UNIVERSIDAD DE CHILE
MAGISTER EN GESTION Y POLITICAS PUBLICAS

ESTADISTICA APLICADA Y ECONOMETRIA

Sara Arancibia C

2012

1

Objetivos

- Comprender y aplicar los conceptos básicos de Econometría y metodologías de Análisis Multivariante, fundamentales para el análisis de información.
- Conocer y manejar el software estadístico SPSS, con énfasis en la resolución de estudios de casos aplicados a la gestión y políticas públicas.

Metodología

- Clases teóricas y prácticas.
- Apoyo de material; transparencias, guías, lecturas complementarias
- Manejo del software SPSS
- En los laboratorios se realizarán estudios de casos apoyados de guías.

2

Evaluación

Tareas semanales (30%), un control (30%), examen (40%)

Bibliografía:

- Introducción a la Econometría. Un enfoque moderno.
Jeffrey y Wooldridge . Ed Thomson Learning
- Econometría. Cuarta Edición
Gujarati Ed. Mc Graw Hill
- Análisis Multivariable para las Ciencias Sociales.
Lévy y Varela Ed Pearson
- Análisis multivariante
Hair-Anderson-Tatham-Black. Ed Prentice Hall.
- Análisis de datos con SPSS 13 Base
Pardo y Ruiz. Ed Mc Graw Hill.
- Análisis Estadístico con SPSS para windows. Estadística Multivariante.
Visauta y Martori. Ed Mc Graw Hill. Segunda Edición

3

Contenidos

Primera sesión

- Introducción a la Econometría
- Introducción al Análisis Multivariable
- Análisis de varianza
- Análisis no paramétrico de H de Kruskall-Wallis
- Guía 1-Estudios de casos

Segunda y Tercera Sesión

- Análisis de regresión lineal simple
- Modelos lin-log y log-lin y semilogaritmicos
- Guía 2- Estudios de casos

Cuarta y Quinta Sesión

- Análisis de regresión múltiple
- Guía 3- Estudios de casos

4

Contenidos

Sexta Sesión
- Modelos de regresión múltiple con variables cualitativas (dami)
- Estimación ponderada
- Guía 4- Estudios de casos

Séptima Sesión
- Regresión logística
- Guía 5- Estudios de casos

Octava Sesión
- Análisis Factorial
- Guía 6- Estudios de casos

5

Introducción

- **Introducción a la Econometría**
- **Introducción al Análisis Multivariante**

6

Introducción a la Econometría

Naturaleza de la Econometría y de los datos económicos

- ¿Qué es la Econometría?
- Funciones de la Econometría
- La metodología de la Econometría
- La regresión es una herramienta fundamental de la Econometría.
- Estructura de los datos económicos
- Relaciones estadísticas vs. Relaciones determinísticas
- Regresión vs. Causalidad
- Regresión vs. Correlación
- Terminología

7

Introducción

Naturaleza de la econometría y de los datos económicos

¿Qué es la econometría?

Literalmente, econometría significa "medición económica".

La econometría se basa en métodos estadísticos para estimar las relaciones económicas, poner a prueba teorías económicas y evaluar y poner en práctica políticas gubernamentales y comerciales.

Aplicaciones de la econometría

Pronóstico de variables macroeconómicas (inflación, el producto interno bruto…)

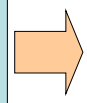
Estudios aplicados a diversos campos de la economía (Ej: estudio de los efectos de los gastos de las campañas políticas en los resultados de las votaciones, en el efecto de los gastos en educación en el rendimiento de los estudiantes, etc)

8

¿Cuáles son las funciones de la econometría?

La econometría tiene básicamente tres funciones estrechamente interrelacionadas.

1) Probar teorías económicas o hipótesis. Por ejemplo, ¿está el consumo directamente relacionado con el ingreso?, ¿está la cantidad demandada de un artículo inversamente relacionada con su precio?.
2) Dar estimaciones numéricas de los coeficientes de las relaciones económicas. Estos son esenciales en la toma de decisiones. Por ejemplo, un asesor gubernamental necesita tener una estimación exacta del coeficiente de la relación entre consumo e ingreso con el fin de determinar el efecto estimulante de una reducción de impuestos propuesta.
3) La predicción de sucesos económicos



La Econometría da contenido empírico a gran parte de la teoría económica

9

La metodología de la Econometría

En términos generales, el análisis econométrico sigue las siguientes líneas generales de acción:
1. Enunciado de la teoría o hipótesis
2. Especificación del modelo econométrico dirigido a probar la teoría
3. Estimación de los parámetros del modelo
4. Verificación o inferencia estadística
5. Predicciones o pronósticos
6. Utilización del modelo para fines de control o formulación de políticas

10

Ejemplo

Consideremos a continuación la teoría keynesiana de la función consumo

Enunciado de la teoría o hipótesis

Keynes plantea

La ley sicológica fundamental consiste en que los hombres están dispuestos, por regla general y en promedio, a aumentar su consumo a medida que aumenta su ingreso, aunque no en la misma proporción al incremento en dicho ingreso.

Keynes afirma que la propensión marginal a consumir (PMC), la tasa de cambio del consumo ante un cambio de una unidad en el ingreso, es mayor que cero pero menor que uno.

11

Especificación del modelo econométrico

Para simplificar, un economista matemático puede sugerir la siguiente forma para la función de consumo de Keynes:

$$Y = \beta_0 + \beta_1 X \qquad 0 < \beta_1 < 1 \qquad (1)$$

en donde

$Y = gastos\ de\ consumo$

$X = ingreso$

$\beta_0 = \text{intersección con el eje } Y$

$\beta_1 = pendiente$

El coeficiente de la pendiente $\beta 1$ representa la propensión marginal a consumir (PMC)

12

La ecuación (1), que afirma que el consumo está relacionado linealmente con el ingreso, es un ejemplo de un modelo matemático.
Si el modelo, como del ejemplo anterior, consta de una sola ecuación, recibe el nombre de modelo uniecuacional; si tiene más de una ecuación, se denomina modelo multiecuacional o modelo de ecuaciones simultáneas.
El modelo matemático de la función de consumo (1) es de limitado interés para el econometrista, por cuanto supone una relación exacta o determinística entre el consumo y el ingreso. Sin embargo, las relaciones existentes entre las variables económicas son generalmente inexactas

13

Para tener en cuenta la existencia de una relación inexacta entre las variables económicas, el econometrista debe modificar la función de consumo determinística de (1), de la siguiente manera

$$Y = \beta_o + \beta_1 X + u \qquad (2)$$

En la que *u* representa el término de perturbación o de error, que es una variable aleatoria (estocástica) con propiedades probabilísticas bien definidas.
El término perturbación, u, suele representar todas aquellas fuerzas que afectan el consumo pero que no se tienen en cuenta de manera explícita en la ecuación

14

Estimación

Habiendo especificado el modelo econométrico, la tarea siguiente del econometrista consiste en obtener estimaciones (valores numéricos) de los parámetros del modelo, a partir de la información disponible, generalmente proporcionada por el estadístico económico. Estas estimaciones le confieren un contenido empírico a la teoría económica. Así por ejemplo, si en el estudio de la función de consumo anteriormente expuesta, se encuentra que $\beta 1$ 0,8, este valor no sólo proporciona una estimación numérica de la PMC sino que corrobora la hipótesis keynesiana según la cual la PMC es menor que 1.

¿Cómo se estiman los parámetros?

La técnica utilizada para obtener dichas estimaciones es el análisis de regresión

15

Verificación (inferencia estadística)

Habiendo obtenido ya estimaciones de los parámetros, la tarea siguiente consiste en desarrollar criterios apropiados dirigidos a establecer si las estimaciones obtenidas están de acuerdo con lo que se espera de la teoría que se está verificando.

La refutación o confirmación de las teorías económicas, basándose en la evidencia empírica, se fundamenta en la inferencia estadística (prueba de hipótesis)

Predicciones o pronósticos

Si el modelo escogido confirma la hipótesis o teoría que se está investigando, se puede entonces proceder a predecir el (los) valor(es) futuro(s) de la variable dependiente Y con base en valores futuros, conocidos o esperados, para la(s) variable(s) explicativa(s) X.

16

Utilización de los modelos para fines de control o formulación de políticas

Supóngase que un economista del gobierno estima la función de consumo keynesiana, obteniendo los siguientes resultados
Y= 5+0,7X
donde el gasto de consumo Y y el ingreso X se miden en miles de millones de dólares. Adicionalmente se supone que el gobierno cree que un nivel de gastos de 1060 (miles de millones de dólares) mantendrá la tasa de desempleo a un nivel relativamente bajo, del orden del 5%. ¿Qué nivel de ingresos (X) garantizará que se obtenga la cantidad presupuestada inicialmente de gastos de consumo?.
Suponiendo que el modelo es aceptable, se tiene que:
1060=5+0,7X o X=1055/ 0,7 =1507

Lo anterior implica que un nivel de ingresos de 1507 (miles de mill de dólares), dada una PMC=0,7, generará un gasto de 1060 (miles de millones de dólares)

17

La regresión es una herramienta fundamental de la econometría.

Interpretación moderna de la regresión

El análisis de regresión está relacionado con el estudio de la dependencia de una variable, la variable dependiente, de una o más variables adicionales, las variables explicativas con la perspectiva de estimar y/ o predecir el valor (poblacional) medio o promedio de la primera en términos de valores conocidos o fijos (en muestreos repetidos) de las segundas.

Debe tenerse siempre en mente que el éxito del análisis de regresión depende de la disponibilidad de información adecuada.

18

Estructura de los datos económicos

Las estructuras de datos más comunes en la econometría aplicada son las de los datos de corte transversal, de series de tiempo, de combinación de cortes transversales, y de panel.

Datos de corte transversal

Un conjunto de datos de corte transversal consta de una muestra de individuos, hogares, empresas, ciudades, estados, países u otras diversas unidades, tomada en un momento determinado. A veces, los datos de todas las unidades no corresponden con exactitud al mismo periodo; por ejemplo, es posible entrevistar a varias familias durante semanas distintas del año. En un análisis de sección cruzada pura, ignoraríamos cualquier diferencia de tiempo mínima en la recopilación de los datos. Si se entrevistó a un grupo de familias en semanas distintas del mismo año, aún veríamos esta información como un conjunto de datos de corte transversal.

19

Tabla 1.1

Conjunto de datos de corte transversal sobre salario y otras características individuales

Obs	*sala*	*educ*	*exper*	*sexo*	*ecivil*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

20

Tabla 1.2

Conjunto de datos sobre las tasas de crecimiento económico y características de los países

obs	país	tpib	Consgob60	Secund60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Bélgica	2.56	13	69
4	Bolivia	1.24	18	12
..	.	.	.	
..	.	.	.	
..	.	.	.	
61	Zimbabwe	2.30	17	6

21

Datos de series de tiempo

Un conjunto de datos de series de tiempo (o datos de series temporales) consta de observaciones, de una o más variables, hechas en el tiempo.
Entre los ejemplos de este tipo de información se encuentran los precios de las acciones, el índice de precios al consumidor, el producto interno bruto, los índices anuales de homicidios y las cifras de venta de automóviles. Como los hechos del pasado pueden tener influencia en los del futuro y los rezagos en el comportamiento son comunes en las ciencias sociales, el tiempo es un factor importante en los datos de series de tiempo. A diferencia del ordenamiento de los datos de corte transversal, la disposición cronológica de las observaciones en una serie temporal proporciona información potencialmente importante.

22

Tabla 1.3

Salario mínimo, desempleo y datos relacionados para Puerto Rico

obs	*año*	*salamin*	*cob*	*desem*	*pib*
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

23

Combinación de cortes transversales

Algunos conjuntos de datos tienen características tanto de corte transversal como de series temporales.
Por ejemplo, supongamos que se realizaron a escala nacional dos encuestas transversales de hogares, una en 1985 y otra en 1990. En 1985 se entrevistó a una muestra aleatoria de hogares sobre variables como ingreso, ahorro, tamaño de la familia, etc. En 1990 se realizó una nueva muestra aleatoria con las mismas preguntas. Con el objeto de aumentar el tamaño de nuestra muestra, podemos formar una combinación de cortes transversales para los dos años. Como las muestras aleatorias se tomaron cada año, sería mera casualidad que el mismo hogar apareciera en la muestra de ambos años. (Por lo regular, el tamaño de la muestra será muy pequeño, en comparación con el de todos los hogares del país). Este importante factor distingue a la combinación de cortes transversales de los conjuntos de datos de panel.

24

Tabla 1.4

Combinaciones de cortes transversales: dos años de precios de la vivienda

obs	año	precio	imptos	piecuad	habit	Baños
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

25

Datos de panel o longitudinales

Un conjunto de datos de panel (o longitudinales) consta de una serie temporal para cada miembro del corte transversal en el conjunto de datos. Como ejemplo, supongamos que tenemos salario, educación y antecedentes de empleo de un grupo de individuos a los que se ha dado seguimiento durante 10 años; o también podríamos reunir información, como datos financieros y de inversiones, sobre el mismo conjunto de empresas durante un periodo de cinco años. De igual forma es posible recopilar datos de panel en unidades geográficas. Por ejemplo, podemos reunir datos de los mismos municipios de un país sobre flujos de migración, tasas impositivas, niveles de salarios, gastos gubernamentales, etc., para los años 1980, 1985 Y 1990.

La característica fundamental de los datos de panel, que los distinguen de las combinaciones de cortes transversales, es el hecho de que se da seguimiento a las mismas unidades

26

Tabla 1.5

Conjunto de datos de panel de dos años sobre estadísticas de delincuencia urbana

obs	ciudad	año	homicidios	población	desem	Policía
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

27

Nota: Los conjuntos de datos que incluyen la dimensión del tiempo, como los de series temporales y de panel, exigen un tratamiento especial por la correlación con el paso del tiempo de la mayor parte de las series de tiempo económicas. Otros temas, como las tendencias y la estacionalidad, surgen en el análisis de los datos de series temporales, pero no en los de corte transversal.

28

Relaciones estadísticas vs. Relaciones determinísticas

- En el análisis de regresión nos interesa lo que se conoce como dependencia estadística entre variables, pero no la funcional o determinística propia de la física clásica.

- En las relaciones estadísticas entre variables tratamos esencialmente con variables aleatorias o estocásticas, esto es variables que tienen distribuciones de probabilidad.

- Por otra parte, en la dependencia funcional o determinística también manejamos variables, pero éstas no son aleatorias o estocásticas.

NOTA: La palabra *estocástica* viene de la palabra griega *stokhos* que significa "centro del blanco". El resultado de lanzar dardos sobre un tablero es un proceso estocástico, esto es, un proceso que permite errores.

29

Relaciones estadísticas vs. Relaciones determinísticas

La dependencia del producto de una cosecha respecto a la temperatura ambiente, la lluvia, el sol y los fertilizantes, por ejemplo, es de naturaleza estadística en el sentido que las variables explicativas, si bien son importantes, no permitirán al agrónomo predecir en forma exacta el producto de la cosecha debido a los errores involucrados en la medición de estas variables y en razón de otra serie de factores (variables), que afectan colectivamente la producción pero pueden ser difíciles de identificar individualmente. De esta manera habrá alguna variabilidad "intrínseca" o aleatoria en la variable dependiente, produc-to de la cosecha, que no puede ser explicada en su totalidad sin importar cuán-tas otras variables explicativas consideremos.

30

Regresión vs Causalidad

Si bien el análisis de regresión tiene que ver con la dependencia de una variable respecto a otras variables, esto no implica causalidad necesariamente.
En palabras de Kendall y Stuart: "Una relación estadística, sin importar qué tan fuerte y sugestiva sea, nunca podrá establecer una conexión causal: nuestras ideas de causalidad deben venir de estadísticas externas y, en último término, de una u otra teoría."

