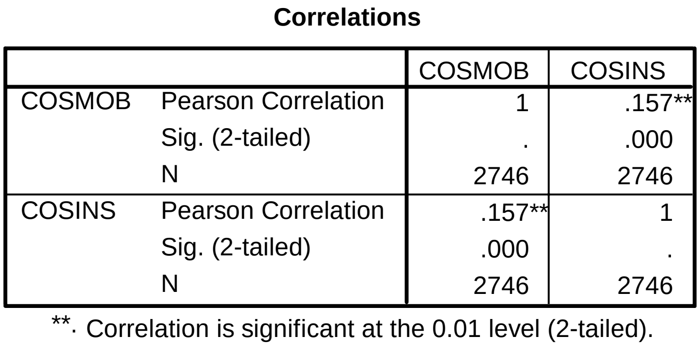
4.4 Correlation Analysis

Correlation analysis examines whether two variables are correlated, which means whether one of them covaries with the other. The correlation coefficient is like a proportional covariance. The correlation can be positive or negative and can have values ranging from -1 to 1. The correlation between variables x_1 and x_2 is defined as:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

In order to obtain the correlation between two continuous variables in SPSS, use the commands: ***Analyze, Correlation, Bivariate, Select variables***

Example of a correlation analysis: Nicaragua 10
Example: Analyze the correlation between the variables labor cost and input cost (COSMOB and CONINS)



Correlations

		COSMOB	COSINS
COSMOB	Pearson Correlation	1	.157**
	Sig. (2-tailed)	.	.000
	N	2746	2746
COSINS	Pearson Correlation	.157**	1
	Sig. (2-tailed)	.000	.
	N	2746	2746

**. Correlation is significant at the 0.01 level (2-tailed).

There is a significant positive correlation between these two variables.

Third day

5. Multiple linear regressions using Ordinary Least Squares (OLS)

In general terms, multiple linear regression with OLS explains the behavior of one variable, called the dependent variable, through the behavior of other variables, called independent or explanatory variables. If y_i is the dependent variable (endogenous) and x_i are the independent variables (exogenous), the linear form of the model for observation *i* is defined as:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_m x_{mi} + u_i$$

Where the β_i are coefficients to be estimated statistically and u_i is an additive random “error” term representing aspects of y_i that could not be explained statistically with the x_{ji} variables.

5.1 Assumptions of the OLS regression

- The dependent variable is continuous,
- All u_i errors are independent of variables $x_{i,}$
- All u_i errors are independent of other u_j errors because $E(u_i u_j)=0$,
- The mean of u_i, $E(u_i)=0$,
- $E(u_i^2)=\sigma^2$.

5.2 Defining a regression model structure

Usually a regression model has a theoretical foundation that posits the causal effect of some independent variables over an outcome variable of interest. An important consideration takes place when the model is incomplete. If an independent variable that affects the dependent variable is omitted, then the coefficient estimates for the other variables will be biased if there is correlation between the omitted variable and the included variables. To avoid such bias it is important to include all the variables that logically could enter into the relationship modeled. However, this will depend on data availability.

Example of model specification: Nicaragua 11

a) **Model definition**

From microeconomic profit maximization, one can derive input demand functions (farmer demand for improved seed, soil technologies, etc) and supply functions (farm production level). According to economic theory, the production input demand depends on output price, input prices, and other relevant variables such as transportation cost and the human, financial and land resources in a typical farm. Similarly, output supply is expected to depend upon the same independent variables, and perhaps some other variables that influence crop production such as weather.

b) **Model specification of an OLS regression**

The process of specifying a regression model is key in empirical work. The challenge is to operationalize the model derived from economic theory. In the case of crop output supply, this is represented by crop yield. Various explanatory variables can be used to explain yield. These variables could arise directly from theory or they could be proxy variables that correlate with those that are desired from theory (but perhaps are difficult to observe).

