STAFF PAPER

INTRODUCCIÓN A LA ESTADÍSTICA PARA ECONOMISTAS AGRÍCOLAS USANDO EL SPSS

Scott M. Swinton y Ricardo Labarta

Staff Paper 2003-13S **September, 2003**

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

Introducción a la Estadística para Economistas Agrícolas Usando el SPSS

Scott M. Swinton y Ricardo Labarta
swintons@msu.edu

Resumen

Este documento es un manual de estadística para economistas aplicados que usan el programa estadístico SPSS. Está diseñado para ser usado dentro de un taller de capacitación de una semana, y tiene como propósito familiarizar a profesionales de la investigación con procedimientos estadísticos básicos, necesarios para realizar análisis socioeconómicos a datos que provienen de encuestas. Este documento es una introducción para los usuarios en la creación y manipulación de las bases de datos, en la estadística exploratoria univariada y bivariada, en la prueba de hipótesis y en la regresión lineal y logística. El texto del documento se apoya en 19 cajas de texto que ilustran como cada procedimiento puede ser aplicado a un grupo de datos provenientes de una encuesta de finca.

36 pages

INTRODUCCIÓN A LA ESTADÍSTICA PARA ECONOMISTAS AGRÍCOLAS USANDO EL SPSS[1]

Scott M. Swinton y Ricardo Labarta[2]

Michigan State University
Department of Agricultural Economics
Staff Paper No. 03-13S

septiembre de 2003

Resumen

Este documento es un manual de estadística para economistas aplicados que usan el programa estadístico SPSS. Está diseñado para ser usado dentro de un taller de capacitación de una semana, y tiene como propósito familiarizar a profesionales de la investigación con procedimientos estadísticos básicos, necesarios para realizar análisis socioeconómicos a datos que provienen de encuestas. Este documento es una introducción para los usuarios en la creación y manipulación de las bases de datos, en la estadística exploratoria univariada y bivariada, en la prueba de hipótesis y en la regresión lineal y logística. El texto del documento se apoya en 19 cajas de texto que ilustran como cada procedimiento puede ser aplicado a un grupo de datos provenientes de una encuesta de finca.

[1] Basado en un taller de capacitación entre el Instituto Nicaragüense de Tecnología Agropecuaria y el Proyecto Bean-Cowpea CRSP, Montelimar, Nicaragua, del 20 al 24 de enero de 2003. Los autores agradecen a Lesbia Rizo por permitir el uso de una base de datos del INTA en los ejemplos.
[2] Scott M. Swinton (swintons@msu.edu) es professor y Ricardo Labarta (labartar@msu.edu) es investigador posgrado asistente en el Department of Agricultural Economics de Michigan State University, East Lansing, MI 48824-1039.

PRIMER DÍA

Introducción al análisis estadístico

Desarrollar un análisis estadístico a través de una herramienta como el SPSS puede compararse a utilizar un nuevo equipo de cocina. Este equipo no tendrá valor si no se utiliza para cocinar. Pero para cocinar, se necesita más que sólo este equipo. También se necesitan ingredientes (que serían los datos en el análisis), otros equipos (como las hojas de cálculo), y sobretodo, el conocimiento para cocinar (en el caso del análisis, el conocimiento de métodos de investigación y del uso de estadística).

El objetivo del taller no es solamente conocer el uso de un programa como el SPSS, sino también aportar ideas sobre cómo diseñar un buen plan de análisis. Para asegurar la participación plena del grupo durante todo el taller, se debe alternar entre cuatro actividades:

a) Información general sobre el programa SPSS
b) Principios de diseño de investigación y análisis estadístico
c) Aplicaciones a los datos de interés de los participantes..
d) Proyectos de investigación grupales de los participantes del taller

1. Presentación general del SPSS

Ventajas para el uso de SPSS. Este programa es de fácil manejo especialmente en la manipulación y en el análisis de grandes bases de datos. El SPSS puede leer directamente información de hojas de cálculo y bases de dato en formatos como DBF, WK1 y XLS. Otra gran ventaja sobre otros programas similares es la diversidad de formas de presentar los resultados tanto en tablas como en gráficos.

Bases de datos en SPSS. Una base de datos consta de una estructura general que se construye sobre la base de variables y observaciones. La idea de una base de datos puede ser asociada con una matriz, como las hojas de cálculo, donde cada fila contiene una observación y cada columna contiene datos sobre una variable particular (a través de muchas observaciones) (Wolf 1990). Cuando se tienen datos de múltiples niveles, es preferible construir una base de datos con una observación diferente por cada unidad de análisis.

Archivos generados en SPSS. El SPSS genera archivos con extensión propia que difieren de acuerdo al tipo de información que contienen cada uno (bases de datos, resultados). Las extensiones más comunes son SAV para archivos de datos, SPS para archivos de sintaxis o comandos y SPO para archivos de resultados.

2. Creación de una base de datos en SPSS

2.1. Ingreso directo de datos. Se puede ingresar la información de encuestas directamente al SPSS generando una base de datos con extensión SAV. Lo primero que se debe hacer es crear las variables que contendrá la base de datos y darles a cada una un nombre de no más de ocho caracteres. Añadir etiquetas informativas es también recomendable para describir las características de cada variable.

2.2 Importación de una base de datos a SPSS. Bases de datos generadas en otros programas informáticos como Excel, Access y FoxPro, pueden ser convertir a archivos SAV. Para realizar la conversión, basta con abrir el archivo dentro del programa de SPSS. Luego este archivo puede ser guardado con extensión SAV.

2.3. Uso de bases de datos transformadas al SPSS. Debemos analizar el tipo de base de datos transformada y determinar si la estructura actual y la información contenida en ella nos permitirán alcanzar los objetivos del análisis. Para este fin es necesario seguir cinco pasos:

a) Determinar cuáles son las características de la base de datos actual: número de variables, tipos de variables, número de observaciones y el nivel de cada observación (Wolf 1990).
b) Verificar si toda la información referida a una unidad de análisis (ej. Agricultor, hogar) está incluida en un sólo
c) Verificar si existen variables redundantes o si todas las variables contienen información única.
d) Verificar si los nombres de las variables y sus etiquetas son los apropiados.

Ejemplo de importación y revisión de archivos diferentes al SPSS: Nicaragua 1
Importación de las bases de datos SISTEMAS.XLS y GENERAL.XLS. Este ejemplo también permite diagnosticar el tipo de estructura que tiene el archivo importado.