Por ejemplo si consideramos el producto de una cosecha, no hay una *razón estadística* para suponer que la lluvia no depende del producto de la cosecha. El hecho de que se trata el producto de la cosecha como dependiente de la lluvia (entre otras cosas) es debido a consideraciones no estadísticas: el sentido común sugiere que la relación no puede revertirse, ya que no podemos contro-lar la lluvia modificando la producción de la cosecha.

31

Regresión vs Correlación

- El análisis de correlación está estrechamente relacionado con el de regresión aunque conceptualmente los dos son muy diferentes.

- En el análisis de correlación el objetivo principal es medir la *fuerza* o el *grado de asociación lineal* entre dos variables.
El coeficiente de correlación, mide esta fuerza de asociación (lineal).

Por ejemplo, se puede estar interesado en encontrar la correlación (el coeficiente) entre el hábito de fumar y el cáncer del pulmón; entre las calificaciones obtenidas en exámenes de estadística y las obtenidas en exámenes de matemáticas; entre las altas calificaciones obtenidas en la escuela secundaria y en la universidad, y así sucesivamente.

32

Regresión vs Correlación

En el análisis de regresión, como ya se mencionó, no estamos interesados en ese tipo de medición.
En cambio, se trata de estimar o de predecir el valor promedio de una variable sobre la base de valores fijos de otras variables. Así, quizás se desee saber si se puede predecir el promedio de las calificaciones en un examen de estadística, conociendo la calificación de un estudiante en un examen de matemáticas.

33

Terminología

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_n X_{ni} + \mu_i$$

En la teoría económica los términos *variable dependiente* y *variable independiente* están descritos de varias maneras; a continuación se presenta una lista representativa de ellas:

y	$X_1, X_2, X_3, \ldots X_k$
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variables de control
Variable predicha	Variables predictora
Regresada	Regresora

34

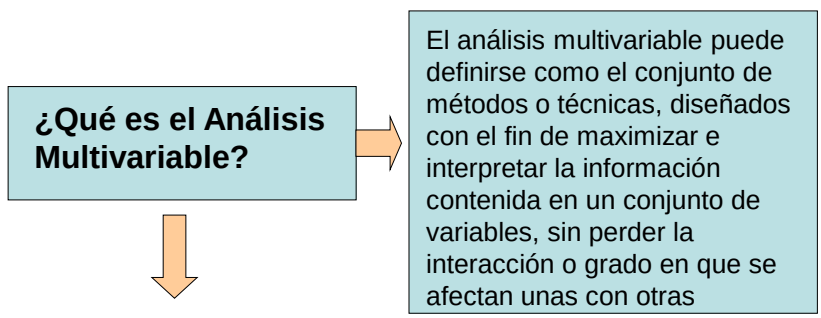
Introducción al Análisis Multivariante

Conceptos y técnicas del Análisis Multivariable

- ¿Qué es el Análisis Multivariable?
- Utilidad del Análisis Multivariable
- Los datos en el Análisis Multivariable
  - Variables y escalas de medida
  - Análisis inicial de datos
- Las técnicas de Análisis Multivariable
  - Técnicas de análisis de la dependencia
  - Técnicas de análisis de la interdependencia

35

¿Qué es el Análisis Multivariable?



El análisis multivariable permite llevar a cabo la resolución de problemas y la toma de decisiones con un enfoque analítico sobre todas las variables que llegan a influir sobre el o los problemas en cuestión.

36

Utilidad del Análisis Multivariable

La complejidad de la realidad socioeconómico-empresarial y el hecho de que en su conocimiento confluyan disciplinas científicas de origen diverso hacen que el contenido de los métodos multivariables se proyecte como un cuerpo de conocimientos de naturaleza interdisciplinaria.

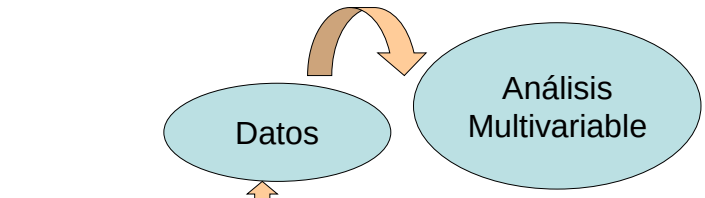
Las necesidades de información de los investigadores y decidores para la planificación, ejecución de acciones o el control de resultados son cada vez mayores.

En el análisis multivariable, se puede encontrar una herramienta práctica, versátil y adaptable a todo tipo de análisis, al permitir extraer información relevante, y eficiente.

37

Los datos en el Análisis Multivariable

Variables y escalas de medida



Valores que toman las **Variables**

Magnitudes que representan distintos conceptos o atributos de individuos u objetos

La precisión de tal representación dependerá directamente de la escala de medida

38

Los datos en el Análisis Multivariable
Escalas de medida

La tipología de escalas de medida distingue cuatro básicas

- Nominal
- Ordinal

} Escalas no métricas o cualitativas

- Intervalo
- Razón

} Escalas métricas o cuantitativas

39

Los datos en el Análisis Multivariable
Escalas de medida

Una variable no métrica puede ser convertida en variable ficticias binarias (dummy). Sería necesario contar con un número de ellas igual al número de categorías de la variable no métrica menos uno.

Ejemplo:
Supóngase que se pretende transformar la variable "medios de transporte más comunes" de tres categorías: 1=autobús, 2=tren y 3=avión.
La conversión podría efectuarse por medio de dos variables ficticias, F1 y F2. Los valores que éstas tomarían para representar cada categoría serían los siguientes:

Categoría	F1	F2
Autobús	1	0
Tren	0	1
Avión	0	0

40

Los datos en el Análisis Multivariable

Análisis inicial de datos

Antes de comenzar con el análisis multivariable, es esencial realizar un examen exhaustivo de los datos.

La detección de problemas ocultos en las matrices de datos supondrá un gran avance en la consecución de resultados lógicos y consistentes.

Es fundamental inspeccionar:

Análisis de datos ausentes (missing values)	▪Analizar si es relevante para el análisis obtener los datos perdidos. ▪Determinar si la información que falta puede ser completada. ▪Sustituir los datos por valores estimados

41

Los datos en el Análisis Multivariable

Análisis inicial de datos

Es fundamental inspeccionar:

Representaciones gráficas para el análisis de datos	▪Histogramas de cada variable ▪Gráficos de dispersión ▪Gráfico de cajas (Boxplot)
Tablas	▪Tablas de frecuencia ▪Tablas de contingencia
Detección de outliers	▪Estudiar los casos atípicos

42

Los datos en el Análisis Multivariable

Análisis inicial de datos

Es fundamental inspeccionar:

Supuestos subyacentes en los métodos multivariables

- Normalidad de las variables
- Linealidad (existencia de asociaciones lineales entre variables)
- Homocedasticidad (Varianza de los errores es constante)

43

Las técnicas del Análisis Multivariable

Tipología de las técnicas

- De análisis de la dependencia
- De análisis de la interdependencia
- Otras técnicas

44

Las técnicas del Análisis Multivariable

- De análisis de la dependencia → Técnicas aplicables cuando una o varias variables dependientes van a ser explicadas por un conjunto de variables independientes que actúan como predictoras

- De análisis de la interdependencia → Técnicas que otorgan la misma consideración a todas las variables objeto de estudio, sin distinguir entre dependientes e independientes, y que tienen como fin descubrir las interrelaciones entre ellas. Son técnicas de clasificación.

- Otras técnicas → Técnicas novedosas que permiten un tratamiento más eficaz y eficiente en grandes cantidades de datos, como análisis con redes neuronales, data mining.

45

Técnicas de análisis de la dependencia

Técnica	Variable dependiente	Variables independientes
Análisis de la varianza y la covarianza	Métrica	No métricas
Análisis discriminante	No métrica	Métricas
Regresión lineal múltiple ídem con variables ficticias	Métrica Métrica	Métricas No métricas
Modelos de elección discreta ídem con variables ficticias	No métrica No métrica	Métricas No métricas
Análisis conjunto	Métrica o no métrica	No métricas
Segmentación Jerárquica	No métrica o métrica	No métricas
Análisis de ecuaciones estructurales	Métrica	Métricas o no métricas
Análisis con clases latentes	No métrica latente	No métricas observables

46

Terminología

Variable ficticia
Variable binaria que se suele emplear para representar una categoría de una variable no métrica.

***Variable métrica* (o cuantitativa)**
Variable medida en escala de intervalo o de razón, capaz de reflejar, por tanto, diferencias de grado o cantidad entre sus elementos. La diferencia entre dos elementos consecutivos es constante a lo largo de toda la escala.

***Variable no métrica* (o cualitativa)**
Variable medida en escala nominal u ordinal que identifica categorías o propiedades. Si es ordinal, los números asignados a cada categoría guardan una relación de orden; pero, por lo demás, son simples etiquetas sin ningún otro significado.

47

ANOVA (o análisis de la varianza)

Método para contrastar si diversas muestras proceden de poblaciones con igual media.

ANCOVA (o análisis de la covarianza)

Proceso que comienza por emplear la regresión para eliminar la variación experimentada por la variable dependiente producida por una variable independiente no controlada (covariable) cuyos efectos se consideran indeseados, y sigue con un ANOVA sobre la variable dependiente ajustada.

48

Análisis discriminante

Técnica de clasificación que permite agrupar a los elementos de una muestra en dos o más categorías diferentes, predefinidas en una variable dependiente no métrica, en función de una serie de variables independientes métricas combinadas linealmente.

Regresión lineal múltiple
Técnica que pretende determinar la combinación lineal de variables independientes cuyos cambios son los mejores predictores de los cambios experimentados por la variable dependiente. Todas las variables que intervienen en la regresión son métricas, aunque admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en variables dami.

49

Modelo logit
Modelo de elección discreta en el que la función de distribución de probabilidad de la variable perturbación es la función logística**.**

Modelo logit multinomial
Modelo logit en el que la variable dependiente es politómica en lugar de dicotómica.

Modelo probit
Modelo de elección discreta en el que la función de distribución de probabilidad de b, variable perturbación es la función normal.

50

Análisis conjunto
Técnica que se emplea para entender cómo conforman los individuos sus preferencias hacia los objetos, normalmente marcas o productos**.**

Segmentación jerárquica
Técnica de análisis de la dependencia que tiene por objeto distinguir grupos de elementos homogéneos en una población a través de un proceso iterativo descendente de partición de la muestra total en sucesivos grupos en virtud del valor adoptado por la variable dependiente, el cual es función de los valores presentados por las variables independientes.

51

Análisis con clases latentes
Técnica que busca distinguir en una muestra grupos de elementos homogéneos en función de los valores que adopta una variable latente no métrica. Tales valores son las categorías de esa variable, las cuales reciben el nombre de clases latentes.

Análisis con ecuaciones estructurales **(o análisis de estructuras de covarianzas)**
Técnica que permite analizar varias relaciones de dependencia que se presentan simultáneamente.

52

Técnicas de análisis de la interdependencia

Se incluyen en esta categoría las siguientes: el análisis factorial y por componentes principales, el análisis de correspondencias, el análisis de conglomerados, el escalamiento multidimensional y el análisis con clases latentes.
En el cuadro siguiente se observan algunas características diferenciadoras entre ellas, como son el tipo de variables que permiten manejar y qué clase de elementos componen los grupos que resultan de la aplicación de cada una.

53

Técnicas de análisis de la interdependencia.

Técnica	Variable	Forma grupos de
Análisis factorial y por componentes principales	Métrica	Variables
Análisis de correspondencias	No métrica	Categorías de variables
Análisis de conglomerados	Métrica y no métrica	Objetos
Escalamiento multidimensional	Métrica y no métrica	Objetos
Análisis con clases latentes	No métricas	Objetos y categorías de variables

54

Análisis factorial
Técnica de análisis de la interdependencia presentada por un cierto número de variables susceptible de ser sintetizada en un conjunto de factores comunes que subyacen tras ella. Dichos factores pueden ser comunes (captan la variabilidad compartida por todas las variables), o específicos (captan la variabilidad propia de cada variable, sin relación con las demás).

Análisis por componentes principales
Técnica de análisis de la interdependencia presentada por un cierto número de variables susceptible de ser sintetizada en un conjunto de factores comunes que subyacen tras ella. Dichos factores o componentes buscan explicar la mayor proporción posible de la variabilidad total, lo que quiere decir que, a diferencia de lo que ocurre en análisis factorial. no existen factores específicos.

55

Análisis de correspondencias
Técnica basada en el estudio de la asociación entre las categorías de múltiples variables no métricas, que persigue la elaboración de un mapa perceptual que ponga de manifiesto dicha asociación en modo gráfico.

Análisis de conglomerados **(o análisis cluster)**
Técnica cuyo fin es clasificar sujetos u objetos en función de ciertas características de
modo que los elementos de cada grupo sean muy similares entre sí.

56

Escalamiento multidimensional
Técnica cuyo fin es elaborar una representación gráfica que permita conocer la imagen que los individuos se crean de un conjunto de objetos por posicionamiento de cada uno en relación a los demás.

Análisis con clases latentes
Técnica que busca distinguir en una muestra grupos de elementos homogéneos en función de los valores que adopta una variable latente no métrica. Tales valores son las categorías de esa variable, las cuales reciben el nombre de clases latentes.

57

Otras técnicas

Elección multicriterio discreta
Conjunto de métodos de ayuda en la resolución de problemas de decisión en los que se han de tener en cuenta diferentes puntos de vista o criterios y en los que se baraja un número finito de alternativas.

Data mining (o minería de datos o extracción de datos)
Proceso mediante el cual se explora y analiza un gran volumen de datos con el fin de descubrir relaciones, reglas o patrones de comportamiento en ellos que sean de utilidad para el usuario en la toma de decisiones.

Análisis con redes neuronales
Técnica cuya forma de proceder pretende replicar el funcionamiento del cerebro humano, intentando aprender de los errores cometidos en aras de la consecución del mejor resultado posible.