For example, in lieu of input prices, the input investment cost that each farmer incurs during crop production might be considered. Thus the unit cost of labor, chemical inputs, and other farm services can explain yield levels. According to microeconomic theory, the crop market price should be included as an explanatory variable because a more valuable crop justifies a greater input investment that increases yields. But this condition requires that farmers anticipate the crop price they will get during the harvest period. Normally, this expected price is related to past prices of the same crop. It means that the model should include the expected price that a farmer has before planting, which turns out to be a function of the prices of previous seasons.

Finally other variables should be included to capture other factors affecting production capacity and incentives. For example, managerial ability related to knowledge, via production experience or awareness of existing research. The production setting also depends upon socioeconomic characteristics of households, farm agroecological characteristics, and the policy environment.

5.3 Goodness of fit of the regression line

The most common measure of how well a regression line estimated by OLS fits the data is the coefficient of determination or R^2. This coefficient measures the percentage of the data variability explained by the regression line. The coefficient of determination can be defined as:

$$R^2 = 1 - \frac{SSE}{SST}$$

Where SSE is the Error Sum of Squares and SST is the Total Sum of Squares. Another measure of goodness of fit is the F statistical test related to the entire regression model.

5.4 The F-test and its implementation

There are two useful types of F-test that can be used. The first one measures the explanatory power of the specified regression model. The F-test is a ratio. The numerator measures the change in the aggregate explanation generated by the complete regression. The denominator measures the total variability of the regression. In both cases, it is important to consider the degree of freedom. In the numerator the degrees of freedom equal the number of variables used in the complete model (K), minus one for the constant (K-1), while in the denominator degrees of freedom equal the number of observations minus the number of variables (n-K). In terms of the coefficient of determination, the F-test can be defined as:

$$F(K-1, n-K) = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

If economic theory underlying the model suggests that certain variables need not necessarily be included in a regression model, a statistical test can help to decide whether the variable(s) contribute to explaining the variability of the dependent variable. For evaluating the contribution of only one variable, a t-test can be used. For evaluating more than one variable, a second type of F-test is needed. This test compares the variability explained by a reduced model (without the excluded variables) with the variability of an entire model (that considers all the original variables)

$$F(J, n-K) = \frac{(SSE_{without} - SSE_{with})/J}{(1-R^2)/(n-K)}$$

If the F-statistic is not significant with respect to the desired threshold value (usually α=5%), then a reduced model can be used without losing meaningful explanatory power.

5.5 Expectations about independent variables in the regression

Before running an SPSS analysis, a good researcher should anticipate results based on theory and experience. For example, it is typically useful to develop hypotheses about how the independent variables will affect the dependent variable, including the mathematical signs of these effects. In a linear regression with a continuous variable, the interpretation of the coefficient on an independent variable is the change in the dependent variable if that independent variable were increased by one unit. In interpreting the coefficient on a binary independent variable, the coefficient measures the change in the dependent variable that would occur if the variable equals one (yes). For example, when looking for the difference between the effects of having a female or male household head, if the variable FEMALE=1 for a female household head, a positive coefficient will imply that having a female household head have a greater effect on the dependent variable than having a male household head. If the coefficient is negative, the difference will be in favor of a male household head.

5.6 The use of OLS regression in SPSS

SPSS makes running regressions quite easy. After specifying a regression model, the next step is to define which variable will be the dependent variable and which ones will be the independent variables.

The commands for running a linear regression are: ***Analyze, Regression, Linear, Define variable, Dependent, Define independent variables, Ok***

Example of a Regression using OLS: Nicaragua 12a

Explaining bean yields as a function of output price (PREVENTA), farm area (AREAFIN), bean plot area (AREA), labor cost (COSMOB), input cost (COSINS), cost of services (COSERV), own tenure (TPROPIA), mass-oriented technical assistance (ATPMA), private technical assistance (ATP1), postrera season (POSTRERA) and apante season (APANTE)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.320(a)	.102	.098	6.857