Analizando la base de datos transformada:

a) Esta base de datos contiene información a nivel de productor, a nivel de cultivo y a nivel de parcelas. Hay más de 7,000 observaciones para cada variable.
b) Existe información importante para los mismos productores en archivos diferentes (SISTEMAS.SAV y GENERAL.SAV).
c) Existe más de un registro por productor. Ejemplo: hay información de diferentes cultivos de un mismo productor en registros separados. También hay parcelas de un mismo cultivo y de un mismo productor en registros separados. Sin embargo, para analizar el comportamiento del productor es preferible una base de datos con un sólo registro por productor.
d) Hay variables redundantes con la misma información (ejemplo código de variables y descripción de las mismas). Esto origina duplicidad, incrementa el tamaño del archivo y es poco práctico.
e) Los nombres de las variables son largos, causan problemas a los programas antiguos y no son prácticos para la ejecución del SPSS. Es mejor asignar a las variables nombres cortos y claros, así como etiquetas con información completa.

3. Creación de una base de datos a nivel de la finca

3.1 Uso de datos pre-existentes en más de una base de datos

La estructura de una base de datos debe ser de la manera más flexible posible. Dependiendo de los objetivos y el tipo de análisis que se desarrollará con la información, el manejo y análisis de datos se facilitaría con más de una base de datos. En este caso, todas las bases de datos deberán incluir una variable indicativa común que permita vincular la información existente en más de una base de datos.

Cuando la información se ingresa de manera directa, el diseño de la estructura es definido desde el inicio. Si la información ya se encuentra ingresada en otro programa informático (Excel, Dbase, etc), es necesario analizar si se quiere conservar la estructura existente o si se desea crear una nueva estructura más apropiada. Si se conserva la estructura existente, sólo es necesario importar el archivo dentro del SPSS como fue explicado previamente. Variar la estructura de un archivo o archivos pre-existentes requiere alguna manipulación dentro del SPSS, dependiendo que ajustes o modificaciones se pretenda hacer a la base de datos. Estas modificaciones pueden incluir la unión de dos o más archivos con información complementaria, la reducción del número de variables, la reducción del número de observaciones y otras. El procedimiento de unir archivos es tal vez el que requiere un mayor trabajo y los pasos son detallados a continuación:

a) ordenar en forma ascendente las bases de datos en SPSS a unir, guiándose por la variable en común entre ambas base de datos. Los comandos requeridos son: ***Datos, Ordenar casos, Ordenar según*** **(VARIABLE COMUN)*, Ordenar de manera ascendente, Ejecutar***.

b) Proceder a fundir archivos. Mantener activado uno de los archivos y usar el comando fundir guiándose por la variable común a ambos archivos: ***Datos, Fundir archivos, Agregar variables, Nombre del archivo a unir, Variable clave (***PRODUCTOR***), ejecutar.***

c) Identificar posibles problemas generados por la forma de la estructura original. Muchas veces las bases de datos unidas no contienen el mismo número de observaciones para las mismas variables comunes.

d) Ajustar la nueva base de datos creada. Si las bases de datos unidas generan una nueva con datos faltantes por diferencias en las estructuras de las bases originales, podría ser necesario hacer ajustes a la nueva base de datos de manera manual.

Ejemplo para unir archivos: Nicaragua 2
Unión de los archivos SISTEMAS.SAV y GENERAL.SAV

a) Ordenar los archivos SISTEMAS.SAV y GENERAL.SAV teniendo como variable común PRODUCTOR. La ejecución es: ***Datos, Ordenar casos, Definir variables guía (*PRODUCTOR***), Ordenar de manera ascendente, Ejecutar***.
b) Tomar como archivo base SISTEMAS.SAV y añadir GENERAL.SAV. La variable común es PRODUCTOR. ***Datos, Fundir archivos, Agregar variables, Archivo*** (GENERAL.SAV)***, Variable común*** (PRODUCTOR)***, Ejecutar.***
c) Revisar si se originaron problemas al unir las bases de datos. Una sugerencia útil es discutir la historia de la encuesta y del ingreso de datos, para poder entender mejor la organización de la base de datos
d) Uno de los problemas se origina por la presencia de varios registros para un mismo productor en una de las bases. El SPSS asignará la nueva información sólo al primer registro de cada productor. Se necesita copiar esta nueva información manualmente en los demás registros usando copiar y pegar

A manera de ejemplo corregir manualmente la variable SEXO. Para esto se procede a copiar el "valor" de SEXO (F o M) que aparece sólo en el primer registro de cada productor, en los demás registros del mismo productor. Con este procedimiento la variable SEXO tendrá un valor para cada uno de los registros del archivo.

De manera similar se pueden corregir las variables: MODATP, MUNICIPIO, REGIÓN, ÉPOCA, NOMTIENE, ANOINGRE y todas las variables que tienen valores faltantes después de unir los archivos. El producto final será la creación de la base de datos que puede ser llamada BASE.SAV que tendrá la estructura a nivel de campo de cultivo de SISTEMAS.SAV y vinculada a los datos de hogar de GENERAL.SAV.

3.2 **Recodificación y generación de nuevas variables**

Este procedimiento permite redefinir variables. Por ejemplo, muchos procedimientos estadísticos funcionan mejor con variables numéricas antes que alfanuméricas o variables de cadena. Todas las variables recodificadas o tipos de variables son generadas a partir de variables existentes.

La recodificación es un proceso muy usado en SPSS y de fácil ejecución. Al igual que cualquier variable en SPSS, es aconsejable dar a las variables recodificadas nombres cortos e indicativos. Además, es necesario acompañar estos nombres con etiquetas que contengan información más descriptiva. La variable recodificada generada puede mantener su nombre original o recibir uno nuevo. En el caso de una variable binaria, es costumbre vincular el nombre a la presencia de un atributo que implique Si=1 y No=0.

Los comandos requeridos para la recodificación son: ***Transformar, Recodificar, En variable diferente, Seleccionar variable, Colocar nombre en variable de salida, Cambiar, Variables viejas y nuevas, Asignar valores (¿cuáles valores nuevos corresponden a valores viejos?)***

Ejemplo para recodificar variables: Nicaragua 3
Recodificación de la variable SEXO (creación de una variable binaria)

Para generar una variable binaria relacionada a la variable SEXO, se puede asociar esta nueva variable con el género del jefe de la familia. La variable para "Jefe de familia femenino" (JEFEFEM) puede ser creada asignándole el valor "1" si corresponde a si y el valor "0" si corresponde a no.

Debido a que la variable SEXO es alfabética, cuando se genera la variable JEFEFEM los antiguos valores "F" y "M" deberán ser reemplazados por "1" y "0" respectivamente.

3.3 Creación de variables categóricas

Las variables categóricas son variables discretas (no continuas) que indican la categoría a la cual pertenece el registro (ejemplo: variedad de semilla, región). Estas variables son muy útiles para hacer tablas cruzadas o estadísticas descriptivas por categoría como se verá más adelante. La generación de variables binarias a partir de estas variables categóricas es más sencilla y facilita el análisis.