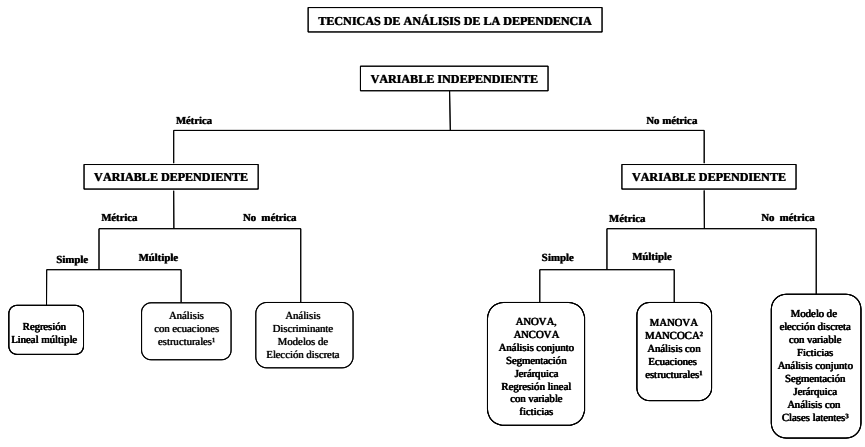
58

La elección de una técnica concreta

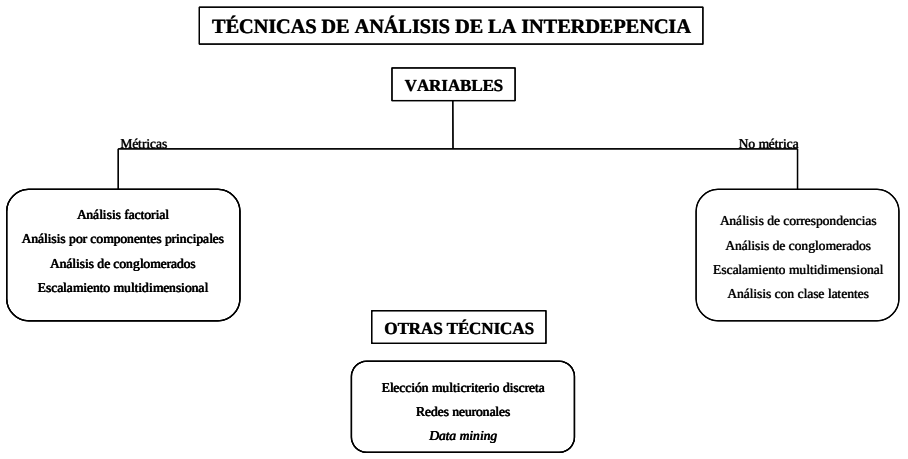
A la luz de lo expuesto en este apartado se deduce que la elección de una determinada técnica de análisis multivariable pasa por dar respuesta previa a preguntas como

- **¿Sigue un fin predictivo o clasificatorio?**
- **¿Se puede distinguir entre variables dependiente e independientes?**
- **¿Cuántas variables dependientes hay?**
- **¿Qué tipo de escalas de medida presentan las variables?**
- **¿Estas se distribuyen normalmente?**

59



60



61

Técnicas a estudiar

Análisis de varianza de un factor	⇨	De análisis de la dependencia En SPSS menú Analizar/Comparar Medias
Regresión lineal simple	⇨	De análisis de la dependencia En SPSS menú Analizar/Regresión
Regresión lineal múltiple	⇨	De análisis de la dependencia En SPSS menú Analizar/Regresión
Regresión logística	⇨	De análisis de la dependencia En SPSS menú Analizar/Regresión/Logística
Análisis Factorial	⇨	De análisis de la interdependencia En SPSS menú Analizar/Reducción de datos

62

Análisis de Varianza de un factor

- El análisis ANOVA de un factor
  - Datos y supuestos
  - Prueba de homogeneidad de Varianzas.
  - Comparaciones post-hoc
- Prueba no parámetrica H de Kruskal-Wallis

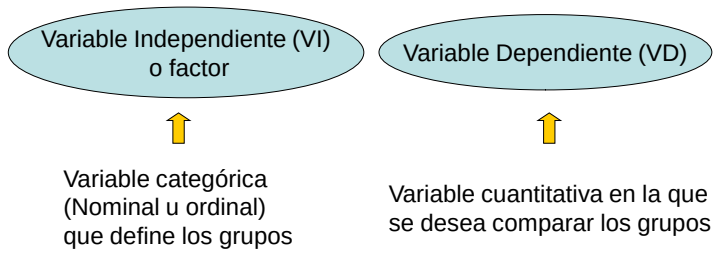
63

Análisis de Varianza

Análisis de varianza de un factor

El análisis ANOVA de un factor es una generalización de la prueba T para dos muestras independientes al caso de diseños con más de dos muestras.

Sirve para comparar varios grupos en una variable cuantitativa.



Variable categórica (Nominal u ordinal) que define los grupos

Variable cuantitativa en la que se desea comparar los grupos

64

Análisis de varianza de un factor

Datos. Los valores de la variable de factor deben ser enteros y la variable dependiente debe ser cuantitativa (nivel de medida de intervalo).

Supuestos. Cada grupo es una muestra aleatoria independiente procedente de una población normal. El análisis de varianza es robusto a las desviaciones de la normalidad, aunque los datos deberán ser simétricos. Los grupos deben proceder de poblaciones con varianzas iguales. Para contrastar este supuesto, utilice la prueba de Levene de homogeneidad de varianzas.

65

Análisis de varianza de un factor

La hipótesis que se pone a prueba en el ANOVA de un factor es que las medias poblacionales (las medias de la VD en cada nivel de la VI) son iguales.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$$

Si las medias poblacionales son iguales, eso significa que los grupos no difieren en la VD y que, en consecuencia, la VI o factor es independiente de la VD.

El procedimiento para poner a prueba la Ho consiste en obtener un estadístico, llamado F, que refleja el grado de parecido existente entre las medias que se están comparando.

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n\hat{\sigma}_{\bar{Y}}^2}{\bar{S}_j^2}$$

66

Análisis de varianza de un factor

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n\hat{\sigma}_{\bar{Y}}^2}{\bar{S}_j^2}$$

El numerador del estadístico F es una estimación de la varianza poblacional basada en la variabilidad existente entre las medias de cada grupo

El denominador del estadístico F es una estimación de la varianza poblacional, basada en la variabilidad existente dentro de cada grupo (j se refiere a los distintos grupos o niveles del factor)

Si las medias poblacionales son iguales, las medias muestrales de los diferentes grupos serán parecidas, existiendo entre ellas tan sólo diferencias atribuibles al azar. En ese caso, la estimación $\hat{\sigma}_1^2$ (basada en las diferencias entre las medias muestrales) reflejará el mismo grado de variación que la estimación $\hat{\sigma}_2^2$ basada en las diferencias entre las puntuaciones individuales dentro de cada grupo) y el cuociente F tomará un valor próximo a 1

67

Análisis de varianza de un factor

Por el contrario, si las medias muestrales son distintas, la estimación $\hat{\sigma}_1^2$ reflejará mayor grado de variación que la estimación $\hat{\sigma}_2^2$, en cuyo caso el cuociente F tomará un valor mayor que 1. Cuanto más diferentes sean las medias muestrales, mayor será el valor de F.

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n\hat{\sigma}_{\bar{Y}}^2}{\bar{S}_j^2}$$

Si las poblaciones muestreadas son normales y sus varianzas iguales, el estadístico F se distribuye según el modelo de probabilidad F de Fisher Snedecor

Los grados de libertad del numerador son el número de grupos menos 1;los del denominador el número total de observaciones menos el número de grupos.

68

Análisis de varianza de un factor

Ejemplo: ANOVA de un factor

Consideremos el archivo de Datos de empleados

Variable dependiente: Salario actual (salario)
Factor: Categoría laboral (catlab)

Descriptivos

Salario actual

	N	Media	Desviación típica	Mínimo	Máximo
Administrativo	363	$27,838.54	$7,567.995	$15,750	$80,000
Seguridad	27	$30,938.89	$2,114.616	$24,300	$35,250
Directivo	84	$63,977.80	$18,244.776	$34,410	$135,000
Total	474	$34,419.57	$17,075.661	$15,750	$135,000

69

Análisis de varianza de un factor

ANOVA

Salario actual

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	89438483925,9	2	44719241962,971	434,481	,000
Intra-grupos	48478011510,4	471	102925714,459		
Total	137916495436	473			

La tabla ANOVA muestra el resultado del estadístico F (cuociente entre dos estimadores diferentes de la varianza poblacional. Uno de los estimadores se obtiene a partir de la variación existente entre las medias de los grupos (variación Inter-grupos). El otro estimador se obtiene a partir de la variación existente entre las puntuaciones dentro de cada grupo (variación Intra-grupos)

La tabla ofrece una cuantificación de ambas fuentes de variación (Suma de cuadrados), los grados de libertad asociados a cada suma de cuadrados (gl) y el valor concreto que adopta cada estimador de la varianza poblacional (medias cuadráticas, que se obtienen dividiendo las sumas de cuadrados entre sus correspondientes grados de libertad)

70

Análisis de varianza de un factor

ANOVA

Salario actual

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	89438483925,9	2	44719241962,971	434,481	,000
Intra-grupos	48478011510,4	471	102925714,459		
Total	137916495436	473			

El cuociente entre las dos medias cuadráticas (la inter-grupos y la intra-grupos) proporciona el valor del estadístico F, el cual aparece acompañado de su correspondiente nivel de significación observado (Sig)

Puesto que el nivel crítico (Sig=0,000) es menor que 0,05, debe rechazarse la hipótesis de igualdad de medias

Puede concluirse que las poblaciones definidas por la variable catlab no poseen el mismo salario medio: hay al menos una población cuyo salario medio difiere del de al menos otra.

71

Análisis de varianza de un factor

Prueba de homogeneidad de las varianzas

El estadístico F del ANOVA de un factor se basa en el cumplimiento de dos supuestos fundamentales:

normalidad **y** **homocedasticidad**

Normalidad significa que la variable dependiente se distribuye normalmente en la J poblaciones muestreadas (tantas como grupos definidos por la variable factor); si los tamaños de los grupos son grandes, el estadístico F se comporta razonablemente bien incluso con distribuciones poblacionales sensiblemente alejadas de la normalidad

Homocedasticidad o igualdad de varianzas significa que la J poblaciones muestreadas poseen la misma varianza; con grupos de distinto, tamaño el incumplimiento de este supuesto debe ser cuidadosamente vigilado.

Prueba de homogeneidad de las varianzas. Prueba de Levene

La prueba de Levene permite contrastar el supuesto de homogeneidad de varianzas, es decir permite contrastar la hipótesis de que los grupos definidos por la variable factor proceden de poblaciones con las misma varianza

Prueba de homogeneidad de varianzas

Salario actual

Estadístico de Levene	gl1	gl2	Sig.
59,733	2	471	,000

La tabla contiene el estadístico de Levene. Puesto que el nivel crítico es menor que 0,05, se debe rechazar la hipótesis de igualdad de varianzas y concluir, que en las poblaciones definidas por las tres categorías laborales, las varianzas de la variable salario no son iguales.

73

Pruebas robustas de igualdad de las medias

Salario actual

	Estadístico[a]	gl1	gl2	Sig.
Welch	162,200	2	117,312	,000
Brown-Forsythe	306,810	2	93,906	,000

a. Distribuidos en F asintóticamente.

El estadístico de Welch y el de Brown-Forsythe contrasta la igualdad de las medias de grupo. Este estadístico es preferible al estadístico F cuando no se puede mantener el supuesto de igualdad de varianzas.

Puesto que el nivel crítico asociado a ambos estadísticos es menor que 0,05, se puede rechazar la hipótesis de igualdad de medias y concluir que los promedios salariales de las poblaciones no son iguales

74

Comparaciones post-hoc

El estadístico F del ANOVA únicamente permite contrastar la hipótesis general de que los J promedios comparados son iguales. Rechazar esa hipótesis significa que las medias poblacionales comparadas no son iguales, pero no permite precisar dónde en concreto se encuentran las diferencias detectadas.

Para saber qué media difiere de qué otra se debe utilizar un tipo particular de contrastes denominados comparaciones múltiples post-hoc

Asumiendo varianzas iguales

Existen varios métodos , el más utilizado es la opción Tukey

75

No asumiendo varianzas iguales

Existen varios métodos , el más utilizado es la opción Games -Howell

En nuestro ejemplo por la prueba de Levene, no podemos asumir que las varianzas poblacionales sean iguales por tanto debe prestarse atención a la opción de Games- Howell

Comparaciones múltiples

Variable dependiente: Salario actual

Games-Howell

(I) Categoría laboral	(J) Categoría laboral	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Administrativo	Seguridad	-$3,100.349*	$568.679	,000	-$4,454.82	-$1,745.88
	Directivo	-$36,139.258*	$2,029.912	,000	-$40,977.01	-$31,301.51
Seguridad	Administrativo	$3,100.349*	$568.679	,000	$1,745.88	$4,454.82
	Directivo	-$33,038.909*	$2,031.840	,000	-$37,881.37	-$28,196.45
Directivo	Administrativo	$36,139.258*	$2,029.912	,000	$31,301.51	$40,977.01
	Seguridad	$33,038.909*	$2,031.840	,000	$28,196.45	$37,881.37

*. La diferencia entre las medias es significativa al nivel .05.

Puede concluirse que todos los promedios comparados difieren significativamente.

76

Análisis no paramétrico
Prueba de H de Kruskal-Wallis

La prueba de Mann- Whitney para dos muestras independientes fue extendida al caso de más de dos muestras por Kruskal y Wallis (1952). La situación experimental que permite resolver esta prueba es similar a la estudiada a propósito del ANOVA de un factor completamente aleatorizado: *J* muestras son aleatoria e independientemente extraídas de *J* poblaciones para averiguar si las *J* poblaciones son idénticas o alguna de ellas presenta promedios mayores que otra.

Las ventajas fundamentales de esta prueba frente al estadístico *F* del ANOVA de un factor son dos:

(1) no necesita establecer supuestos sobre las poblaciones originales tan exigentes como los del estadístico *F* (normalidad, homocedasticidad); y

(2) permite trabajar con datos ordinales.

Si se cumplen los supuestos en los que se basa el estadístico *F,* la potencia de éste es mayor que la que es posible alcanzar con el estadístico *H* de Kruskal- Wallis.

77

Ejemplo:
Pruebas no paramétricas /Varias muestras independientes
H de Kruskal-Wallis

Rangos

	Categoría laboral	N	Rango promedio
Salario actual	Administrativo	363	190,37
	Seguridad	27	278,98
	Directivo	84	427,85
	Total	474	

Estadísticos de contraste[a,b]

	Salario actual
Chi-cuadrado	207,679
gl	2
Sig. asintót.	,000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Categoría laboral

La primera tabla ofrece el tamaño de cada grupo (N) y los rangos promedios resultantes de la asignación de rangos a las puntuaciones de los tres grupos.

En la segunda tabla, puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis de igualdad de medias poblacionales y concluir que las poblaciones comparadas difieren en salario actual.

78

Análisis de regresión lineal

- **Análisis de regresión lineal simple**
- **Análisis de regresión lineal múltiple**

79

Análisis de regresión lineal simple

- Análisis de regresión con dos variables: Algunas ideas básicas
  - Concepto de función de regresión poblacional
  - Significado del término lineal
  - Especificación estocástica de la FRP
  - Función de regresión muestral (FRM)
- Análisis de regresión con dos variables: problema de estimación.
  - Método de Mínimos cuadrados ordinarios (MCO)
  - Modelo clásico: Supuestos detrás del método MCO
  - Precisión o errores estándar de MCO
  - Propiedades de los estimadores de MCO
  - Coeficiente de determinación r2:una medida de bondad de ajuste
  - Coeficiente de correlación muestral y propiedades de r
  - Interpretación de la pendiente

80

Análisis de regresión lineal simple

- Modelo clásico de regresión lineal normal (MCRLN)

- Regresión con dos variables: estimación de intervalos y pruebas de hipótesis.
  - Intervalos de confianza
  - Pruebas t
- Aplicación problemas de predicción
  - Predicción del valor de la media condicional
  - Predicción de un valor individual

- Formas funcionales de los modelos de regresión
  - Modelo log-lineal
  - Modelos semilogaritmicos

81

Análisis de regresión lineal simple

Algunas ideas básicas

El análisis de regresión se relaciona en gran medida con la estimación y/o predicción de la media (de la población) o valor promedio de la variable dependiente, con base en los valores conocidos o fijos de las variables explicativas.

Consideremos los datos de la tabla siguiente, la que se refiere a la población total de 60 familias de una comunidad hipotética, así como a su ingreso semanal (X) y a su gasto de consumo semanal (Y), dados en dólares.