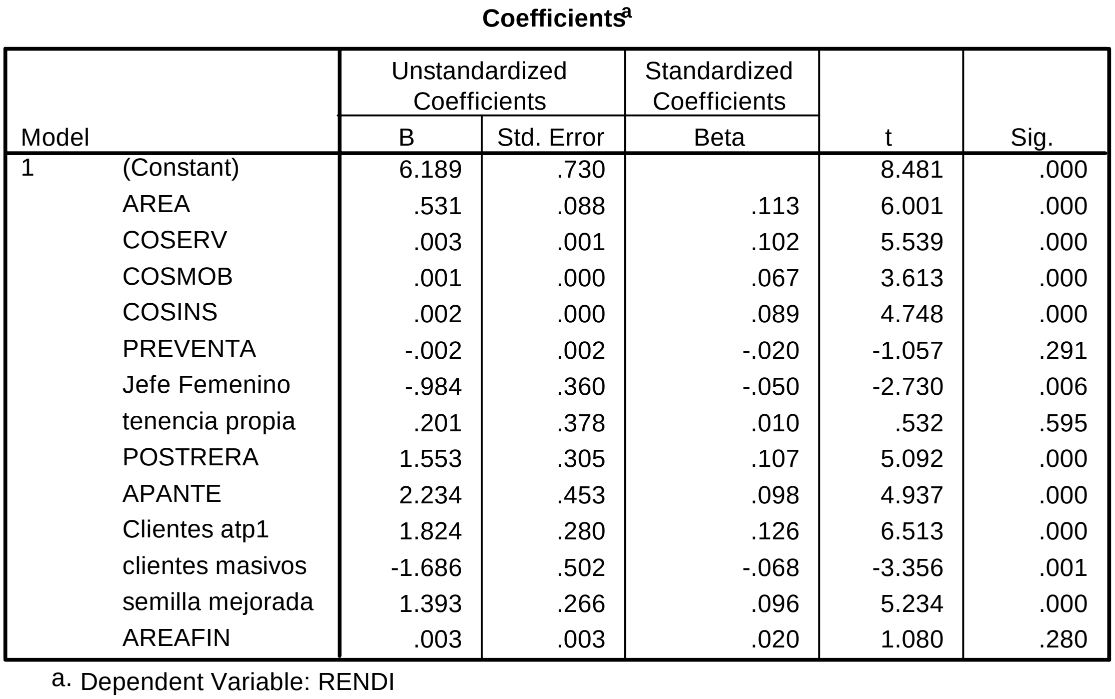
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14611.027	13	1123.925	23.907	.000[a]
	Residual	128205.3	2727	47.013		
	Total	142816.3	2740			

a. Predictors: (Constant), AREAFIN, semilla mejorada, PREVENTA, tenencia propia, COSERV, Jefe Femenino, COSMOB, clientes masivos, APANTE, AREA, COSINS, Clientes atp1, POSTRERA

b. Dependent Variable: RENDI

Example of coefficient estimation using OLS: Nicaragua 12b



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6.189	.730		8.481	.000
	AREA	.531	.088	.113	6.001	.000
	COSERV	.003	.001	.102	5.539	.000
	COSMOB	.001	.000	.067	3.613	.000
	COSINS	.002	.000	.089	4.748	.000
	PREVENTA	-.002	.002	-.020	-1.057	.291
	Jefe Femenino	-.984	.360	-.050	-2.730	.006
	tenencia propia	.201	.378	.010	.532	.595
	POSTRERA	1.553	.305	.107	5.092	.000
	APANTE	2.234	.453	.098	4.937	.000
	Clientes atp1	1.824	.280	.126	6.513	.000
	clientes masivos	-1.686	.502	-.068	-3.356	.001
	semilla mejorada	1.393	.266	.096	5.234	.000
	AREAFIN	.003	.003	.020	1.080	.280

[a] Dependent Variable: RENDI

Example of coefficient interpretation: Nicaragua 13
First results

From the first table, the most important result is the coefficient of determination (R^2). In the model specified, the independent variables included explain 10.2% of the variability of bean yields. Although this coefficient is not high, usually cross sectional models of agricultural yields have low R^2 values.

The second table (ANOVA) summarizes the significance level of the whole model. The joint F-test with a value of 23.907 rejects the null hypothesis that the explanatory variables have no effect over the dependent variable.