Los comandos necesarios para crear variables categóricas son: ***Transformar, Recodificación automática, Variable a cambiar, Nombre de variable a ser transformada, Ejecutar***

Ejemplo para generar variable categórica: Nicaragua 4
Transformación de la variable alfabética TENENCIA en una nueva variable categórica NTENEN.
Este procedimiento produjo los siguientes valores:

1 = Alquilada
2 = Mediaría
3 = Prestada
4 = Propia
5 = R.A.

El mismo procedimiento puede ser aplicado a las variables TPROPIA, POSTRERA, APANTE, ATP1, ATPMA, SEMMEJ, OCCSA

3.4 Revisión de la base de datos y creación de sub-bases

Después de generar y recodificar las variables necesarias, es necesario revisar la nueva base de datos y decidir si está completa y si todos los datos serán usados en el análisis. Muchas veces una base de datos contiene mayor información que la que requieren análisis específicos y retarda el proceso. Por ejemplo, si se quiere analizar específicamente la información sobre un cultivo, la información sobre otros cultivos es superflua. En este caso hay dos caminos a seguir:

a) Se mantiene la base de datos completa y se restringen los datos que se usarán en cada análisis. Este procedimiento restringirá el análisis a observaciones especificadas. Los comandos necesarios para esto son: ***Datos, Seleccionar casos, Condición de si (definir la variable y el valor de la misma que queremos restringir), Continuar, Ejecutar***.

b) Dividir la muestra en sub-bases de acuerdo al tipo de información que se quiere usar. Por ejemplo se puede crear una sub-base con información referida sólo a campos de cultivo con un cultivo específico para reducir el tamaño de la base de datos. Este procedimiento requiere ordenar la base de datos usando una variable clave como factor (ejemplo: cultivo). Luego se eliminan todas las observaciones que no pertenezcan a un cultivo determinado. Finalmente se graba el archivo resultante con un nuevo nombre.

SEGUNDO DÍA

3.5 Depuración de los datos

Este procedimiento mejora la calidad de los datos que se tiene. En particular es importante ver si la base de datos contiene observaciones con valores no esperados. Estas observaciones particulares son conocidas como valores extremos. El SPSS brinda varios procedimientos que permiten detectar su presencia y corregirlos. Cómo se verá en la siguiente sección, estos procedimientos se ejecutan de acuerdo al tipo de variable que queramos analizar. Los valores extremos no son obligatoriamente erróneos. Si son ciertos, pueden ser muy informativos. Pero si estos valores provienen de algún tipo de error, hay que proceder a corregirlos o borrarlos (ver el comando ***explorar*** en la página 14).

4. Introducción a la estadística descriptiva

4.1 Base Teórica

El propósito de la estadística descriptiva es ayudar a sacar inferencias de una población, observando miembros de una muestra de la población. Las estadísticas descriptivas más informativas son las medidas de ubicación de una distribución y de dispersión de la misma.

4.1.1 Medidas de ubicación de una distribución

La media es la principal medida de ubicación y por definición, depende de la probabilidad de cada caso. Su formulación poblacional es:

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

Donde x_i es el valor de cada observación y $p(x_i)$ es su probabilidad asociada. En el caso de una muestra, se asume que la probabilidad de ocurrencia es igual para cada observación por lo que la formula se convierte en:

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

Otras medidas de ubicación son la mediana (el valor del punto medio en un grupo de datos ordenado) y la moda (el valor de ocurrencia más frecuente).

4.1.2 Medidas de dispersión

Las medidas de dispersión más usadas son la varianza, la desviación estándar y el coeficiente de variación (CV). Para una población específica con media poblacional μx y probabilidad de ocurrencia $p(x_i)$ para cada observación la varianza se define como:

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

Por su parte la varianza muestral se define como:

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

La desviación estándar permite tener una medida de dispersión con las mismas unidades que la media:

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

Finalmente el coeficiente de determinación se calcula como la razón entre la desviación estándar y la media:

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

Un coeficiente de variación mayor a 0.5 implica que la media no es diferente de cero con 95% de confianza si los datos siguen una distribución de probabilidad normal (Ver la sección 4.3.3 abajo)

4.2 Las estadísticas descriptivas en el SPSS

El objetivo de esta sección es llegar a conocer formas para explorar los datos usando métodos tanto gráficos como estadísticos. Un análisis exploratorio de datos es un primer paso importante antes de proceder con métodos más formales.

4.2.1 **Gráficos**. El SPSS ofrece una serie de posibilidades para hacer un análisis gráfico con la información de una base de datos. Las opciones más usadas son: ***Histogramas, Barras de error*** y ***Gráficos de dispersión***. Todas estas opciones son accesibles a partir del comando ***Gráficos***.

4.2.2 **Diagnósticos estadísticos**. Estos procedimientos permiten un análisis numérico y están agrupados bajo los comandos ***Analizar, Estadísticas descriptivas***. Esta opción ofrece un buen grupo de opciones para ejecutar análisis univariado y bivariado, tanto para variables categóricas como para variables continuas. Los cuatro procedimientos a continuación son muy útiles.

4.2.2.1 **Frecuencias**. Este procedimiento permite conocer la distribución de frecuencias de las variables categóricas. Para ejecutar este procedimiento se usan los comandos: ***Analizar, Estadísticas descriptivas, Frecuencias, Variable, Ejecutar***

Ejemplo de la distribución de frecuencias de una variable categórica: Nicaragua 5
Ejemplo: Frecuencias de la variable tecnología de semillas (TECSEM).

TECSEM

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		1251	45.6	45.6	45.6
	0	2	.1	.1	45.6
	1	12	.4	.4	46.1
	2	3	.1	.1	46.2
	4	3	.1	.1	46.3
	7	1	.0	.0	46.3
	8	2	.1	.1	46.4
	9	705	25.7	25.7	72.1
	10	49	1.8	1.8	73.9
	11	270	9.8	9.8	83.7
	12	258	9.4	9.4	93.1
	13	28	1.0	1.0	94.1
	14	43	1.6	1.6	95.7
	15	63	2.3	2.3	98.0
	16	19	.7	.7	98.7
	17	25	.9	.9	99.6
	18	10	.4	.4	99.9
	20	2	.1	.1	100.0
	Total	2746	100.0	100.0	

En la tabla se puede notar un gran número de casos con valores faltantes (estos representan las tecnologías que no incluyen semillas

4.2.2.2. **Descriptivas**: Este procedimiento permite conocer las principales medidas estadísticas de variables continuas: media, desviación estándar, valor mínimo, valor máximo. Los comandos necesarios son: ***Analizar, estadísticas descriptivas, descriptivas, variables, ejecutar***

Ejemplo de estadísticas descriptivas de variables continuas: Nicaragua 6
Ejemplo: estadísticas descriptivas de rendimiento (RENDI) y área de fríjol (AREA):

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
RENDI	2746	0	75	11.40	7.217
AREA	2746	0	23	1.55	1.533
N válido (según lista)	2746				