82

Algunas ideas básicas

Tabla 1

Gastos de consumo familiar semanal Y,$. Ingreso familiar semanal X,$

Y\X	80	100	120	140	160	180	200	220	240	260
	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
		88		113	125	140		160		185
				115				162		191
Total	325	462	445	707	678	750	685	1043	777	1211
Medias	65	77	89	101	113	125	137	149	155	173

Las 60 familias se dividen en 10 grupos de ingresos (de $80 a $260). Se tienen 10 valores fijos de X y los correspondientes valores de Y para cada uno de los valores X; así que hay 10 subpoblaciones Y

83

Algunas ideas básicas

Se tienen 10 valores medios para las 10 subpoblaciones de Y.

A estos valores medios se les denomina valores esperados condicionales, en vista de que dependen de los valores dados a la variable condicional X. Se denota por E(Y/X)

Resulta importante distinguir dichos valores condicionales esperados del valor esperado incondicional del gasto de consumo semanal, E(Y).

E(Y)=7272/60=121,2
Es incondicional en el sentido de que para obtener esta cifra se omiten los niveles de ingresos de las diversas familias

84

Algunas ideas básicas

¿Cuál es el valor esperado del gasto de consumo semanal de una familia?

La media incondicional: $121,20

¿Cuál es el valor esperado del gasto de consumo semanal de una familia cuyo ingreso mensual es, digamos, $140?

La media condicional: $101

Saber el nivel de ingreso nos permite predecir mejor el valor medio del gasto de consumo

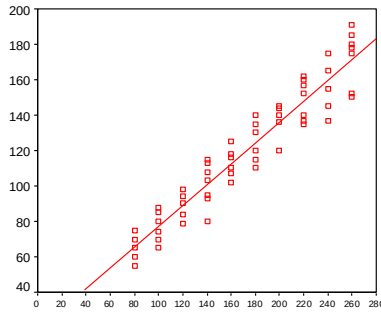
85

Algunas ideas básicas

Se puede observar en él gráfico de dispersión, al unir las medias condicionales la recta de regresión poblacional (RRP). (o regresión de Y sobre X).
El adjetivo "poblacional " se debe al hecho de que en este ejemplo se consideró una población de 60 familias.

Gráfico de dispersión

Gasto de consumo v/s Ingreso



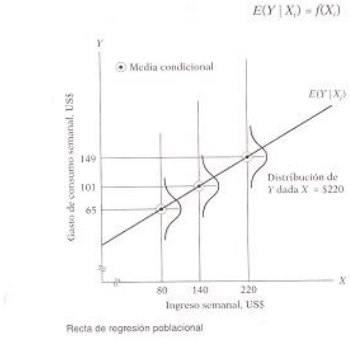
Ingreso semanal

A pesar de la variabilidad del gasto para cada ingreso, en promedio el consumo semanal se incrementa en la misma medida que el ingreso

86

Curva de regresión poblacional

Desde el punto de vista geométrico, **una curva de regresión poblacional** es simplemente el lugar geométrico de las medias condicionales de la variable dependiente para los valores fijos de la (s) variables explicativa(s).



Recta de regresión poblacional

Es la curva que conecta las medias de las subpoblaciones de Y que corresponden a los valores del regresor X

Concepto de función de regresión poblacional (FRP)

Es claro que cada media condicional E(Y/Xi) es función de Xi, donde Xi es un valor dado de X.

$$E(Y/Xi)=f(Xi) \quad \textbf{(1)}$$

y f(Xi) denota alguna función de la variable explicativa X.

¿Qué forma toma la función f(Xi)?

En una situación real no tenemos la totalidad de la población para efectuar el análisis.
La forma funcional de la FRP es, una pregunta empírica, aunque en casos específicos la teoría puede tener algo que decir. Por ejemplo, un economista podría plantear que el gasto de consumo está relacionado linealmente con el ingreso.

Por tanto, como una primera aproximación podemos suponer que la FRP es una función lineal de Xi

$$E(Y / X_i) = \beta_1 + \beta_2 X_i$$

Ecuación de regresión poblacional FRP

$$E(Y / X_i) = \beta_1 + \beta_2 X_i \quad \Leftarrow$$ Ecuación de regresión poblacional FRP **(2)**

Donde β_1 y β_2 son parámetros no conocidos pero fijos que se denominan coeficientes de regresión.

En el análisis de regresión el interés es estimar la FRP, es decir estimar los valores de β_1 y β_2 no conocidos con base en las observaciones de Y y X

89

Significado del término lineal

Linealidad en las variables

Se dice que una función Y=f(X) es lineal en X si X aparece elevado a una potencia o índice de 1 solamente y dicha variable no está multiplicada ni dividida por alguna otra variable

$$E(Y / X_i) = \beta_1 + \beta_2 X_i \quad \Leftarrow$$ es lineal en Xi.

Geométricamente la curva de regresión es una línea recta

Linealidad en los parámetros

Se dice que una función es lineal en el parámetro, β_1 por ejemplo si β_1 aparece elevado a una potencia o índice de 1 solamente y no está multiplicado ni dividido por ningún otro parámetro.

$$E(Y / X_i) = \beta_1 + \beta_2 X_i^2 \quad \Leftarrow$$ Es lineal en los parámetros pero no es lineal en la variable X

90

Especificación estocástica de la FRP

¿Qué podemos decir sobre la relación entre el gasto de consumo de una familia individual y un nivel dado de ingresos? Se observa en la figura , que dado el nivel de ingresos de Xi, el gasto de consumo de una familia individual está agrupado alrededor del consumo promedio de todas las familias en ese nivel de Xi, esto es, alrededor de su esperanza condicional. Por consiguiente, podemos expresar la desviación de un Yi individual alrededor de su valor esperado de la siguiente manera:

$$u_i = Y_i - E(Y / X_i) \quad \text{o} \quad Y_i = E(Y / X_i) + u_i \quad \textbf{(3)}$$

Donde la desviación u_i es una variable aleatoria no observable que toma valores positivos o negativos. Técnicamente **, u_i es conocida como perturbación estocástica o término de error estocástico.**

91

Especificación estocástica de la FRP

Se puede decir que el gasto de una familia individual, dado su nivel de ingresos, puede ser expresado como la suma de dos componentes

$$Y_i = E(Y / X_i) + u_i \quad \textbf{(4)}$$

La media del gasto de consumo de todas las familias con el mismo nivel de ingresos.

Componente aleatorio . Es un sustituto para todas aquellas variables que son omitidas del modelo pero que colectivamente afectan a Y

92

Especificación estocástica de la FRP

Si se supone que $E(Y / X_i)$ es lineal en X_i como en la ec (2) la ecuación (3) puede escribirse como

$$Y_i = E(Y / X_i) + u_i = \beta_1 + \beta_2 X_i + u_i \quad \textbf{(5)}$$

La ecuación plantea que el gasto de consumo de una familia está relacionado linealmente con su ingreso, más el término de perturbación. Así los gastos de consumo individual, dado X=US$80, pueden ser expresados como

$$Y_1 = 55 = \beta_1 + \beta_2 (80) + u_2$$
$$Y_2 = 60 = \beta_1 + \beta_2 (80) + u_2$$
$$Y_3 = 65 = \beta_1 + \beta_2 (80) + u_3$$
$$Y_4 = 70 = \beta_1 + \beta_2 (80) + u_4$$
$$Y_5 = 75 = \beta_1 + \beta_2 (80) + u_5$$

93

Especificación estocástica de la FRP

Ahora, si se toma el valor esperado de (5), obtenemos

$$Y_i = E(Y / X_i) + u_i \qquad (5)$$
$$E(Y_i / X_i) = E[E(Y / X_i)] + E(u_i / X_i)$$
$$= E(Y/X_i) + E(u_i / X_i)$$

Puesto que $E(Y_i / X_i)$ es lo mismo que $E(Y / X_i)$

Implica que $E(u_i / X_i) = 0$ **(6)**

Así, el supuesto de que la recta de regresión pasa a través de las medias condicionales de Y implica que los valores de la media condicional de u_i son cero.

94

Especificación estocástica de la FRP

La especificación estocástica

$$Y_i = E(Y / X_i) + u_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(7)}$$

Tiene la ventaja que muestra claramente otras variables además del ingreso, que afectan el gasto de consumo y que un gasto de consumo de familias individuales no puede ser explicado en su totalidad solamente por la(s) variable(s) incluida(s) en el modelo de regresión.

95

Función de regresión muestral (FRM)

En la práctica lo que se tiene al alcance no es más que una muestra de valores de Y que corresponden a algunos valores fijos de X. Por consiguiente la labor ahora es estimar la FRP con base en información muestral.
Supóngase que no se conocía la población de la tabla 1 y que la única información que se tenía era una muestra de valores de Y seleccionada aleatoriamente para valores dados de X tal como se presenta en la tabla 2

De la muestra de la tabla 2, ¿se puede predecir el gasto de consumo semanal promedio Y para la población correspondiente a los valores de X seleccionados?

¿Se puede estimar la forma FRP a partir de la información muestral?

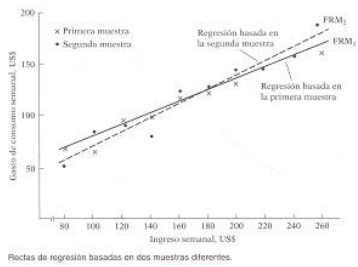
Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Tabla 2 Primera muestra

96

Función de regresión muestral (FRM)

Consideremos otra muestra tomada de la población de la tabla 1. Las rectas de la figura se conocen como rectas de regresión muestral. En general, se podrían obtener N FRM diferentes para N muestras diferentes y estas FRM no necesariamente son iguales



Rectas de regresión basadas en dos muestras diferentes.

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

Tabla 3 Segunda muestra

97

Ahora, en forma análoga a la FRP en la cual se basa la recta de regresión poblacional, se puede desarrollar el concepto de función de regresión muestral.
La contraparte muestral de (1) puede escribirse como

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Es la contraparte de $E(Y / X_i) = \beta_1 + \beta_2 X_i$

Donde

$\hat{Y}_i = estimador\ \text{de E(Y/X)}$

$\hat{\beta}_1 = \text{estimador de } \beta_1$

$\hat{\beta}_2 = \text{estimador de } \beta_2$

Un **estimador,** conocido también como estadístico (muestral) es simplemente una regla, o método que dice cómo estimar el parámetro poblacional a partir de la información suministrada por la muestra disponible. Un valor numérico particular obtenido por el estimador en una aplicación es conocido como **estimado.**

98

Función de regresión muestral (FRM) en su forma estocástica

La FRM en su forma estocástica se puede expresar como

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i \quad \textbf{(8)}$$

Donde $\hat{\mu}_i$ denota el término residual (muestral)

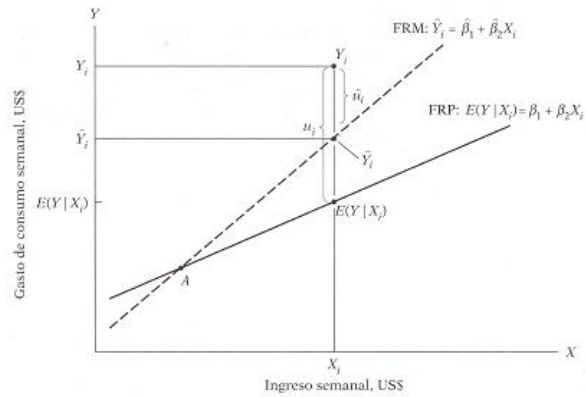
Conceptualmente es análogo a ui y puede ser considerado como un estimado de ui

El objetivo principal en el análisis de regresión es estimar la FRP $Y_i = \beta_1 + \beta_2 X_i + \mu_i$

Con base en la FRM $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$

99

Rectas de regresión muestral y poblacional



Rectas de regresión muestral y poblacional.

Debido a fluctuaciones muestrales el estimado de la FRP basado en FRM es, en el mejor de los casos, una aproximación.

100

Rectas de regresión muestral y poblacional

Para X=Xi, se tiene una observación muestral Y=Yi. En términos de la FRM, la Yi observada puede ser expresada como

$$Y_i = \hat{Y}_i + \hat{\mu}_i$$

Y en términos de la FRP, puede ser expresada como

$$Y_i = E(Y / X_i) + \mu_i$$

Dado que la FRM es apenas una aproximación de la FRP, ¿se puede diseñar un método que haga que esta aproximación sea lo más ajustada posible?

101

Función de regresión simple: problema de estimación

La tarea consiste en estimar la función de regresión poblacional (FRP) con base en la función de regresión muestral (FRM) en la forma más precisa posible.
Los dos métodos de estimación que suelen utilizarse son:

1) Los mínimos cuadrados ordinarios (MCO)
2) La máxima verosimilitud (MV).

El método de MCO es el que más se emplea en el análisis de regresión por ser en gran medida más intuitivo y matemáticamente más simple.

102

Método de mínimos cuadrados ordinarios (MCO)

El método MCO se atribuye a Carl Friedrich Gauss un matemático alemán. Bajo ciertos supuestos el método tiene algunas propiedades estadísticas muy atractivas que lo han convertido en uno de los más eficaces y populares del análisis de regresión.

Primero se estima $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$ **(9)**

que muestra que los residuos son simplemente las diferencias entre los valores observados y los estimados de Y.
Ahora, dados n pares de observaciones de Y y X, se está interesado en determinar la FRM de tal manera que esté lo más cerca posible a la Y observada.

103

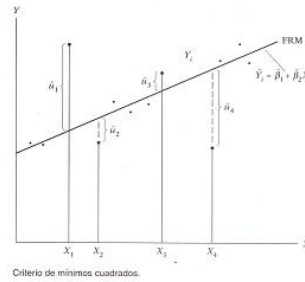
Método de mínimos cuadrados ordinarios (MCO)

Con este fin se puede adoptar el siguiente criterio: seleccionar la FRM de tal manera que la suma de los residuos :

$$\sum \hat{u}_i = \sum \left(Y_i - \hat{Y}_i \right)$$

sea la menor posible.

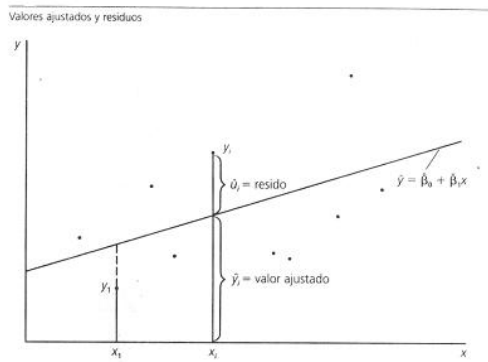
Este criterio, no es muy bueno porque a todos los residuos se les da la misma importancia sin considerar qué tan cerca o qué tan dispersas estén las observaciones individuales de la FRM. Debido a lo anterior, es muy posible que la suma algebraica de los residuos sea pequeña (aun cero) a pesar de que las $\hat{u}_i$ están bastante dispersas alrededor de FRM.



Criterio de mínimos cuadrados.

104

Valores ajustados y residuos



105

Método de mínimos cuadrados ordinarios (MCO)

Se puede evitar este problema si se adopta el criterio de mínimos cuadrados, el cual establece que la FRM puede determinarse en forma tal que

$$\sum \hat{u}_i^{\,2} = \sum \left(Y_i - \hat{Y}_i \right)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right)^2 \qquad \textbf{(10)}$$

sea la menor posible. Este método da más peso a los residuos tales como $\hat{u}_1$ y $\hat{u}_4$ que a los residuos $\hat{u}_2$ y $\hat{u}_3$

El procedimiento de MCO genera las siguientes ecuaciones para estimar β_1 y β_2 donde n es el tamaño de la muestra

106

Método de mínimos cuadrados ordinarios (MCO)

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

Ecuaciones normales

Resolviendo las ecuaciones normales simultáneamente se obtiene

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \hat{\beta}_1 = \overline{\mathrm{Y}} - \hat{\beta}_2 \overline{\mathrm{X}}$$

Estimadores de mínimos cuadrados

107

Modelo clásico de regresión lineal: supuestos detrás del método MCO

El modelo de Gauss, modelo clásico o estándar de regresión lineal (MCRL) el cual es el cimiento de la mayor parte de la teoría econométrica, plantea 10 supuestos.