The third table contains information about coefficient estimates. The first step is to determine which variables have significant individual effects over the dependent variable and which ones do not. Results show that there is not statistical evidence that the variables TPROPIA, AREAFIN and PREVENTA explain the level of bean yields. For the remaining variables the individual t-tests reject the individual null hypotheses that these variables do not have a significant effect on bean yields.

Example of the use of the F-test: Nicaragua 14

How would the model be affected by the elimination of variables with insignificant t-statistics? An F-test can offer an answer

TPROPIA, AREAFIN and PREVENTA are candidates to be removed from the model. For implementing the F-test, two regressions were estimated. The first included the 13 original variables, and in the second, is a reduced model where the three “insignificant” variables were excluded. The R^2 of both regressions are needed for calculating this F-statistic. In the complete regression the R^2 is 0.102 and in the restricted regression 0.101. The test measures the effect on the model’s explanatory power of removing these 3 variables out of the 13 original variables. The null hypothesis states that the 3 variables to be excluded do not have a jointly significant effect on bean yields. In this example, the F-statistic is defined as:

$$F = \frac{(0.102 - 0.101)/3}{(1 - 0.102)/(2040)} = 0.7572 < F_{\alpha=0.05}(3{,}2040) = 3.00$$

According to this result, the test fails to reject the hypothesis that the 3 variables do not have a significant effect on the dependent variable. So these variables can be eliminated.

Results of the reduced model: Nicaragua 15a

R^2 and ANOVA

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.319(a)	.101	.098	6.855

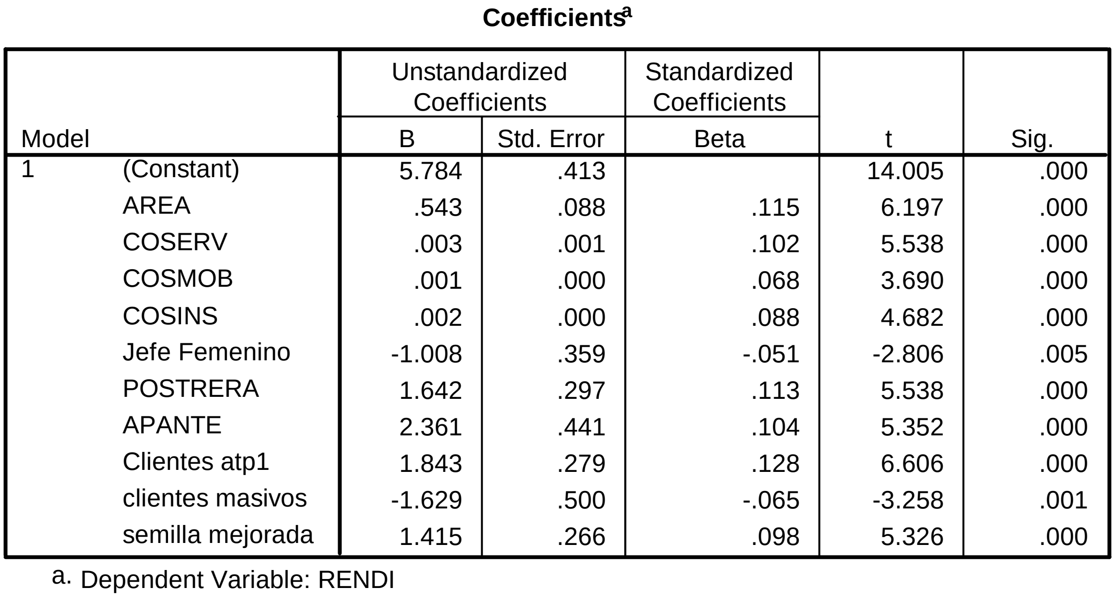
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14508.818	10	1450.882	30.875	.000[a]
	Residual	128477.9	2734	46.993		
	Total	142986.7	2744			

a. Predictors: (Constant), semilla mejorada, APANTE, COSMOB, Jefe Femenino, COSERV, clientes masivos, AREA, COSINS, Clientes atp1, POSTRERA