4.2.2.3. **Tablas cruzadas (de contingencia)**. Este comando genera la distribución conjunta de dos variables categóricas. Se usan los comandos: ***Analizar, Estadísticas descriptivas, Tabla de contingencia, Variable de fila, Variable de columna, Ejecutar***

Ejemplo de generación de tablas cruzadas: Nicaragua 7
Ejemplo: generar una tabla cruzada de las regiones según grupo ATP (Asistencia técnica). Uso de variables NREGION y NOMATP

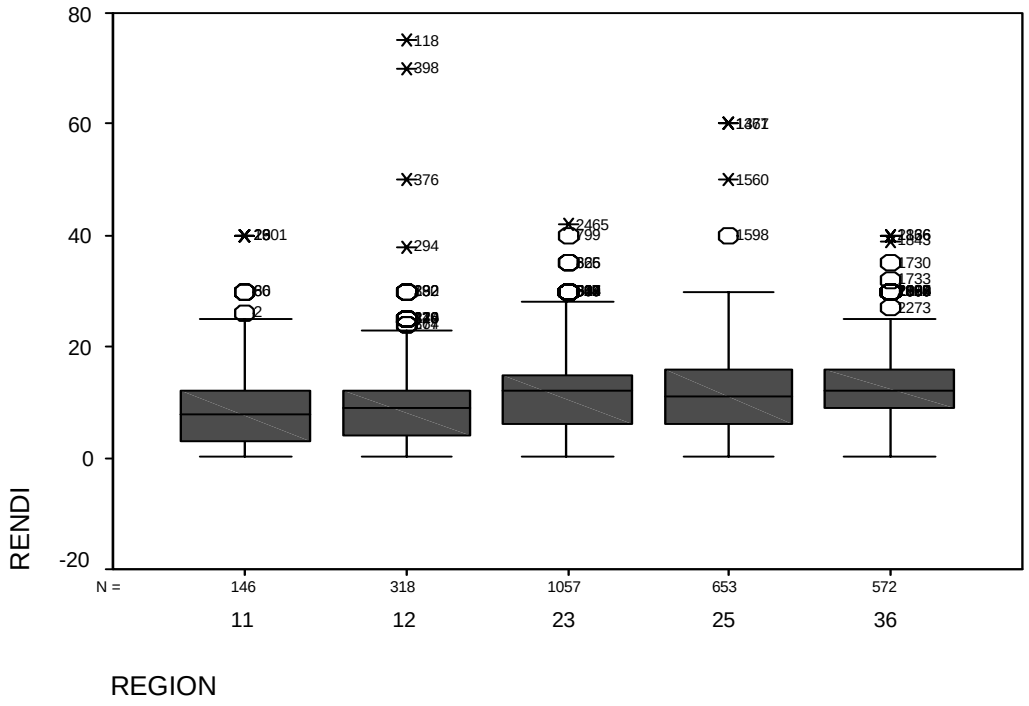
Tabla de contingencia REGION * NOMATP

Recuento

		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. **Explorar**: Este comando muestra la distribución de probabilidad empírica de una variable continua. Una parte de este procedimiento es el sub-comando análisis de tallo y hojas. El procedimiento se ejecuta los comandos: ***Analizar, Estadísticas Descriptivas, Explorar, Lista dependiente, Lista de factor***, ***Ejecutar***

Ejemplo para explorar variables continuas según una variable categórica (análisis de tallo y hojas): Nicaragua 8
Ejemplo: Explorar la variable rendimiento de acuerdo a región. (usar: RENDI lista dependiente y REGION lista de factor)



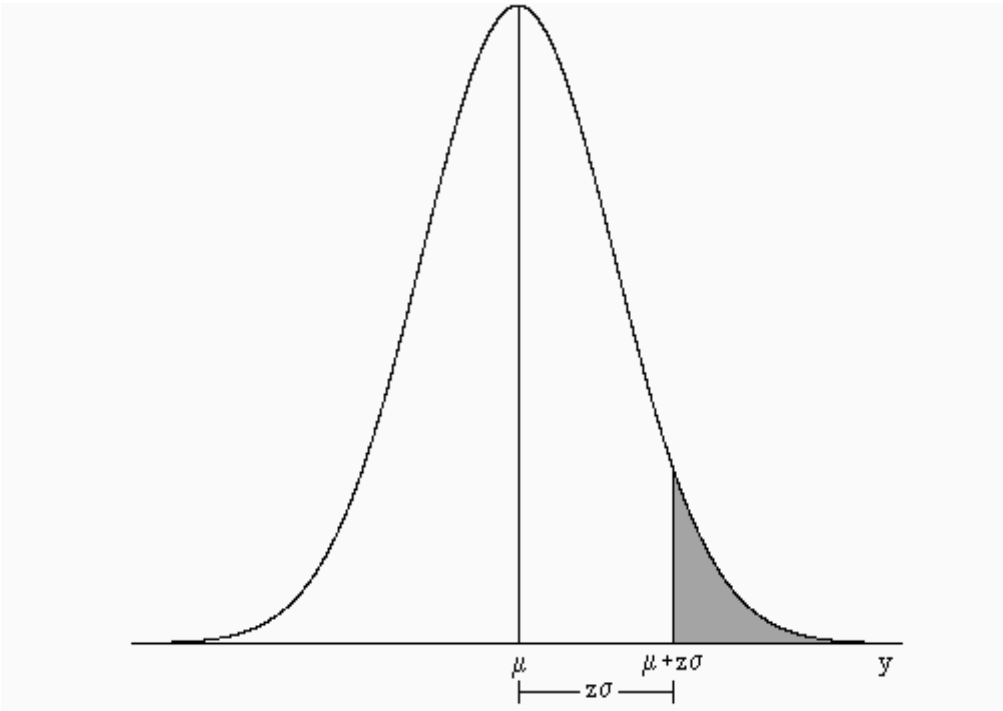
De acuerdo a las figuras, se puede ver que un análisis de tallo y hojas puede ser muy útil para identificar los casos de valores extremos.

4.3 **Inferencia Estadística**

El propósito de la inferencia estadística es inferir o averiguar las características de una población a partir de las características que se encuentran en muestras de la misma población. Este procedimiento se usa con variables continuas (no categóricas). La inferencia estadística está sujeta a error pues la información disponible corresponde sólo a una parte de la población.

4.3.1. Tipos de errores estadísticos

Son dos los errores estadísticos que se pueden producir cuando se realiza una inferencia estadística. El error de Tipo I se refiere a la probabilidad que existe de rechazar una hipótesis cierta. Esta probabilidad asociada es denominada nivel de significancia y se le denota por α. En el siguiente gráfico se describe el valor de α para una distribución normal con una media de población μ y una desviación estándar de σ.