Supuesto 1: Modelo de regresión lineal
El modelo de regresión es lineal en los parámetros

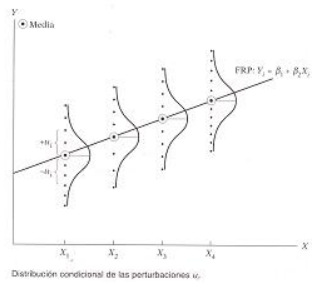
$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$ modelo simple

Supuesto 2: Los valores de X son fijos en muestreo repetido.
Significa que el análisis de regresión es un análisis de regresión condicional, esto es, condicionado a los valores dados del (los) regresor X.

108

Supuesto 3: El valor medio de la perturbación ui es igual a cero.
Dado el valor de X, el valor esperado del término aleatorio de perturbación ui es cero.

$$E(u_i / X_i) = 0$$



Distribución condicional de las perturbaciones u_i

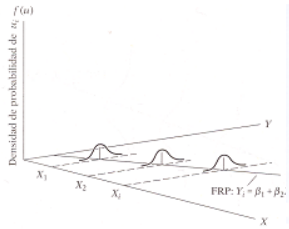
Nótese que el supuesto E(ui/Xi)=0 implica que

$$E(Y / X_i) = \beta_1 + \beta_2 X_i$$

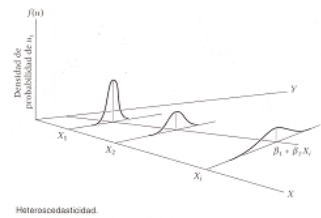
109

Supuesto 4: Homocedasticidad o igual varianza de ui.
Dado el valor de X, la varianza de ui es la misma para todas las observaciones, es decir, las varianzas condicionales de ui son idénticas.

$$\text{var}(ui / Xi) = \sigma^2$$



Homocedasticidad.



Heterocedasticidad.

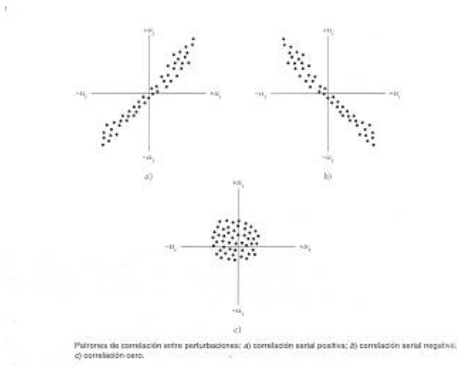
Homocedasticidad **Heterocedasticidad**

110

Supuesto 5: No existe auto correlación entre las perturbaciones.

Dados dos valores cualquiera de X, Xi y Xj , la correlación entre dos ui y uj es cero.

$$\text{cov}(ui, uj / Xi, X_j) = 0$$



Patrones de correlación entre perturbaciones: a) correlación serial positiva; b) correlación serial negativa; c) correlación cero.

111

Supuesto 6: La covarianza entre ui y Xi es cero o E(uiXi)=0

$$\text{cov}(ui, X_i) = 0$$

Supuesto 7: El número de observaciones n debe ser mayor que el número de parámetros por estimar.

Supuesto 8: Variabilidad en los valores de X.

No todos los valores de X en una muestra dada deben ser iguales.

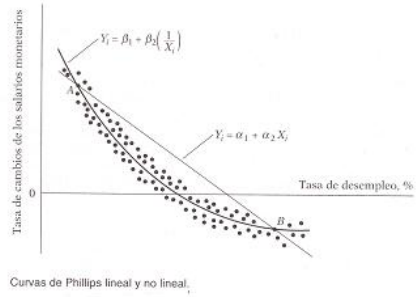
$$\text{var}(X) > 0$$

Recordar que la varianza muestral de X es

$$\text{var}(X) = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

112

Supuesto 9: El modelo de regresión está correctamente especificado.



Curvas de Phillips lineal y no lineal.

Supuesto 10:No hay multicolinealidad perfecta.
No hay relaciones perfectamente lineales entre las variables explicativas.

Precisión o errores estándar de los mínimos cuadrados estimados

Lo que se requiere es alguna medida de "confiabilidad" o precisión de los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$. En estadística la precisión de un valor estimado es medida por su error estándar (ee). Los errores estándar de los MCO estimados pueden obtenerse de la siguiente manera

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \qquad \text{ee}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \qquad \textbf{(11)}$$

Nota: El error estándar es la desviación estándar de la distribución muestral del estimador, y la distribución muestral es una distribución del conjunto de valores del estimador obtenidos de todas las muestras posibles de igual tamaño de una población dada.

Precisión o errores estándar de los mínimos cuadrados estimados

Nota: σ^2 es estimada mediante la fórmula

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \quad \text{(12)}$$

Suma de residuos al cuadrado (SRC)

Número de grados de libertad

Donde $\hat{\sigma}^2$ es el estimador de MCO de la verdadera σ^2 .

El término número de grados de libertad significa el número total de observaciones n menos el número de restricciones puestas en ellas.

115

Error estándar de la regresión

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}} \quad \text{(13)}$$

Es la desviación estándar de los valores de Y alrededor de la recta de regresión estimada, la cual es utilizada como una medida resumen de la bondad del ajuste de dicha recta

116

Propiedades de los estimadores de mínimos cuadrados: Teorema de Gauss-Markov

Dados los supuestos del modelo de regresión lineal clásica, los estimativos de mínimos cuadrados poseen propiedades ideales u óptimas, las cuales se encuentran resumidas en el teorema de Gauss Markov

Un estimador $\hat{\beta}_2$ de MCO es el mejor estimador lineal insesgado (MELI) de β_2 si:

1. Es lineal, es decir, una función lineal de una variable aleatoria tal como la variable dependiente Y en el modelo de regresión.

117

Propiedades de los estimadores de mínimos cuadrados: Teorema de Gauss-Markov

2. Es insesgado, es decir, su valor promedio o esperado, $E(\hat{\beta}_2)$ es igual al valor verdadero, $E(\hat{\beta}_2) = \beta_2$
3. Tiene varianza mínima entre la clase de todos los estimadores lineales insesgados; a un estimador insesgado con varianza mínima se le conoce como estimador eficiente

118

Teorema de Gauss-Markov

En el contexto del análisis de regresión se puede demostrar que los estimadores de MCO son MELI

Teorema de Gauss-Markov: Dados los supuestos del modelo clásico de regresión lineal, los estimadores de mínimos cuadrados, en la clase de estimadores lineales insesgados, tienen varianza mínima; es decir son MELI

Coeficiente de determinación r^2

Una medida de la bondad del ajuste

La cantidad r^2 se conoce como coeficiente de determinación (muestral) y es la medida más frecuente utilizada de la bondad del ajuste de una recta de regresión.

Mide la proporción o el porcentaje de la variación total en Y explicada por el modelo de regresión

Coeficiente de determinación r^2

Para calcular r^2 , para cada i se escribe:

$$y_i = \hat{y}_i + \hat{\mu}_i$$

Elevando la expresión al cuadrado en ambos lados y sumando sobre la muestra, se obtiene

$$\begin{aligned}\sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2\end{aligned} \quad \textbf{(14)}$$

puesto que $\sum \hat{y}_i \hat{u}_i = 0$ y $\hat{y}_i = \hat{\beta}_2 x_i$

121

Coeficiente de determinación r^2

Las diversas sumas de cuadrados que aparecen en la expresión anterior pueden describirse de la manera siguiente

$$\sum y_i^2 = \sum (Y_i - \bar{Y})^2 \quad \textbf{(STC)}$$

variación total de los valores reales de *y* con respecto a su media muestral, los cuales pueden ser llamados **suma total de cuadrados (STC)**

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2 \quad \textbf{(SEC)}$$

variación de los valores *Y* estimados alrededor de su media $(\hat{Y} = \bar{Y})$ que apropiadamente puede llamarse la suma de los cuadrados debida a la regresión [es decir, debida a la(s) variable(s) explicativa(s)], o explicada por ésta, o simplemente **la suma explicada de cuadrados (SEC).**

122

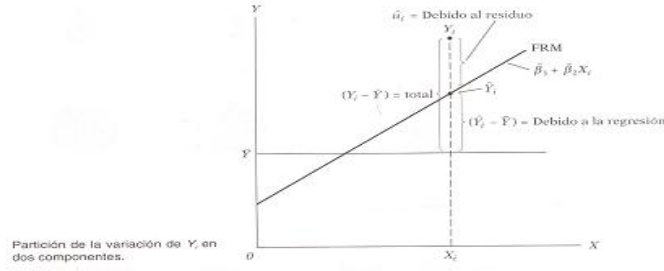
Coeficiente de determinación r^2

$$\sum \hat{u}_i^2 \qquad \textbf{(SRC)}$$

la variación residual o no explicada de los valores de *Y* alrededor de la recta de regresión, o simplemente la suma de residuos al cuadrado (SRC).

Así, (14) es

STC = SEC + SRC



Partición de la variación de *Y*, en dos componentes.

123

Coeficiente de determinación r^2

muestra que la variación total en los valores *Y* observados alrededor del valor de su media puede ser dividida en dos partes, una atribuible a la recta de regresión y la otra a fuerzas aleatorias, puesto que no todas las observaciones *Y* caen sobre la recta ajustada. Ahora dividiendo por la STS en ambos lados, se obtiene

⇐ **STC = SEC + SRC**

$$1 = \frac{\text{SEC}}{\text{STC}} + \frac{\text{SRC}}{\text{STC}}$$

⇒

$$= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$$

Ahora, se define r^2 como

$$r^2 = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2} = \frac{SEC}{STC}$$

124

Coeficiente de determinación r^2

O en forma alterna

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{SRC}{STC} \quad \textbf{(15)}$$

Coeficiente de determinación

La cantidad r^2 así definida se conoce como el coeficiente de determinación (muestral) y es la medida más frecuentemente utilizada de la bondad del ajuste de una recta de regresión

r^2 *mide la proporción o el porcentaje de la variación total en Y explicada por el modelo de regresión.*

125

Coeficiente de correlación muestral

Una cantidad estrechamente relacionada con r^2 pero conceptualmente muy diferente de ésta es el coeficiente de correlación, el cual, es una medida del grado de asociación entre dos variables. Puede ser calculado a partir de

$$r = \pm\sqrt{r^2}$$

O a partir de su definición

$$r = \frac{\sum x_i y_i}{\left(\sum x_i^2 \sum y_i^2\right)} = \frac{n\sum x_i y_i - \left(\sum x_i \sum y_i\right)}{\sqrt{\left[\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

(16)

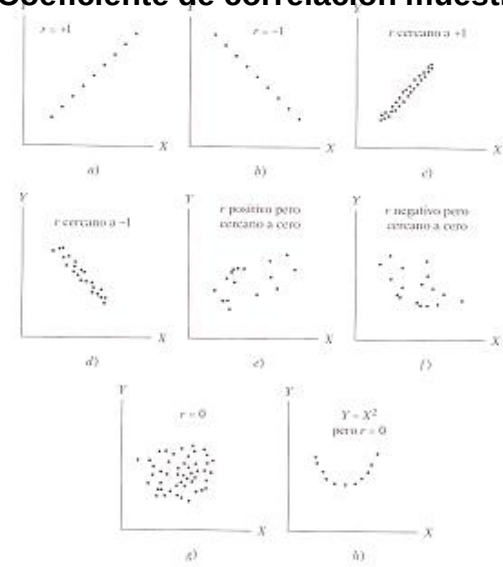
126

Propiedades de r

- Puede tener signo positivo o negativo, dependiendo del signo del término en el numerador de (16), el cual mide la *covariación* muestral de dos variables.
- Cae entre los límites de -1 y 1
- Es simétrico por naturaleza; es decir, el coeficiente de correlación entre X y Y (r_{xy}) es el mismo que entre *Y y* X (r_{yx}).
- Es independiente del origen y de la escala
- Si X y *Y* son estadísticamente independientes, el coeficiente de correlación entre ellos es cero; pero si *r* = O, esto no significa que las dos variables sean independientes. En otras palabras, una correlación igual a cero no necesariamente implica independencia.
- Es una medida de *asociación lineal* o *dependencia lineal* solamente; su uso en la descripción de relaciones no lineales no tiene significado.

127

Coeficiente de correlación muestral



Patrones de correlación (adaptado de Henri Theil, *Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, N.J., 1978, p. 86).

128

Interpretación de la pendiente:

Puesto que el coeficiente de la pendiente es simplemente la tasa de cambio, se mide en las unidades de la siguientes proporción

$$\frac{\text{unidades de la variable dependiente (Y)}}{\text{unidades de la variable explicativa (X)}}$$

La interpretación del coeficiente de la pendiente β_2 es que si X cambia en una unidad , la Y cambia en promedio en β_2 unidades

129

Ejemplo; Gasto de consumo familiar e ingreso familiar

Considerando una muestra de una población donde X representa ingreso familiar por semana e Y gastos de consumo familiar por semana, se obtienen los siguientes cálculos

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

$\hat{\beta}_1 = 24{,}4545 \quad se(\hat{\beta}_1) = 6{,}4138$

$\hat{\beta}_2 = 0{,}5091 \quad se(\hat{\beta}_2) = 0{,}0357$

$r^2 = 0{,}9621 \quad r = 0{,}9809$

Por tanto la línea de regresión estimada es

$$\hat{Y}_i = 24{,}4545 + 0{,}5091 X_i$$

130

Ejemplo; Gasto de consumo familiar e ingreso familiar

Resultados en SPSS

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,981[a]	,962	,957	6,493

a. Variables predictoras: (Constante), X

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	24,455	6,414		3,813	,005
	X	,509	,036	,981	14,243	,000

a. Variable dependiente: Y

131

Ejemplo; Gasto de consumo familiar e ingreso familiar

Interpretación:

El valor de $\hat{\beta}_1 = 0{,}5091$ que mide la pendiente de la línea, muestra que dentro del rango de la muestra de X comprendido entre $80 y $260 semanales, a medida que X aumenta, digamos en $1, el aumento estimado en el promedio de gastos de consumo semanales es de aproximadamente 51 centavos. El valor de $\hat{\beta}_1 = 24{,}45$, el cual corresponde a la intersección de la línea, indica el nivel promedio de los gastos de consumo semanales cuando el ingreso semanal es cero. No obstante, esta es una interpretación mecánica de la intersección. En el análisis de regresión esta interpretación literal del intercepto no es siempre significativa, aunque en el ejemplo que estamos considerando se puede argumentar que una familia sin ingreso alguno(ya sea por desempleo, despido, etc.) puede mantener algún nivel mínimo de gastos de consumo, ya sea tomando dinero prestado o utilizando sus ahorros.

132

Ejemplo; Gasto de consumo familiar e ingreso familiar

Sin embargo en general, se debe apelar al sentido común para interpretar la intersección puesto que es muy común que el rango que ha tomado la muestra de valores de X no haya incluido el valor cero como uno de los valores observados.
Quizá sea mejor interpretar la intersección como el efecto medio o promedio que tienen todas las variables omitidas del modelo de regresión sobre el valor de Y. El valor de 0,9621 para r cuadrado significa que cerca del 96% de la variación en los gastos de consumo semanales se explica por la variable ingreso; puesto que r cuadrado puede tener un valor máximo de 1 solamente, el r cuadrado observado sugiere que la línea de regresión muestral se ajusta muy bien a la información. El coeficiente de correlación de 0,9809 muestra que las dos variables, gastos de consumo e ingreso, están muy positivamente correlacionadas.