b. Dependent Variable: RENDI

Results of the reduced model: Nicaragua 15b



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.784	.413		14.005	.000
	AREA	.543	.088	.115	6.197	.000
	COSERV	.003	.001	.102	5.538	.000
	COSMOB	.001	.000	.068	3.690	.000
	COSINS	.002	.000	.088	4.682	.000
	Jefe Femenino	-1.008	.359	-.051	-2.806	.005
	POSTRERA	1.642	.297	.113	5.538	.000
	APANTE	2.361	.441	.104	5.352	.000
	Clientes atp1	1.843	.279	.128	6.606	.000
	clientes masivos	-1.629	.500	-.065	-3.258	.001
	semilla mejorada	1.415	.266	.098	5.326	.000

[a] Dependent Variable: RENDI

Example of OLS coefficient interpretation: Nicaragua 16

The coefficients in the third table above have a special interpretation in terms of changes in the dependent variables.

The constant of the model states that a farmer without improved seed, with no labor, input or services investment, without any technical assistance from the project, and who plants beans in the Primera season, has an average bean yield of 5.784 quintals per manzana of land (NB: 1 manzana = 0.7 hectares).

To illustrate the interpretation of a continuous variable, consider the variable AREA (bean planted area). According to the regression, if a farmer increases the bean area by 1 manzana, the farmer can expect an increase in bean yield of 0.543 quintals per manzana. In other words, bean yields tend to rise with planted area.

The interpretation of the coefficient of a binary variable differs slightly. Using the variable ATP1, the estimated coefficient states that on average a farmer who receives private technical assistance achieves yields that are 1.843 quintals per manzana more than a farmer without any kind of technical assistance, keeping everything else constant (both farmers use the same improved seed, farm in the same season, have the same bean area, and invest the same amount in labor, inputs and services).

5.7 Specifying another functional form

Many times linear regression models do not represent accurately the relationship between a dependent variable and the explanatory variables. In this case it may be necessary to evaluate other functional forms that allow better representation of the cause-effect relationship between both kinds of variables. Apart from linear regression, quadratic models and logarithmic models are the most preferred.

Example: Rationale for choosing a quadratic functional form. Nicaragua 17

Given the available data, a quadratic model can be specified. There could be two hypotheses about how the size of the bean area influences the yield level. One hypothesis claims that the greater the bean area, the greater the bean yields because of economies of scale. The second hypothesis, contrarily, claims that the smaller the area planted with beans, the greater the yields because of more intensive management. By adding a quadratic term for the area planted in beans, to the original model, we can test the hypothesis that yields may increase with size up to some point, and then decline.

Example of the estimation of a quadratic model using OLS: Nicaragua 18a
Including a quadratic term for the bean area (AREA2)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.340(a)	.116	.112	6.803