El error de Tipo II, asociado con la denominada potencia de prueba, se refiere a la probabilidad de aceptar una hipótesis que es falsa. Dada la estructura de las pruebas de hipótesis estadísticas, el error de tipo II se asocia más con la probabilidad de fallar en rechazar una hipótesis falsa.

4.3.2 Intervalos de Confianza

Este estimador se define como la probabilidad de no cometer un Error de Tipo I en cuanto a la ubicación de la media. Su notación formal se define como:

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

Si se estandariza la distribución normal sobre la base de una media de cero, la formula se convierte en:

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

4.3.3 La Prueba "t"

Esta prueba permite evaluar la validez estadística de un estimador poblacional dentro de una muestra. Por ejemplo se puede pensar que la media de una distribución tiene un valor de "c". La prueba t se formula de tal manera que permita averiguar si la media muestral es estadísticamente diferente al valor "c". En términos formales la prueba se define como:

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

La distribución estadística simétrica que se usa para analizar esta prueba se denomina t de Student, la cual tiene características particulares:

- El 67% de las observaciones de su distribución se encuentran dentro de una desviación estándar de la media.
- El 95% de las observaciones de su distribución se encuentran dentro de dos desviaciones estándar de la media.
- La significancia es la probabilidad de que un valor de t sea mayor que el valor de la prueba t.
- En muestras grandes (mayor a 30), la distribución t se aproxima a la normal.

4.3.4 La covarianza

La covarianza permite conocer si dos variables varían conjuntamente. La covarianza poblacional entre las variables x_1 y x_2 se define como:

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

La covarianza muestral se define como:

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

4.3.5. Comparación de dos medias

Esta prueba permite formular la hipótesis sobre la igualdad de dos medias, lo que es lo mismo averiguar si la diferencia entre dos medias es cero. La prueba se define como:

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

En este caso la desviación estándar requerida para la prueba difiere de la desviación estándar de cualquiera de las otras distribuciones (es menor). La desviación estándar de una diferencia es la raíz cuadrada de la varianza de una diferencia, que es definida como:

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

Si las muestras son independientes, la covarianza es de cero. Esto hace que la varianza de una diferencia dependa de si los dos grupos o muestras comparten la misma varianza poblacional o si provienen de dos poblaciones distintas.

Los comandos para realizar estos tipos de pruebas son: ***Analizar, Comparar medias***, luego hay que seleccionar uno de los siguientes comandos ***Medias, Prueba t para una muestra, Prueba t para muestras independientes, Prueba t para muestras emparejadas*** y ***Prueba ANOVA***

Ejemplo para calcular medias muestrales: Nicaragua 9a
Calcular las medias del rendimiento de frijol que tienen tanto jefes femeninos como masculinos. Usar las variables RENDI Y JEFFEM

Informe

RENDI

Jefe Femenino	Media	N	Desv. típ.
0	11.60	2302	7.271
1	10.34	444	6.845
Total	11.40	2746	7.217

Ejemplo para una prueba t individual: Nicaragua 9b
¿Es el rendimiento medio significativamente diferente de 10 quintales por manzana?

Prueba para una muestra

	Valor de prueba = 10					
					95% Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
RENDI	10.134	2745	.000	1.40	1.13	1.67

Los resultados muestran que la hipótesis nula es rechazada. Hay evidencia estadística que el rendimiento del fríjol es diferente de 10 quintales por manzana.

Ejemplo para evaluar la diferencia entre dos medias: Nicaragua 9c
Diferencias entre los rendimientos de fríjol entre familias con jefe femenino y masculino

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	ig. (bilateral	Diferencia e medias	rror típ. de a diferencia	Inferior	Superior
RENDI	Se han asumido varianzas iguales	.063	.802	-3.361	2744	.001	-1.26	.373	-1.987	-.523
	No se han asum varianzas iguales			-3.502	650.869	.000	-1.26	.358	-1.959	-.551

Esta prueba rechaza la hipótesis nula que las medias son iguales. La diferencia de 1.26 quintales por manzana entre hogares con jefe femenino y masculino es significativamente diferente de cero, sin importar si se asume que ambas categorías de hogares comparten una varianza común

4.4 **Análisis de correlación**.

Permite conocer si dos variables independientes están correlacionadas, es decir si el comportamiento de una tiene relación con el comportamiento de una segunda variable. El coeficiente de correlación viene a ser como una covarianza proporcional. La correlación puede ser positiva o negativa y los valores van desde -1 a 1. La correlación entre las variables x_1 y x_2 se define como:

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

Para obtener la correlación entre dos variables continuas se usan los comandos: ***Analizar, Correlación, Bivariada, Seleccionar variables.***

Ejemplo de un análisis de correlación: Nicaragua 10
Ejemplo: Analizar la correlación entre las variables costo de mano de obra y costo de insumos (COSMOB Y COSINS)

Correlaciones

		COSMOB	COSINS
COSMOB	Correlación de Pearson	1	.157**
	Sig. (bilateral)	.	.000
	N	2746	2746
COSINS	Correlación de Pearson	.157**	1
	Sig. (bilateral)	.000	.
	N	2746	2746

**. La correlación es significativa al nivel 0,01 (bilateral).

En este caso existe una correlación positiva significativa entre ambas variables.

TERCER DIA

5. Regresión lineal múltiple por Mínimos Cuadrados Ordinarios (MCO)

En términos generales, por medio de la regresión lineal múltiple de MCO se trata de explicar el comportamiento de una variable, denominada dependiente o explicada, a través del comportamiento de otras variables, llamadas independientes o explicativas. Si y_i es la variable

independiente (endógena) y x_i son las variables independientes (exógenas), la forma lineal del modelo para la observación *i* es definida como:

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

Donde los β_i son coeficientes que serán estimados estadísticamente y u_i es un "error" al azar de adición que representa los aspectos de y_i que no pueden ser explicados estadísticamente con las variables $x_{ji.}$

5.1 Supuestos de la regresión por MCO

- La variable dependiente es continua
- Todos los errores u_i son independientes de las variables X_i
- Todos los errores u_i son independientes de los demás u_j pues $E(u_i u_j)=0$
- La Media de los u_i, $E(u_i)=0$
- $E(u_i^2)=\sigma^2$

5.2 Definición de la estructura de un modelo de regresión

Normalmente un modelo tiene una base teórica a través de la cuál se establece el efecto de causalidad de ciertas variables sobre una variable de interés. Un punto importante a considerar es cuando un modelo está incompleto. Si se omite una variable independiente que afecta la variable dependiente, puede haber un sesgo en los coeficientes de las demás variables si hay una correlación entre la variable omitida y las incluidas. Para evitar el sesgo se debe procurar incluir todas las variables que lógicamente pueden entrar en la relación modelada. Sin embargo esto dependerá de la disponibilidad de la información.