133

Ejemplo: Salario y educación

De la población de trabajadores en 1976, sea *y* = *sala*, en la que sala se mide, en dólares por hora. Así, para una persona cualquiera, si *sala* = *6.75*, el salario por hora es de 6.75 dólares. Sea *x* = *educ* los años de escolaridad; por ejemplo, *educ* = 12 corresponde a la educación preparatoria completa. Puesto que el salario promedio de la muestra es de 5.90 dólares, el índice de precios al consumidor indica que esta suma es equivalente a 16.64 dólares de 1997.
Con los datos de SALA 1.RAW, en los que *n* = 526 individuos, obtenemos la siguiente línea de regresión de MCO (o función de regresión muestra!):

$$\hat{sala} = -0.90 + 0.54\ educ.$$

134

Ejemplo: Salario y educación

Debemos interpretar con cuidado la ecuación. La intercepción -0.90 significa literalmente que una persona sin instrucción recibe un salario pronosticado de -90 centavos de dólar por hora, lo que, desde luego, es una tontería. Resulta que ningún miembro de la muestra tiene menos de ocho años de educación, lo que explica el pronóstico descabellado de una escolaridad de 0 años.
Para una persona con ocho años de escolaridad, el salario pronosticado es

$\widehat{sala}$ = -0.90 + 0.54(8) = 3.42, o 3.42 dólares por hora (en dólares de 1976).

La estimación de la pendiente implica que un año más de educación aumenta el salario promedio en 54 centavos de dólar por hora.

135

Ejemplo: Resultados electorales y gastos de campaña

El archivo VOTE 1.RAW contiene datos sobre los resultados electorales y los gastos de campaña de 173 contiendas bipartidistas para la Cámara de los Representantes estadounidense en 1988.
En cada contienda hay dos candidatos, A y B.
Sea voto*A* el porcentaje de los votos recibidos por el candidato A y part*A* el porcentaje de participación de los gastos de su campaña, ambos en el total correspondiente.
Además de part*A*, muchos otros factores influyen en los resultados electorales (entre ellos la calidad de los candidatos y posiblemente las sumas gastadas por A y B).
No obstante, podemos estimar un modelo de regresión simple para averiguar si gastar más que el contrario produce un porcentaje mayor en la votación.

136

Ejemplo: Resultados electorales y gastos de campaña

La ecuación estimada con las 173 observaciones es

$$\hat{votoA} = 40.90 + 0.306\ partA.$$

Esto significa que, si la participación de los gastos del candidato A aumenta un punto porcentual, éste casi obtiene un tercio de punto porcentual más de la votación total. En la ecuación de los resultados electorales $R^2 = 0.505$. Así, la participación en los gastos de campaña explica algo más de 50 por ciento de la variación en los resultados de esta muestra, lo cual es una proporción bastante considerable.

137

El supuesto de normalidad: El modelo clásico de regresión lineal normal

Recordemos que con los supuestos vistos anteriormente los estimadores de MCO $\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2$ satisfacían diferentes propiedades estadísticas muy deseables, tales como insesgamiento y varianza mínima . Si nuestro objetivo es únicamente la estimación puntual el método de MCO será suficiente, sin embargo la estimación puntual es sólo la formulación de un aspecto de la inferencia estadística.

Nuestro interés no consiste solamente en estimar la función muestral de regresión (FRM), sino también en utilizarla para obtener inferencias respecto a la función de regresión poblacional (FRP).

138

El supuesto de normalidad: El modelo clásico de regresión lineal normal

La regresión lineal normal clásica supone que cada ui, está normalmente distribuida con

$$Media: \quad E(u_i) = 0$$

$$Varianza: \quad E\left[u_i - E(u_i)\right]^2 = E(u_i^2) = \sigma^2$$

$$Cov(u_i, u_j): \quad E\left[u_i - E(u_i)\right]\left[u_j - E(u_j)\right] = E(u_i u_j) = 0 \quad i \neq j$$

Estos supuestos pueden expresarse en forma más compacta como

$$u_i \sim \mathrm{N}(0, \sigma^2)$$

El supuesto de normalidad

La regresión lineal normal clásica supone que la distribución probabílistica de ui es normal.
La suposición de normalidad permite utilizar las pruebas estadísticas t, F, χ^2

Consideremos el ejemplo consumo e ingreso.

$$\hat{Y}_i = 24{,}4545 + 0{,}5091 X_i$$

Obtuvimos que la PMC estimada es de 0,5091, correspondiente a una sola estimación puntual de la PMC de la poblacional desconocida.

¿Qué tan confiable es esta estimación?.

Debido a fluctuaciones muestrales, es posible que una sola estimación difiera del valor verdadero, aunque en un muestreo repetido se espera que su valor medio sea igual al valor verdadero $E(\hat{\beta}_2) = \beta_2$

Estimación de intervalos

Ahora, en estadística, la confiabilidad de un estimador puntual se mide por su error estándar. Por consiguiente, en lugar de depender de un solo estimador puntual, se puede construir un intervalos alrededor del estimador puntual, por ejemplo, dentro de dos o tres errores estándar a cada lado del estimador puntual, tal que este intervalo tenga, digamos, 95% de probabilidad de incluir el verdadero valor del parámetro. Esta es la idea básica de la estimación de intervalos.

141

Estimación de intervalos

Consideremos el ejemplo hipotético consumo-ingreso. La ecuación

$$\hat{Y}_i = 24{,}4545 + 0{,}5091 X_i$$

muestra que la propensión marginal a consumir (PMC) estimada es 0,5091, la cual constituye una única estimación (puntual) de la PMC poblacional desconocida β_2 que es un (punto) estimado de la población desconocida PMC β_2 .

¿Qué tan confiable es esta estimación?
Debido a las fluctuaciones muestrales, es probable que una sola estimación difiera del valor verdadero, aunque en un muestreo repetido se espera que el valor de su media sea igual al valor verdadero (Nota: $E(\hat{\beta}_2) = \beta_2$)

142

Estimación de intervalos

Ahora, en estadística, la confiabilidad de un estimador puntual se mide por su error estándar. Por consiguiente, en lugar de depender de un solo estimador puntual, se puede construir un intervalo alrededor del estimador puntual, por ejemplo, dentro de dos o tres errores estándar a cada lado del estimador puntual, tal que este intervalo tenga, digamos, 95% de probabilidad de incluir el verdadero valor del parámetro. Ésta es, a grandes rasgos, la idea básica de la estimación de intervalos.
Para ser más específico, supóngase que se desea encontrar qué tan" cerca" está por ejemplo, $\hat{\beta}_2$ de β_2

Con este fin, tratamos de encontrar dos números positivos, δ y α, este último situado entre 0 y 1, tal que la probabilidad de que el intervalo aleatorio $\left(\hat{\beta}_2 - \delta,\ \hat{\beta}_2 - \delta\right)$ contenga el verdadero β_2 sea 1 - α.

143

Estimación de intervalos

Simbólicamente

$$\Pr\left(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta\right) = 1 - \alpha$$

Tal intervalo, si existe, se conoce como intervalo de confianza; a 1 - α se le denomina coeficiente de confianza; y α (0 < α < 1) se conoce como el nivel de significancia.
Los puntos extremos del intervalo de confianza se conocen como límites de confianza (también denominados valores *críticos),* siendo $\hat{\beta}_2 - \delta$ el *límite* de confianza inferior y $\hat{\beta}_2 + \delta$ el *límite* de confianza superior.
Obsérvese que en la práctica α y 1 - α son expresados frecuentemente en forma porcentual como 100α y 100(1 -α *)%.*

144

Intervalos de confianza para los coeficientes de regresión β_1 y β_2

Intervalo de confianza de 100(1-α) por ciento para β

$$\hat{\beta} \pm t_{\alpha/2} se(\hat{\beta})$$

Al regresar a nuestro ejemplo ilustrativo de consumo e ingreso encontramos que

$$\hat{\beta}_2 = 0{,}5091 \qquad se(\hat{\beta}_2) = 0{,}0357$$

Si suponemos que que α =5%, es decir un coeficiente de confianza del 95% entonces la tabla t muestra que para 8 gl, el t crítico es $t_{0{,}025} = 2.306$

145

Intervalos de confianza para los coeficientes de regresión β_1 y β_2

Al sustituir esos valores se obtiene que el intervalo de confianza del 95% para β2 es el siguiente:

$$0{,}4268 \leq \beta_2 \leq 0{,}5914$$

La interpretación de este intervalo de confianza es: dado un coeficiente de confianza del 95%, a largo plazo, en 95 de cada cien casos, intervalos como (0,4268 ; 0,5914) contendrán el verdadero β_2 .

Como se advirtió antes, obsérvese que no se puede decir que la probabilidad de que el intervalo específico (0,4268 ; 0,5914) contenga el verdadero β_2 . de 95% porque este intervalo es ahora fijo y no aleatorio;por consiguiente β_2 se encontrará o no dentro de él.

146

Intervalos de confianza para los coeficientes de regresión β_1 y β_2

Para el ejemplo consumo-ingreso, el intervalo de confianza para β_1 al 95% es:

$$9{,}6643 \leq \beta_1 \leq 39{,}2448$$

Utilizando

$$\hat{\beta} \pm t_{\alpha/2} se(\hat{\beta})$$

Se tiene $24{,}4545 \pm 2{,}306(6{,}4138)$

Se debe ser cauteloso al interpretar el intervalo de confianza (9,6643; 39,2448). A largo plazo, en 95 de cada 100 casos, intervalos como (9,6643; 39,2448) contendrán el verdadero β_1; la probabilidad de que este intervalo fijo incluya el verdadero β_1 es 1 o 0

147

Prueba de hipótesis. Prueba t

La idea fundamental detrás de las pruebas de significancia consiste en utilizar un estadístico de prueba (estimador).
Bajo el supuesto de normalidad la variable

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)}$$

sigue la distribución t con N-2 grados de libertad. Si el valor verdadero de $\beta 1$ se especifica en la hipótesis nula, el valor t puede calcularse fácilmente a partir de la muestra disponible, pudiendo servir por tanto como estadístico de prueba

148

Prueba de hipótesis. Prueba t

Consideremos nuevamente el ejemplo de consumo -ingreso. Sabemos que

$$\hat{\beta}_1 = 0{,}5091 \qquad \text{se}(\hat{\beta}_1) = 0{,}0357$$

Si Ho: $\beta 1 = 0{,}3$ y H1: $\beta 1 \neq 0{,}3$

$$t = \frac{0{,}5091 - 0{,}3}{0{,}0357} = 5{,}86$$

$$Si \ \alpha = 5\%,,, \text{gl} = 8 \quad \text{entonces} \quad t_{0{,}025} = 2.306$$

luego el t calculado es mayor al t de tabla y por lo tanto se rechaza la hipótesis nula

El procedimiento anterior se denomina prueba t. En el lenguaje de pruebas de significancia, se dice que un estadístico es estadísticamente significativo si el valor del estadístico de prueba se encuentra en la región crítica. En nuestro ejemplo, el estadístico t es significativo y procedemos a rechazar la hipótesis nula.

149

Aplicación problema de predicción

Con base en los datos muestrales, se obtuvo la siguiente regresión muestral.

$$\hat{Y}_i = 24{,}4545 + 0{,}5091 X_i$$

Donde $\hat{Y}_t$ es el estimador del verdadero $E(Y_i)$ correspondiente a X dada. ¿Qué uso se puede dar a esta regresión histórica? Uno es "predecir" o "pronosticar" el gasto de consumo futuro Y correspondiente a algún nivel dado de ingreso X.

Ahora, hay dos clases de predicciones:

1) la predicción del valor de la media condicional de Y correspondiente a un valor escogido X, por ejemplo, que es el punto sobre la recta de regresión poblacional misma, y
2) predicción de un valor individual Y correspondiente a X_0 .

Se llamarán estas dos predicciones de predicción media y la predicción individual.

150

Aplicación problema de predicción

Supóngase que Xo = 100 y se desea predecir
E(Y I Xo = 100). Ahora, puede demostrarse que la regresión histórica

$$\hat{Y}_i = 24{,}4545 + 0{,}5091 X_i$$

proporciona la estimación puntual de esta predicción media de la siguiente forma:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$
$$= 24.4545 + 0.5091(100) = 75.3645$$

Donde $\hat{Y}_0$ = estimador de *E(Y* I Xo). Puede demostrarse que este predictor puntual es el mejor estimador lineal e insesgado (MELI).

Puesto que $\hat{Y}_0$ es un estimador, es probable que éste sea diferente de su verdadero valor. La diferencia entre los dos valores dará alguna idea sobre el error de predicción o de pronóstico.

151

Aplicación problema de predicción

se demuestra que en la ecuación $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$

$\hat{Y}_0$ está normalmente distribuida con media $\beta_1 + \beta_2 X_0$ y con una varianza dada por la siguiente fórmula:

$$\text{var}(\hat{Y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right)$$

Al reemplazar σ^2 desconocida por su estimador insesgado se cumple que la variable

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_2 X_0)}{ee(\hat{Y}_0)}$$

sigue una distribución *t* con *n* - 2 g de l. La distribución *t* puede ser utilizada por consiguiente para construir intervalos de confianza para el verdadero *E(Yo* I Xo) y para hacer pruebas de hipótesis acerca de tal valor de la manera usual, a saber,

$$\left(\hat{\beta}_1 + \hat{\beta}_2 X_0 \right) \pm t_{\alpha/2} ee(\hat{Y}_0)$$

152

Aplicación problema de predicción

Para los datos del ejemplo (tabla 3.3 anexo 1)

$$\text{var}(\hat{Y}_0) = 42.159\left[\frac{1}{10} + \frac{(100-170)^2}{33000}\right] = 10.4759$$

$$y \quad \text{ee}(\hat{Y}_0) = 3.2366$$

Por consiguiente, el intervalo de confianza al 95% para el verdadero $E(Y / X_0) = \beta_1 + \beta_2 X_0$

es $67.9010 \leq E(Y / X = 100) \leq 82.8381$

Por tanto, dada X_0 =100, en muestreo repetido, en 95 de cada 100 intervalos como el anterior estará incluido el verdadero valor medio; la mejor estimación del verdadero valor medio es, por supuesto, la estimación puntual 75.3645

153

Predicción individual

Si nuestro interés está en predecir un valor individual Y, Y_0 correspondiente a un valor dado X, digamos X_0, entonces el mejor estimador lineal insesgado de Y_0 está dado también por

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 \qquad \textbf{(17)}$$
$$= 24.4545 + 0.5091(100) = 75.3645$$

Pero su varianza es la siguiente

$$\text{var}(Y_0 - \hat{Y}_0) = E(Y_0 - \hat{Y}_0)^2 = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right) \qquad \textbf{(18)}$$

Puede demostrarse además que Y_0 también sigue una distribución normal con media y varianza dadas por (17) y (18), respectivamente. Sustituyendo $\hat{\sigma}^2$ desconocida por σ^2 se cumple que

$$t = \frac{Y_0 - \hat{Y}_0}{ee(Y_0 - \hat{Y}_0)}$$ también sigue una distribución t

154

Predicción individual

Por consiguiente, la distribución *t* puede utilizarse para hacer inferencia sobre la verdadera Y_o. Al continuar con nuestro ejemplo consumo-ingreso, se ve que la predicción puntual de Y_o es 75.3645, igual a $\hat{Y}_0$ y su varianza es 52.6349. Por consiguiente, el intervalo de confianza al 95% para Y_o correspondiente a Xo =100 es

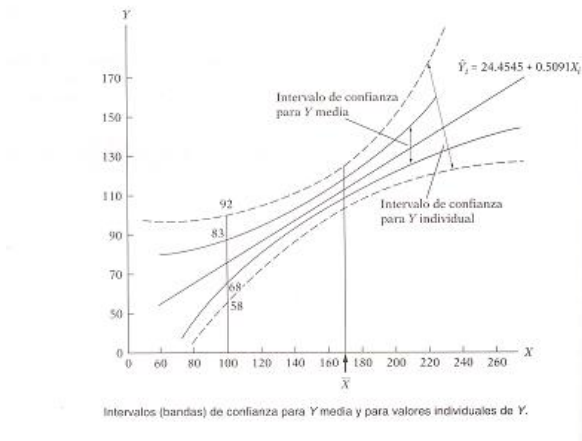
$$(58.6345 \leq Y_0 / X_0 = 100) \leq 92.0945)$$

Comparando este intervalo con

$$67.9010 \leq E(Y / X = 100) \leq 82.8381$$

Se ve que el intervalo de confianza para el Y_0 individual es más amplio que el intervalo para el valor medio de Y_0

Intervalos de confianza para Y media y para valores individuales de Y



Intervalos (bandas) de confianza para *Y* media y para valores individuales de *Y*.