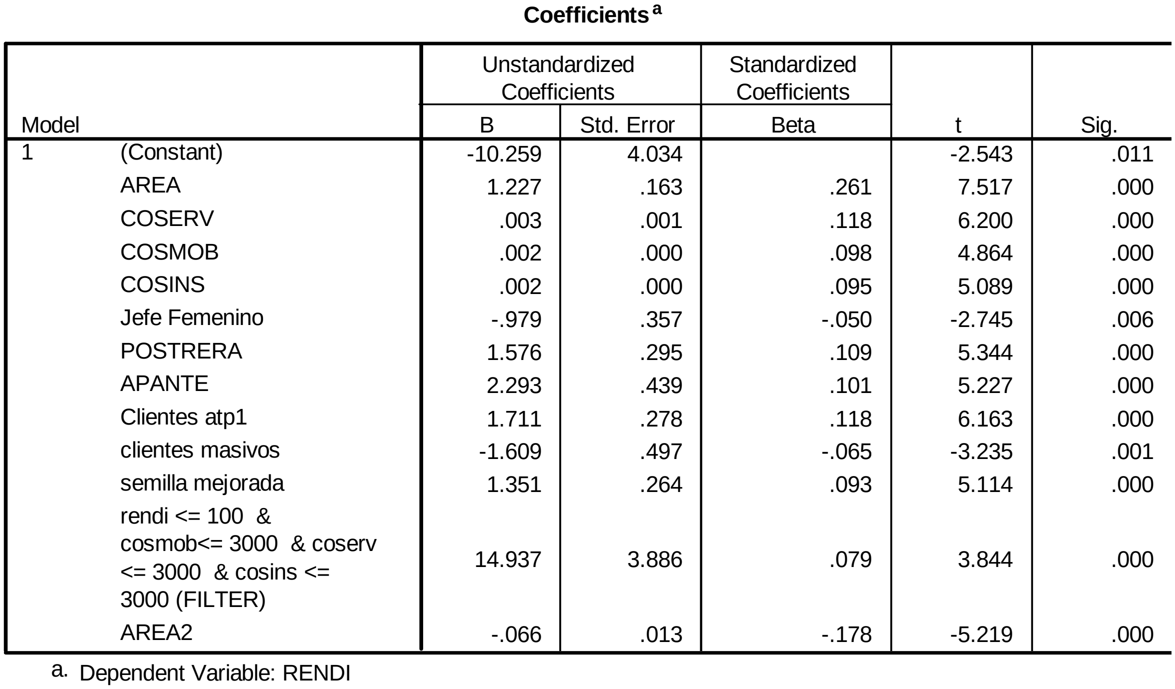
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16529.761	12	1377.480	29.759	.000[a]
	Residual	126456.9	2732	46.287		
	Total	142986.7	2744			

a. Predictors: (Constant), AREA2, rendi <= 100 & cosmob<= 3000 & coserv <= 3000 & cosins <= 3000 (FILTER), POSTRERA, Jefe Femenino, semilla mejorada, Clientes atp1, COSINS, COSERV, APANTE, clientes masivos, COSMOB, AREA

b. Dependent Variable: RENDI

Example of the estimation of a quadratic model using OLS: Nicaragua 18a
Including a quadratic term in the bean area (AREA2)



Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-10.259	4.034		-2.543	.011
	AREA	1.227	.163	.261	7.517	.000
	COSERV	.003	.001	.118	6.200	.000
	COSMOB	.002	.000	.098	4.864	.000
	COSINS	.002	.000	.095	5.089	.000
	Jefe Femenino	-.979	.357	-.050	-2.745	.006
	POSTRERA	1.576	.295	.109	5.344	.000
	APANTE	2.293	.439	.101	5.227	.000
	Clientes atp1	1.711	.278	.118	6.163	.000
	clientes masivos	-1.609	.497	-.065	-3.235	.001
	semilla mejorada	1.351	.264	.093	5.114	.000
	rendi <= 100 & cosmob<= 3000 & coserv <= 3000 & cosins <= 3000 (FILTER)	14.937	3.886	.079	3.844	.000
	AREA2	-.066	.013	-.178	-5.219	.000

[a] Dependent Variable: RENDI

Model results show a significant effect of the quadratic term. In combining the effects of the variables AREA and AREA2, the effect of the bean area on the yields of this crop is increasing at a diminishing rate. According to this trend there will be a point beyond which greater bean area will reduce the bean yield. From differential calculus, the inflexion point up to which increasing the bean area still has a positive effect is around 9 manzanas. After that, greater bean area will reduce the average yield. Additionally, note from the higher R^2 that this model explains greater variability of the bean yields.

6. Regression with binary dependent variables

Binary regression models have as a dependent variable a binary variable that takes the value “1” or “0”. An important use of this kind of model is for “explaining” the determinant factors of the adoption (or non adoption) of a new technology. The theoretical foundation of these models can be found in the theory of “random utility”. This concept presupposes a level of satisfaction (utility) that a consumer achieves when consuming something. Such consumption can include the consumption of production inputs, like a new technology. Unfortunately, the true level of satisfaction cannot be observed. Only the consumer’s choice can be observed, that is, whether or not the consumer buys a good or service. This choice could be the adoption (value of “1”) or not (value of “0”) of a technology.