Ejemplo de especificación de modelo: Nicaragua 11
a) Definición del modelo
A partir del problema de maximización de ganancias, se deriva la función de demanda por insumos (demanda de productores por semilla mejorada, por tecnologías de suelos, etc) y la función de oferta de los productores (qué factores determinan el nivel de producción de un productor individual). Durante el taller se coincidió que según la teoría económica, la demanda por un insumo productivo depende de: el precio del producto final, el precio de los insumos de producción y otras variables adicionales que se pensó influirán tales como el costo de transporte y los recursos humanos financieros y de tierra de la finca. De manera similar, se derivó de la teoría económica, que la producción de un cultivo es determinada por las mismas variables independientes previas y tal vez alguna más que influencie la producción de un cultivo.

b) Especificación del modelo de regresión de MCO
El proceso de especificar un modelo de regresión es clave en un trabajo empírico. El desafío es operacionalizar el modelo derivado de la teoría económica. En el caso de la oferta de cultivos que tiene una finca, ésta está representada por los rendimientos que alcanza un productor en determinado cultivo. Para explicar el nivel de rendimiento se pueden utilizar diversas variables explicativas que directamente sugiere la teoría o que dan una idea aproximada del comportamiento de las mismas.
Por ejemplo, para incluir los precios de los insumos de producción se pueden incluir los valores de los costos (inversión) que cada productor incurre en la producción de un cultivo. Así los costos unitarios de mano de obra, de insumos y de otros servicios pueden explicar los niveles de rendimiento. Según la teoría se debería incluir el precio de mercado del cultivo como variable explicativa porque un cultivo que vale más justifica mayor inversión en insumos que aumentan el rendimiento. Pero esta condición requiere que el productor anticipe el precio que recibirá por su producto al momento de la cosecha. Normalmente este precio esperado está relacionado con los niveles pasados del precio del cultivo analizado y el conocimiento de los mismos por parte del productor. Es decir se debe incluir en el modelo el precio esperado que tiene un productor antes de iniciar la producción de un cultivo que muchas veces es una función del(os) precio(s) de la campaña(s) anterior(es).
Finalmente se deben incluir otras variables que por experiencia o trabajos similares se sepa influyen en los niveles de rendimiento de un cultivo. Aquí información sobre características socioeconómicas de la familia del productor, características agroecológicas de la finca y entornos de política deben ser incluidas.

5.3 Representatividad de la línea de regresión (bondad de ajuste)

La medida más usada para medir cuan representativa es la línea que se estima por MCO es el coeficiente de determinación o R^2. Este coeficiente mide el porcentaje de la variabilidad en los datos explicada por la regresión. El coeficiente de determinación puede ser definido como:

$$R^2 = 1 - \frac{SEC}{STC}$$

Donde SEC es la suma de errores cuadrados y STC es la suma total de cuadrados. Otra medida de la bondad de ajuste es la prueba estadística F referida al modelo de regresión entero.

5.4. La prueba F y su implementación

Hay dos formas útiles de la prueba F que se pueden usar analizar la base de datos. La primera mide el poder explicativo del modelo de regresión especificado. La prueba F es una razón. En el numerador se mide el cambio en la explicación agregada por la regresión completa. El denominador mide la variabilidad total de la regresión. En ambos casos es necesario considerar los grados de libertad. En el numerador, estos son iguales al número de variables usadas en el modelo completo (K) menos uno por la constante (K-1), mientras que en el denominador los grados de libertad igualan al número de observaciones menos el número de variables (n-K). De esta manera se puede definir la prueba F en términos del coeficiente de determinación como:

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

Si según la teoría no es obligatorio que ciertas variables disponibles se incluyan en un modelo de regresión, una prueba estadística puede ayudar a decidir si la(s) variable(s) contribuye(n) a explicar la variabilidad de la variable dependiente. Para evaluar la contribución de una sola variable se puede utilizar la prueba t de su respectivo coeficiente. Para evaluar más de una variable se necesita utilizar una segunda forma de la prueba F que compara la variabilidad

explicada por un modelo reducido (sin las variables excluidas) con la de un modelo entero (donde se consideran todas las variables originales).

$$F(J, n-K) = \frac{(ECS_{\text{sin}} - ECS_{con})/J}{(1-R^2)/(n-K)}$$

Si el estadístico F no es significativo respecto al valor umbral determinado (comúnmente α=5%) entonces se puede utilizar el modelo reducido sin perder mucho poder explicativo.

5.5 Expectativas de las variables independientes de la regresión

Antes de ejecutar en SPSS los comandos para la regresión, un buen investigador debe anticipar algunos efectos que por teoría o experiencia espera sobre el comportamiento de las variables. Por ejemplo es muy importante determinar qué efectos se esperan de las variables independientes sobre las dependientes, incluyendo los signos de estos efectos. En una regresión lineal con variables continuas, la interpretación del coeficiente de una variable independiente es el cambio en la variable dependiente si se agrega una unidad adicional a la variable independiente. La interpretación del coeficiente de una variable independiente binaria, es el cambio en la variable dependiente que ocurriría si la variable independiente fuera igual a uno (si). Por ejemplo si se busca la diferencia del efecto de tener jefes de familia femeninos o masculinos, y se define el valor de 1 para JEFFEM cuando el jefe de familia es, un coeficiente positivo implicará que tener un jefe femenino tiene un efecto mayor en la variable dependiente que el tener un jefe de familia masculino. Si el coeficiente es negativo, la diferencia será a favor de los jefes de familia masculinos.

5.6 El uso de MCO en el SPSS

El programa SPSS permite realizar regresiones de manera muy sencilla. Una vez especificado el modelo de regresión, sólo hay que definir dentro del programa cuál es la variable dependiente y cuáles son las variables independientes.

Los comandos para ejecutar la regresión son: ***Analizar, Regresión, Lineal, Definir variable dependiente, Definir variables independientes, Ejecutar***

Ejemplo de regresión de MCO: Nicaragua 12a

Explicando los rendimientos en función de PRECIO DE VENTA, ÁREA DE LA FINCA, ÁREA DE LA PARCELA DE FRÍJOL, COSTO DE MANO DE OBRA, COSTO DE INSUMOS, COSTO DE SERVICIOS, SEMILLA MEJORADA, JEFE FEMENINO, TENENCIA PROPIA, CLIENTES MASIVOS, CLIENTES ATP1, ÉPOCA DE POSTRERA y ÉPOCA DE APANTE

El R^2 y el cuadro ANOVA

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.320(a)	.102	.098	6.857

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	14611.027	13	1123.925	23.907	.000[a]
	Residual	128205.315	2727	47.013		
	Total	142816.343	2740			