Formas funcionales de los modelos de regresión

Consideremos algunos modelos de regresión que pueden ser no lineales en las variables pero que son lineales en los parámetros o que pueden serlo mediante transformaciones apropiadas de las variables.

En particular, consideremos los modelos de regresión:
1. El modelo log-lineal
2. Modelos semilogarítmicos

157

Cómo medir la elasticidad: Modelo Log-Lineal

Considérese el siguiente modelo, conocido como el modelo de regresión exponencial:

$$Y_i = \beta_1 \, \mathrm{X_i}^{\beta_2} \mathrm{e}^{\mu_i}$$

El cual puede ser expresado alternativamente

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + \mu_i$$

Si escribimos como

$$\ln Y_i = \alpha + \beta_2 \ln X_i + \mu_i$$

Donde $\alpha = \ln \beta_1$ este modelo es lineal en los parámetros α y β_2 y lineal en los logaritmos de las variables Y y X y puede ser estimado por regresión MCO

158

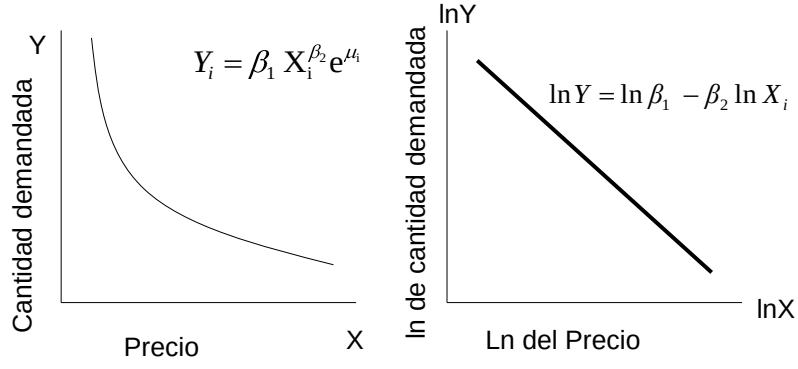
Cómo medir la elasticidad: Modelo Log-Lineal

Una característica importante del modelo log-log, que lo ha hecho muy popular en el trabajo empírico, es que el coeficiente de la pendiente β_2 mide la elasticidad de *Y* con respecto a X, es decir, el cambio porcentual en *Y* ante un pequeño cambio porcentual en X dado. Así, si Yrepresenta la cantidad demandada de un bien y X su precio unitario, β_2 mide la elasticidad-precio de la demanda, un parámetro de gran interés en economía.

159

Modelo de elasticidad constante

Si la relación entre la cantidad demandada y el precio es como se muestra en la figura *(a)* la transformación doble-log presentada en la figura (b) dará entonces la estimación de la elasticidad-precio (-β_2)



160

Ejemplo
Gasto en bienes duraderos respecto al gasto de consumo personal total

Consideremos datos sobre el gasto de consumo personal total (GCPERT), el gasto en bienes duraderos (GASBD), el gasto en bienes perecederos (GASBPER) y el gasto en servicios (GASERV), todos medidos en millones de dólares de 1992. (tabla 6.3-Anexo 1)

Su póngase que se desea calcular la elasticidad del gasto en bienes durables respecto al gasto de consumo personal total. Al graficar el logaritmo del gasto en bienes durables en comparación con el logaritmo del gasto de consumo personal total, se observará que la relación entre las dos variables es lineal. Por tanto, el modelo del doble logaritmo podría resultar adecuado. Los resultados de la regresión son:

161

$$\ln GASBD = -9.6971 + 1.9056 \ln GCPERT,$$
$$ee = (0.4341) \quad (0.0514)$$
$$t = (-22.3370)^* \quad (37.0962)^* \quad r^2 = 0.9849$$

donde * indica que el valor *p* es extremadamente pequeño.

Todos estos resultados muestran que la elasticidad de GASBD respecto a GCPERT es de casi 1.90, lo que sugiere que si el gasto personal total aumenta 1 %, en promedio, el gasto en bienes duraderos se incrementa casi 1.90%. En consecuencia, el gasto en bienes duraderos es muy sensible a los cambios en el gasto de consumo personal. Ésta es una razón por la que los productores de bienes duraderos siguen muy de cerca los cambios en el ingreso personal y el gasto de consumo personal.

162

Ejemplo: Salario y ventas

Podemos estimar un modelo de elasticidad constante que relacione el salario del director ejecutivo con las ventas de la empresa. Sea vtas las ventas anuales de la compañía, medidas en millones de dólares. Un modelo de elasticidad constante es

$$\ln(sala) = \beta_0 + \beta_1 \ln(vtas) + u$$

en el que es la elasticidad de sala en relación con vtas. Este modelo se encuentra entre los de regresión simple, al definir la variable dependiente como $y = \log(sala)$ y la independiente como $x = \log(vtas)$. La estimación de esta ecuación mediante MCO da

$$\widehat{\ln(sala)} = 4.822 + 0.257 \ln(vtas)$$
$$n = 209, R^2 = 0.211.$$

El coeficiente de ln(vtas) es la elasticidad estimada de sala con respecto a vtas. Implica que un incremento de uno por ciento en las ventas de la compañía aumenta el salario del director ejecutivo en alrededor de 0.257 por ciento, que es la interpretación usual de elasticidad.

163

Cómo medir la tasa de crecimiento: Modelo Log-Lin

Los economistas, la gente de negocios y los gobiernos frecuentemente están interesados en encontrar la tasa de crecimiento de ciertas variables económicas. tales como población, PNB, oferta monetaria, empleo, productividad, déficit comercial. etc.

Supóngase que se desea saber la tasa de crecimiento del gasto de consumo personal en servicios.

Sea *Y,* el gasto real en servicios en el tiempo *t,* y *Yo* el valor inicial del gasto en servicios.

Recordemos la muy conocida fórmula del interés compuesto, vista en los cursos básicos de economía.

$$Y_i = Y_0 (1+\mathrm{r})^{\mathrm{t}} \qquad \textbf{(1)}$$

Donde r es la tasa de interés compuesta de Y

164

Cómo medir la tasa de crecimiento: Modelo Log-Lin

Tomando el logaritmo natural, podemos escribir

$$\ln Y_i = \ln Y_0 + \mathrm{t}\,\ln(1+\mathrm{r}) \qquad \textbf{(2)}$$

Ahora sea $\beta_1 = \ln \mathrm{Y}_0 \qquad \beta_2 = \ln(1+\mathrm{r})$

Se puede escribir (2) así

$$\ln Y_i = \beta_1 + \beta_2\,\mathrm{t} \qquad \textbf{(3)}$$

Agregando el término de perturbación, se obtiene

$$\ln Y_i = \beta_1 + \beta_2\,\mathrm{t} + \mu_\mathrm{i} \qquad \textbf{(4)}$$

Este modelo es igual a cualquier otro modelo de regresión lineal en el sentido de que los parámetros β_1 y β_2 son lineales. La única diferencia es que la variable dependiente o regresada es el logaritmo de *Y y* el regresor o variable explicativa es el "tiempo", que adquiere valores de 1 ,2,3, etc.

165

Cómo medir la tasa de crecimiento: Modelo Log-Lin

Modelos como $\ln Y_i = \beta_1 + \beta_2\,\mathrm{t} + \mu_\mathrm{i}$

se denominan modelos **semilog** porque solamente una variable (en este caso la regresada) aparece en forma logarítmica.
Para fines descriptivos, un modelo en el cual la variable regresada es logarítmica se denominará modelo log-lin.

En este modelo *el coeficiente de la pendiente mide el cambio proporcional constante* o *relativo en Y para un cambio absoluto dado en el valor del regresor* (en este caso la variable t)*,* es decir;

$$\beta_2 = \frac{\text{cambio relativo en Y}}{\text{cambio absoluto en X}}$$

166

Cómo medir la tasa de crecimiento: Modelo Log-Lin

Si se multiplica el cambio relativo en Y por 100, β_2 nos dará entonces el cambio porcentual, o *la tasa de crecimiento,* en *Y* ocasionada por un cambio absoluto en X, el regresor.

Es decir, 100 por β_2 da como resultado la tasa de crecimiento en *Y;* 100 por β_2 se conoce en la literatura como la semielasticidad de *Y* respecto a X.

167

Cómo medir la tasa de crecimiento: Modelo Log-Lin

Ejemplo: Para ilustrar el modelo de crecimiento

$$\ln Y_i = \beta_1 + \beta_2 \, t + \mu_i$$

consideremos los datos sobre el gasto en servicios proporcionados en (tabla 6.3-Anexo 1). Los resultados de la regresión son los siguientes:

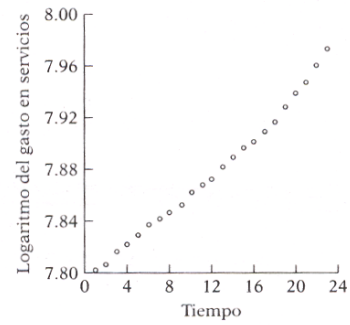
$$\begin{aligned} \ln GES_t &= 7.7890 + 0.00743t \\ ee &= (0.0023) \quad (0.00017) \\ t &= (3.387.619)^* \quad (44.2826)^* \quad r^2 = 0.9894 \end{aligned}$$

Nota: GES significa gasto en servicios y el asterisco (*) denota que el valor p es extremadamente pequeño.

168

Cómo medir la tasa de crecimiento: Modelo Log-Lin

La interpretación de la ecuación es que durante un periodo de un trimestre, el gasto en servicios se incrementó a una tasa (trimestral) de 0.743%. Aproximadamente esto es igual a un crecimiento anual de 2.97%. Puesto que 7.7890 = ln(GES) al comienzo del periodo de análisis, si se toma su antilogaritmo se tiene 2.41390 (billones de dólares), como el valor inicial de GES (es decir, el valor al final del último trimestre de 1992).



169

Cómo medir la tasa de crecimiento: Modelo Log-Lin

Ejemplo: Salario y educación

Recuerde el ejemplo del salario y la educación, en el que hicimos la regresión del salario por hora sobre los años de escolaridad. Obtuvimos una estimación de la pendiente de 0.54, que significa que pronosticamos que cada año adicional de instrucción aumenta en promedio el salario por hora en 54 centavos de dólar.

A causa del carácter lineal de

$$\widehat{sala} = -0.90 + 0.54\ educ.$$

0, 54 centavos es el incremento tanto para el primer año como para el vigésimo, lo que acaso no sea razonable.

Ahora, consideremos ln(sala) como la variable dependiente, obtenemos la siguiente relación:

$$\widehat{\ln(sala)} = 0.584 + 0.083\ educ$$
$$n = 526,\ R^2 = 0.186$$

170

Cómo medir la tasa de crecimiento: Modelo Log-Lin

El coeficiente de *educ* tiene una interpretación porcentual cuando se multiplica por 100: sala aumenta 8.3 por ciento por cada año adicional de escolaridad. Es lo que entienden los economistas cuando se refieren al "rendimiento de otro año de estudios".

Es importante recordar que la principal razón para tomar el logaritmo de sala es imponer un efecto porcentual constante de la educación en sala.

La intercepción no es muy significativa, ya que da el log(sala) pronosticado cuando educ = 0. La R cuadrada muestra que educ explica alrededor de 18.6 por ciento de la variación en log(sala) (que no es sala).

171

El modelo Lin-Log

A diferencia del modelo de crecimiento recién estudiado, en el cual se estaba interesado en encontrar el crecimiento porcentual en Y, ante un cambio unitario absoluto en X, ahora hay interés en encontrar el cambio absoluto en Y debido a un cambio porcentual en X. Un modelo que puede lograr este propósito puede escribirse como

$$Y_i = \beta_1 + \beta_2 \ln X_i + \mu_i$$

Para fines descriptivos, llamamos a este modelo un modelo lin-log.

172

El modelo Lin-Log

Interpretación de la pendiente

$$\beta_2 = \frac{cambio\ en\ Y}{cambio\ en\ \ln\ X} = \frac{cambio\ en\ Y}{cambio\ relativo\ en\ X}$$

Simbólicamente, se tiene $\beta_2 = \dfrac{\Delta Y}{\Delta X / X}$

En forma equivalente $\Delta Y = \beta_2 (\Delta X / X)$

Esta ecuación plantea que el cambio absoluto en Y (= ΔY) es igual a la pendiente multiplicada por el cambio relativo en X.

173

El modelo Lin-Log

Si este último es multiplicado por 100 entonces

$$\Delta Y = \beta_2 (\Delta X / X)$$

da el cambio absoluto en Y ocasionado por un cambio porcentual en X. Así, si ΔX/X cambia en 0.01 unidades (o 1%), el cambio absoluto en Y es 0.01(β_2).

Por tanto, si en una aplicación se encuentra que β_2 = 500, entonces el cambio absoluto en Y es (0.01)(500), o 5.0.

Por consiguiente, cuado se utiliza MCO para estimar regresiones como en

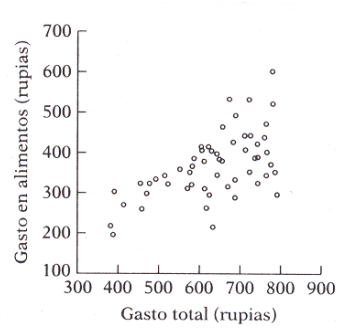
$$Y_i = \beta_1 + \beta_2 \ln X_i + \mu_i$$

se debe multiplicar el valor del coeficiente de la pendiente estimado, β_2 por 0.01 o, dividido entre 100.

174

El modelo Lin-Log

Ejemplo:
Como ejemplo del modelo lin-log, consideremos el gasto alimenticio en India, (tabla 2.8-Anexo 1). Si se grafican los datos, se obtiene la gráfica de la figura . Tal y como esta figura sugiere, el gasto alimenticio se incrementa en forma más lenta, conforme el gasto total aumenta, lo cual quizá proporcione sustento a la ley de Engels.



Nota: ¿Cuándo resulta útil un modelo lin-log ? Se ha encontrado una interesante aplicación en los así conocidos modelos de gasto Engel [nombrados en honor del estadístico alemán Emst Engel (1821-1896). Engel postuló que "el gasto total que se dedica a los alimentos tiende a incrementarse en progresión aritmética, mientras que el gasto total aumenta en progresión geométrica.