As the binary regression has a dependent variable that takes values of 0 and 1, the estimated regression model predicts the probability that the dependent variable takes a value of 1 (the probability that a farmer adopt a technology). The coefficients of the independent variables state how these variables influence the probability that the dependent variable occurs (e.g., the probability of adopting a new technology).

Binary models are useful for explaining technology adoption processes when there are no continuous variables available for the analysis. There are many models of binary dependent variable, but one of the most used one is logit regression.

6.1 Logit regression

In this case the probability that the dependent variable *y* (conditioned to the independent variables x_i) will be 1 can be expressed by the following logistic form:

$$P_i = E(y = 1 \mid x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots)}}$$

The logarithm of the odds ratio (P/(1-P)) (called a “logit”) has a linear functional form:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

The interpretation of the estimated coefficient in the regression differs from the traditional interpretation of OLS. As with OLS, the direction of the effect follows the sign of the estimated coefficient. However, the marginal effect varies with the magnitude of the independent variables.

Example of a Logit Model: Nicaragua 19
From economic theory we expect that the demand for a production input will depend on the market output price, the production input prices, and other variables that affect incentives and capacity to use the input. However, for certain inputs, one can only observe whether a farmer adopts the input or not. In this example, the dependent variable is the adoption of improved seed. This adoption process is explained by output price, farm area, bean area, labor cost, input cost, services cost, improved seed, female household head, own tenure, mass-oriented technical assistance, private technical assistance, Postrera season and Apante season.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3710.079(a)	.024	.033

.

Classification Table [a]

			Predicted		
			semilla mejorada		Percentage Correct
	Observed		0	1	
Step 1	semilla mejorada	0	478	769	38.3
		1	380	1114	74.6
	Overall Percentage				58.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	PREVENTA	-.001	.001	2.507	1	.113	.999
	AREAFIN	-.001	.001	.578	1	.447	.999
	JEFEFEM	-.254	.106	5.705	1	.017	.776
	TPROPIA	.122	.112	1.200	1	.273	1.130
	POSTRERA	.167	.091	3.415	1	.065	1.182
	APANTE	.162	.134	1.444	1	.229	1.175
	ATP1	.495	.083	35.936	1	.000	1.640
	ATPMA	.371	.149	6.225	1	.013	1.449
	COSMOB	.000	.000	.263	1	.608	1.000
	COSINS	.000	.000	4.671	1	.031	1.000
	COSERV	.001	.000	8.821	1	.003	1.001
	AREA	.015	.027	.315	1	.575	1.015
	Constant	-.308	.216	2.031	1	.154	.735

a. Variable(s) entered on step 1: PREVENTA, AREAFIN, JEFEFEM, TPROPIA, POSTRERA, APANTE, ATP1, ATPMA, COSMOB, COSINS, COSERV, AREA.

The model has a very low coefficient of determination and its prediction power is poor. The model predicts adoption of improved seed correctly only 58% of the time. It predicts adoption correctly in 74.6% of cases, but non-adoption in only 38.3% of cases.

Fourth and fifth days

These days are devoted to developing group projects. The topic of each project should be proposed by the participants, based on their personal and institutional interest. Some examples of projects are:

- Effect of technical assistance in the adoption of specific crop varieties,
- Factors that influence the adoption of a specific technology,
- Effects of technology adoption on crop yields.

After working in groups a summary and a public presentation are suggested in order to discuss the statistical methods used and the results gotten from the statistical analysis.

Useful References

Beals, R. E. (1972) *Statistics for Economists: An Introduction*. Chicago: Rand McNally.

Mendenhall, W., R. L. Scheaffer, and D. D. Wackerly (1986). *Mathematical Statistics with Applications*. 3rd ed. Boston: Duxbury.

Snedecor, G. W., and W. G. Cochran (1967). *Statistical Methods*. Sixth ed. Ames, IA: Iowa State University Press.

Weisberg, S. (1985) *Applied Linear Regression*. New York: Wiley.

Wolf, C. (1990). “Computer Analysis of Survey Data: File Organization for Multi-Level Data”. Agricultural Economics Computer Service, Michigan State University. http://www.aec.msu.edu/agecon/fs2/survey/levels.pdf