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

b. Variable dependiente: RENDI

Ejemplo de Estimación de coeficientes usando MCO: Nicaragua 12b

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	6.189	.730		8.481	.000
	PREVENTA	-.002	.002	-.020	-1.057	.291
	AREA	.531	.088	.113	6.001	.000
	AREAFIN	.003	.003	.020	1.080	.280
	COSMOB	.001	.000	.067	3.613	.000
	COSERV	.003	.001	.102	5.539	.000
	COSINS	.002	.000	.089	4.748	.000
	semilla mejorada	1.393	.266	.096	5.234	.000
	Jefe Femenino	-.984	.360	-.050	-2.730	.006
	tenencia propia	.201	.378	.010	.532	.595
	clientes masivos	-1.686	.502	-.068	-3.356	.001
	Clientes atp1	1.824	.280	.126	6.513	.000
	POSTRERA	1.553	.305	.107	5.092	.000
	APANTE	2.234	.453	.098	4.937	.000

a. Variable dependiente: RENDI

Ejemplo de interpretación de resultados de regresión: Nicaragua 13
Los primeros resultados

Del primer cuadro el dato más importante es el coeficiente de determinación. En el modelo especificado las variables independientes incluidas explican el 10.2% de la variabilidad de los rendimientos de fríjol. Si bien este coeficiente no es muy alto, normalmente modelos de análisis de rendimientos agrícolas de corte transversal tienen como característica bajos coeficientes de determinación.

El segundo cuadro resume el nivel de significancia del modelo como un todo. La prueba conjunta de F con el valor de 23.907 está rechazando la hipótesis nula de que las variables explicativas no tienen ningún efecto sobre la variable dependiente.

El tercer cuadro contiene la información sobre la estimación de coeficientes. El primer paso es determinar cuáles variables tienen un efecto individual significativo sobre la variable dependiente y cuáles no. Los resultados muestran que no existe evidencia estadística de que las variables TPROPIA, ÁREAFIN Y PREVENTA expliquen el nivel de los rendimientos de fríjol. Para las demás variables las pruebas t individuales rechazan las hipótesis nulas de que estas variables no tienen un efecto significativo en los rendimientos de fríjol.

Ejemplo del uso de la prueba F: Nicaragua 14
¿Cómo afecta al modelo la eliminación de un grupo de variables con estadísticos t no significativos? Una prueba F puede explicar porque

En el modelo TPROPIA, ÁREAFIN Y PREVENTA son candidatas a ser eliminadas. Para implementar la prueba F se estimaron dos regresiones. En la primera se incluyeron las 13 variables originales y la segunda, es un modelo reducido donde las tres variables "insignificantes" fueron incluidas. Los R^2 de ambas regresiones son necesarios para calcular el F estadístico de la prueba. En la regresión completa el R^2 era de 0.102 y en la regresión restringida de 0.101. La prueba mide el efecto en el poder explicativo del modelo cuando se remueven estas 3 variables de un total de 13 variables originales. La hipótesis nula de la prueba indica que las tres variables a excluir no tienen un efecto significativo en conjunto sobre los rendimientos de fríjol. En este ejemplo el estadístico es definido como:

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

De acuerdo a este resultado se falla en rechazar la hipótesis de que los 3 coeficientes no tienen un efecto significativo en la variable dependiente, entonces las variables pueden ser eliminadas.

Resultados del modelo reducido: Nicaragua 15a
El R^2 y el cuadro ANOVA

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.319(a)	.101	.098	6.855

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	14508.818	10	1450.882	30.875	.000[a]
	Residual	128477.868	2734	46.993		
	Total	142986.686	2744			

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

b. Variable dependiente: RENDI

Resultados del modelo reducido: Nicaragua 15b

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5.784	.413		14.005	.000
	AREA	.543	.088	.115	6.197	.000
	COSMOB	.001	.000	.068	3.690	.000
	COSERV	.003	.001	.102	5.538	.000
	COSINS	.002	.000	.088	4.682	.000
	semilla mejorada	1.415	.266	.098	5.326	.000
	Jefe Femenino	-1.008	.359	-.051	-2.806	.005
	clientes masivos	-1.629	.500	-.065	-3.258	.001
	Clientes atp1	1.843	.279	.128	6.606	.000
	POSTRERA	1.642	.297	.113	5.538	.000
	APANTE	2.361	.441	.104	5.352	.000

a. Variable dependiente: RENDI

Ejemplo de interpretación de coeficientes MCO: Nicaragua 16

Los coeficientes de la tabla 15b tienen una interpretación especial en términos de cambios en la variable dependiente.

La constante del modelo nos indica que un agricultor que no usa semilla mejorada de fríjol, que no invierte en mano de obra, servicios e insumos, que no recibe asistencia técnica del proyecto y que siembra el fríjol en la campaña primera, tiene en promedio un rendimiento de fríjol de 5.784 quintales por manzana. (1 manzana = 0.7 hectáreas)

Para ilustrar el caso de la interpretación de una variable continua, tomamos el caso de la variable AREA. Según nuestra regresión, si un agricultor aumenta el área de fríjol de su finca en 1 manzana, este agricultor puede esperar que los rendimientos de dicho cultivo aumenten en 0.543 quintales por manzanas. Es decir conforme mayor área le dedique un productor al fríjol, más productivo será en este cultivo.

La interpretación del coeficiente de una variable binaria difiere un poco. Tomando el caso de la variable ATP1, su coeficiente estimado nos indica que en promedio un agricultor que recibe asistencia técnica privada obtiene rendimientos de fríjol mayores en 1.843 quintales por manzanas al que recibe un agricultor sin ningún tipo de asistencia técnica, manteniendo todas las demás variables constantes (usan la misma semilla mejorada, siembran en la misma campaña agrícola, tienen la misma área de fríjol, e invierten la misma cantidad de dinero en mano de obra, servicios e insumos).

5.7 Especificando otra forma funcional

Muchas veces un modelo de regresión lineal no representa de la mejor manera la relación entre una variable dependiente y las variables explicativas. En este caso es necesario evaluar otras formas funcionales que permitan representar mejor la relación causa y efecto entre una variable independiente y una variable dependiente. Adicionalmente a la regresión lineal usada, los modelos cuadráticos y los modelos logarítmicos son los preferidos.

Ejemplo de cómo especificar otra forma funcional: Nicaragua 17
Dada la información disponible se puede especificar un modelo de regresión cuadrático. Podrían haber dos hipótesis acerca de cómo el tamaño del área dedicada a la producción de fríjol influencia el nivel de rendimientos. Una hipótesis sostiene que a mayor el área de fríjol en producción, mayor el rendimiento del cultivo, por la presencia de economías de escala. La hipótesis opuesta sostiene que por el contrario, a menor área, mayor probabilidad de tener una mayor producción por un manejo más intensivo. Al adicionar un término cuadrático para el área de fríjol sembrada al modelo original, se puede probar la hipótesis de que los rendimientos pueden incrementarse al crecer el área hasta cierto punto y después declinar.

Ejemplo de estimación de un modelo cuadrático con MCO: Nicaragua 18a
Incluyendo el Área como término cuadrático. El R^2 y el cuadro ANOVA

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.333(a)	.111	.107	6.821

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	15845.861	11	1440.533	30.965	.000[a]
	Residual	127140.825	2733	46.521		
	Total	142986.686	2744			

a. Variables predictoras: (Constante), AREA2, clientes masivos, semilla mejorada, COSMOB, Jefe Femenino, APANTE, COSERV, COSINS, Clientes atp1, POSTRERA, AREA

b. Variable dependiente: RENDI

Ejemplo de estimación de un modelo cuadrático usando MCO: Nicaragua 18b
Incluyendo el Área como término cuadrático

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5.159	.427		12.078	.000
	AREA	1.281	.163	.272	7.862	.000
	COSMOB	.001	.000	.066	3.601	.000
	COSERV	.003	.001	.098	5.347	.000
	COSINS	.002	.000	.088	4.727	.000
	semilla mejorada	1.335	.265	.092	5.043	.000
	Jefe Femenino	-.971	.358	-.050	-2.715	.007
	clientes masivos	-1.718	.498	-.069	-3.451	.001
	Clientes atp1	1.735	.278	.120	6.237	.000
	POSTRERA	1.544	.296	.107	5.222	.000
	APANTE	2.235	.440	.099	5.084	.000
	AREA2	-.068	.013	-.184	-5.361	.000

a. Variable dependiente: RENDI

Los resultados de este modelo muestran un efecto significativo del término cuadrático. Al combinar los signos de la variable ÁREA y de ÁREA CUADRADA, se ve que el efecto del área de fríjol en el rendimiento de este cultivo es un incremento pero a una tasa decreciente. A través del cálculo diferencial podemos calcular que el incremento de área aumenta los rendimientos de fríjol hasta un área de alrededor de 9 manzanas. Luego, una mayor área de fríjol disminuirá los rendimientos promedios. Adicionalmente se aprecia que este modelo de regresión explica una mayor variabilidad de los rendimientos de fríjol por el alza de su coeficiente de determinación (ver el R^2).

6. Regresión de variables dependientes binarias

Los modelos de regresión binaria tienen como variable dependiente una variable binaria que toma el valor de "1" o "0". Un uso importante de tales modelos es el de "explicar" los factores determinantes de la adopción (o no) de una nueva tecnología. La base teórica de estos modelos se encuentra en la teoría de "utilidad aleatoria". Este concepto supone que existe un nivel de satisfacción (utilidad) que un consumidor alcanza cuando consume algo.

Este consumo puede referirse también al consumo de insumos productivos, por ejemplo, la adopción de una nueva tecnología. Sin embargo, la utilidad verdadera no puede ser observada. Solamente se puede observar la elección de un consumidor, es decir si compra o no un bien o servicio. Esta elección puede ser la adopción (valor "1") o no (valor "0") de una tecnología.

Puesto que la regresión binaria tiene una variable dependiente que toma valores de 0 y 1, la regresión estimada predice la probabilidad de que la variable dependiente valga 1 (la probabilidad de que un productor adopte una tecnología). Los coeficientes de las variables independientes indican cómo influyen éstas en la probabilidad de que la variable dependiente ocurra (Ej. la probabilidad de adoptar una tecnología).

Los modelos binarios son útiles para explicar procesos de adopción de tecnología cuando no se cuentan con variables continuas para el análisis. Existen varios modelos de variables dependientes binarias, pero uno de los más usados es la regresión Logit.

6.1 Regresión LOGIT

En este caso la probabilidad de que la variable dependiente *y* (condicionada por las variables independientes x_i) sea 1 se puede representar de la siguiente forma logística:

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

El logaritmo de la razón de probabilidades (P/(1-P)) (llamado un “logit”) tiene una forma funcional lineal:

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

La interpretación de los coeficientes de la regresión estimada difiere de la tradicional interpretación de MCO. Al igual que en el caso de MCO, la dirección del efecto sigue el signo del coeficiente estimado. Sin embargo, el efecto marginal varia con la magnitud de las variables independientes.

Ejemplo de un modelo LOGIT: Nicaragua 19

A partir de la teoría económica se sabe que la demanda por un insumo productivo depende del precio del cultivo en el mercado, los precios de los insumos de producción y otras variables que afecten los incentivos y la capacidad de usar el insumo. Sin embargo en el caso de ciertos insumos, sólo se puede observar si un productor adopta o no el insumo. Este ejemplo la variable dependiente es la adopción de semillas mejoradas. Este proceso de adopción es explicado por las variables precio de venta, área de la finca, área de la parcela de fríjol, costo de mano de obra, costo de insumos, costo de servicios, semilla mejorada, jefe femenino, tenencia propia, clientes masivos, clientes atp1, época de postrera y época de apante.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3710.080	.024	.033

Classification Table [a]

Observed			Predicted		
			semilla mejorada		Percentage Correct
			0	1	
Step 1	semilla mejorada	0	477	770	38.3
		1	380	1114	74.6
	Overall Percentage				58.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	COSERV	.001	.000	8.750	1	.003	1.001
	COSMOB	.000	.000	.259	1	.611	1.000
	COSINS	.000	.000	4.670	1	.031	1.000
	AREAFIN	-.001	.001	.578	1	.447	.999
	AREA	.015	.027	.317	1	.574	1.015
	PREVENTA	-.001	.001	2.507	1	.113	.999
	JEFEFEM	-.254	.106	5.703	1	.017	.776
	ATP1	.495	.083	35.924	1	.000	1.640
	ATPMA	.371	.149	6.220	1	.013	1.449
	TPROPIA	.122	.112	1.199	1	.274	1.130
	POSTRERA	.167	.091	3.409	1	.065	1.182
	APANTE	.161	.134	1.442	1	.230	1.175
	Constant	-.307	.216	2.022	1	.155	.735

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

El modelo tiene un coeficiente de determinación muy bajo y su poder de predicción es pobre. El modelo predice correctamente el resultado de adopción de semilla mejorada sólo el 58% del tiempo. Predice correctamente adopción el 74.6% de los casos y no adopción

CUARTO Y QUINTO DÍAS

Estos días son dedicados al desarrollo de proyectos grupales. El tema de cada proyecto debería ser propuesto por los mismos participantes del taller, teniendo en cuenta sus intereses personales e institucionales. Algunos ejemplos de proyectos son:

1. Efecto de la asistencia técnica sobre la adopción de nuevas variedades de cultivos,
2. Factores que influyen en la adopción de tecnologías específicas,
3. Efecto de de la adopción de tecnologías sobre los rendimientos de cultivos.

Al terminar los análisis grupales, se recomienda unas presentaciónes públicas al taller para poder discutir los métodos estadísticos usados y los resultados sacados.

Referencias útiles

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, Chris (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