175

El modelo Lin-Log

Los resultados de ajustar el modelo lin-log a los datos son los siguientes:

$$GASAL_i = -1\,283.912 + 257.2700 \ln GASTOT_i$$
$$t = (-4.3848)^* \quad (5.6625)^* \quad r^2 = 0.3769$$

Interpretado de la forma antes descrita, el coeficiente de la pendiente, que vale casi 257, significa que un incremento en el gasto total en alimentos de 1%, en promedio, propicia un incremento de casi 2.57 rupias en el gasto en alimento de las 55 familias incluidas en la muestra. (Nota: se dividió el coeficiente estimado de la pendiente entre 100.)

176

Análisis de regresión múltiple

- Análisis de regresión múltiple: problema de la estimación
  - Notación y supuestos
  - Interpretación de la ecuación de regresión múltiple
  - Significado de los coeficientes de regresión parcial
  - Estimación MCO de los coeficientes de regresión parcial
  - El coeficiente de determinación múltiple R2
  - El coeficiente de correlación múltiple R
- Análisis de regresión múltiple: el problema de la inferencia
  - El supuesto de normalidad
  - Prueba de hipótesis en regresión múltiple
    - Prueba t para coeficientes individuales
    - Prueba F de significación global
- Modelos de regresión con variables dicotómicas
- Problemas en el análisis de regresión
- Estimación ponderada

177

Análisis de regresión múltiple

El modelo de dos variables, con frecuencia es inadecuado en la práctica. Es el caso del ejemplo consumo-ingreso, en donde se supuso implícitamente que solamente el ingreso *X* afecta el consumo *Y.* Pero la teoría económica rara vez es tan simple, ya que, además del ingreso, existen muchas otras variables que probablemente afectan el gasto de consumo.

Por consiguiente, se necesita ampliar el modelo simple de regresión con dos variables para considerar modelos que contengan más de dos variables.

La adición de variables conduce al análisis de los modelos de regresión múltiple, es decir, a modelos en los cuales la variable dependiente, o regresada, *Y,* depende de dos o más variables explicativas, o regresoras.

178

Modelo de tres variables

Generalizando la función de regresión poblacional (FRP) de dos variables se puede escribir la FRP de tres variables así:

$$Y_i = \beta_1 + \beta_2 \mathrm{X}_{2\mathrm{i}} + \beta_3 \mathrm{X}_{3\mathrm{i}} + \mu_{\mathrm{i}}$$

donde *Y es* la variable dependiente, X_2 y X_3 las variables explicativas (o regresoras). u_i es el término de perturbación estocástica, e i la iésima observación.
Los coeficientes se denominan coeficientes de regresión parcial

Se continúa operando dentro del marco del modelo clásico de regresión lineal (MCRL).

179

Modelo de tres variables

Supuestos

Específicamente. se supone lo siguiente

- Valor medio de ui, igual a cero

$$E(u_i / X_{2i}, X_{3i}) = 0 \quad \text{para} \quad \text{cada} \quad \mathrm{i}$$

- No correlación serial

$$\mathrm{cov}(u_i, u_j) = 0 \quad \mathrm{i} \neq \mathrm{j}$$

- Homocedasticidad

$$\mathrm{var}(u_i) = \sigma^2$$

180

Supuestos

- Covarianza entre ui y cada variable X igual a cero

$$\text{cov}(u_i, X_{2i}) = \text{cov}(u_i, X_{3i}) = 0$$

- No hay sesgo de especificación
  El modelo está especificado correctamente
- No hay colinealidad exacta entre las variables X

No hay relación lineal exacta entre X_2 y X_3

Adicionalmente, se supone que el modelo de regresión múltiple es lineal en los parámetros, que los valores de las regresoras son fijos en muestreos repetido y que hay suficiente variabilidad en dichos valores..

181

Interpretación de la ecuación de regresión múltiple

Dados los supuestos del modelo de regresión clásico, se cumple que, al tomar la esperanza condicional de *Y* *a* ambos lados de

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i$$

se obtiene

$$E(Y_i / X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

Expresado en palabras, de la expresión anterior se obtiene la media condicional o el valor esperado de *Y* condicionado a los valores dados o fijos de las variables X_2 y X_3. Por consiguiente, igual que en el caso de dos variables, el análisis de regresión múltiple es el análisis de regresión condicional, sobre los valores fijos de las variables explicativas, y lo que obtenemos es el valor promedio o la media de *Y,* o la respuesta media de *Y* a valores dados de las regresoras X.

Nota: Las propiedades de los estimadores MCO del modelo de regresión múltiples son similares a aquellas del modelo con dos variables

182

Significado de los coeficientes de regresión parcial

Los coeficientes de regresión β_2 y β_3 se denominan coeficientes de regresión parcial.

β_2 mide el cambio en el valor de la media de Y, E(Y) por unidad de cambio en X_2 permaneciendo X_3 constante.

β_3 mide el cambio en el valor medio de Y, E(Y) por unidad de cambio en X_3 cuando el valor de X_2 se conserva constante.

183

El coeficiente de determinación múltiple R^2

En el caso de tres variables nos gustaría conocer la proporción de la variación en Y explicada por las variables X y X conjuntamente. La medida que da esta información es conocida como el coeficiente de determinación múltiple y se denota por R^2;conceptualmente se asemeja a r^2.

$$R^2 = 1 - \frac{SRC}{STC} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

R^2, al igual que r^2, se encuentra entre 0 y 1.

Se dice que el ajuste del modelo es “mejor” entre más cerca esté R^2 de 1

184

El coeficiente de correlación múltiple R

Recuérdese que en el caso de dos variables, se definió *r* como el coeficiente de correlación y se indicó que éste mide el grado de asociación (lineal) entre las dos variables.
El análogo de *r* para tres o más variables es el coeficiente de correlación múltiple, denotado por *R,* el cual es una medida del grado de asociación entre *Y* y todas las variables explicativas conjuntamente.
Aun cuando r puede ser positivo o negativo, *R* siempre se considera positivo. En la práctica, sin embargo, *R* tiene poca importancia. La medida de mayor significado es R^2.

185

Ejemplo: Mortalidad Infantil respecto al PIB per cápita y a la tasa de alfabetización en las mujeres

Consideremos como ejemplo el comportamiento de la mortalidad infantil (MI) en relación con el PIB *per cápita* (PIBPC) y el alfabetismo femenino medido por la tasa de alfabetización en las mujeres (TAM). A *priori.* se espera que la TAM también ejerza un impacto negativo en la MI al igual que el PIBPC. cuando se introducen ambas variables en el modelo, se requiere eliminar la influencia neta de cada regresora. Es decir. se necesita estimar los coeficientes de regresión (parcial) de cada regresora. Por lo tanto. el modelo es:

$$MI_i = \beta_1 + \beta_2 PIBPC_i + \beta_3 TAM_i + u_i$$

186

Ejemplo: Mortalidad Infantil respecto al PIB per cápita y a la tasa de alfabetización en las mujeres

Téngase en cuenta que MI es el número de muertes de niños menores de 5 años por cada 1 000 nacimientos vivos. el PIBPC es el PIB *per cápita* en 1980 y que la TAM se mide en porcentaje. La muestra se realizó en 64 países.
Utilizando un paquete estadístico se obtienen los siguientes resultados:

$$MI_i = 263{,}6416 - 0{,}0056 PIBPC_i - 2{,}2316 TAM_i$$

$$ee = \quad (11{,}5932) \quad (0{,}0019)$$

$$R^2 = 0{,}7077 \qquad \overline{R^2} = 0{,}6981$$

187

Interpretación

El coeficiente de regresión parcial -0,0056 del PIBPC indica que si la influencia de la TAM se mantiene constante, conforme el PIBPC se incrementa, digamos en un dólar, en promedio, la mortalidad infantil disminuye en 0.0056 unidades. Para hacerlo interpretable desde el punto de vista económico, si el PIB *per cápita* se incrementara mil dólares, en promedio, el número de muertes de niños menores de 5 años se reduciría a 5.6 por cada 1000 nacimientos vivos.
El coeficiente -2.2316 señala que si la influencia del PIBPC se mantiene constante, el número de muertes de niños menores de 5 años disminuiría, en promedio, 2.23 por cada mil nacimientos vivos, en tanto que la tasa de alfabetización en las mujeres subiría un punto porcentual. El valor de la intersección de casi 263, si se interpretara de una forma mecanicista, significaría que si los valores del PIBPC y de la TAM fuesen cero, la mortalidad infantil promedio sería de aproximadamente 263 muertes por cada mil nacimientos vivos.

188

El valor de la intersección de casi 263, si se interpretara de una forma mecanicista, significaría que si los valores del PIBPC y de la TAM fuesen cero, la mortalidad infantil promedio sería de aproximadamente 263 muertes por cada mil nacimientos vivos. Por supuesto, tal interpretación debería tomarse con mucho cuidado.
El valor de *R2* de casi 0.71 significa que casi 71 % de la variación en la mortalidad infantil se explica mediante el PIBPC y la TAM, lo cual es un gran porcentaje si se considera que el valor máximo que puede tener *R2* es 1l. De todo lo dicho hasta aquí, los resultados de la regresión tienen sentido.

189

Análisis de regresión múltiple: el problema de la inferencia

El supuesto de normalidad

Como ya se sabe, si el único objetivo es la estimación puntual de los parámetros de los modelos de regresión, será suficiente el método de mínimos cuadrados ordinarios (MCO), el cual no hace supuestos sobre la distribución de probabilidad de las perturbaciones u_i. Pero si el objetivo no sólo es la estimación sino además la inferencia, entonces, como se analizó para el modelo de regresión simple, se debe suponer que las u_i siguen alguna distribución de probabilidad.
Se supuso que las u_i seguían la distribución normal con media cero y varianza constante. Se mantiene el mismo supuesto para los modelos de regresión múltiple. Con el supuesto de normalidad , se halla que los estimadores MCO de los coeficientes de regresión parcial, son los mejores estimadores lineales insesgados (MELI)..

190

El supuesto de normalidad

Consideremos nuevamente el ejemplo de la regresión de la mortalidad infantil (MI) sobre el PIB *per cápita* (PIBP) y la tasa de analfabetismo en las mujeres (TAM) para una muestra de 64 países. Los resultados de la regresión se reproducen a continuación.

$$MI_i = 263{,}6416 - 0{,}0056 PIBPC_i - 2{,}2316 TAM_i$$
$$ee = \quad (11{,}5932) \quad (0{,}0019)$$
$$R^2 = 0{,}7077 \qquad \overline{R^2} = 0{,}6981$$

¿Qué hay respecto a la significancia estadística de los resultados observados? Considérese por ejemplo el coeficiente del PIBP (-0.0056). ¿Es estadísticamente significativo este coeficiente; es decir, es estadísticamente diferente de cero? ¿Ambos coeficientes son estadísticamente significativos?

191

Prueba de hipótesis sobre coeficientes individuales de regresión parcial

Bajo el supuesto de que $u_i \sim N(0, \sigma^2)$ entonces, se puede utilizar la prueba *t* para demostrar una hipótesis sobre cualquier coeficiente de regresión parcial *individual* .

Para ilustrar el procedimiento, considérese la regresión sobre la mortalidad infantil**.**

La hipótesis nula establece que, manteniendo X3 constante (la tasa de alfabetismo en las mujeres), el ingreso personal disponible no tiene influencia (lineal) sobre el gasto personal de consumo. Para probar la hipótesis nula, se utiliza la prueba t donde:

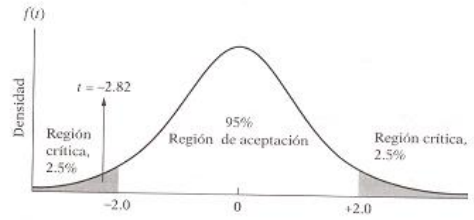
$$H_0 : \beta_2 = 0 \qquad \text{y} \qquad H_1 : \beta_2 \neq 0$$

192

Prueba de hipótesis sobre coeficientes individuales de regresión parcial

Para el ejemplo considerado se tiene

$$t = \frac{-0.0056}{0.0020} = -2.8187$$



Intervalo de confianza al 95% para *t* (60 g de l).

193

Prueba de hipótesis sobre coeficientes individuales de regresión parcial

Puesto que el valor t calculado de 2.8187 (en términos absolutos) excede el valor crítico t de 2, se puede rechazar la hipótesis nula de que el PIBP no tiene ningún efecto sobre la mortalidad infantil. Para expresarlo en términos más positivos, si se mantiene la tasa de analfabetismo para las mujeres constante, el PIB *per cápita* tiene un efecto significativo (negativo) sobre la mortalidad infantil, como se podría *esperar a priori.* De forma gráfica, la situación es la que se muestra en la figura anterior.

En la práctica, no se tiene que suponer un valor particular de α para llevar a cabo la prueba de hipótesis. Uno simplemente utiliza el valor *p* dado, que en el caso actual es de 0.0065. La interpretación de este valor *p* (es decir, el nivel exacto de significancia) es que si la hipótesis nula fuese verdadera, la probabilidad de obtener un valor t igual a 2.8187 o mayor (en términos absolutos) es de sólo 0.0065 o 0.65%. que de hecho es una probabilidad pequeña, mucho menor que el valor artificial mente adoptado de α = 5%.

194

Prueba de hipótesis sobre coeficientes individuales de regresión parcial

Existe una conexión muy estrecha entre la prueba de hipótesis y la estimación del intervalo de confianza. Para este ejemplo, el intervalo de 95% de confianza para β es

$$\hat{\beta} \pm t_{\alpha/2} se(\hat{\beta})$$

que para β_2 de este ejemplo se convierte en

$$-0.0096 \leq \beta_2 \leq -0.0016$$

o sea, el intervalo de -0.0096 a -0.0016 incluye al verdadero coeficiente β_2 con un coeficiente de confianza del 95%. Por tanto, si 100 muestras de tamaño 64 se seleccionan y 100 intervalos de confianza como el anterior se forman, entonces se espera que 95 de ellos contengan el verdadero parámetro de población β_2 .Puesto que el intervalo no incluye el valor cero de la hipótesis nula, se puede rechazar tal hipótesis (que el verdadero β_2 es cero con 95% de confianza).

195

Prueba de la significación global de la regresión

La prueba t hace referencia a la prueba de significancia individual de los coeficientes de regresión parcial estimados, es decir, bajo la hipótesis separada de que cada uno de los verdaderos coeficientes de regresión parcial de la población era cero. Pero ahora considérese la siguiente hipótesis:

$$H_0 : \beta_2 = \beta_3 = 0$$

Esta hipótesis nula es conjunta de que β_2 y β_3 son iguales a cero en forma conjunta o simultánea. Una prueba de tal hipótesis se denomina prueba de significancia global de la recta de regresión observada o estimada, es decir, si
Y está relacionada o no linealmente con X_2 y X_3 a la vez..

196

Prueba de la significación global de la regresión

La significación global de la regresión se puede probar con la relación de la varianza explicada a la varianza no explicada: Esta sigue una distribución F con k-1 y n-k grados de libertad, donde n es el número de observaciones y k es el número de parámetros estimados.

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

Si la relación F calculada excede el valor tabulado de F al nivel especificado de significación y grados de libertad, se acepta la hipótesis de que los parámetros de la regresión no son todos iguales a cero y que R cuadrado es significativamente diferente de cero.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... \beta_n = 0$$

$$H_1 : \text{No todas las } \beta \text{ son cero}$$

197

Modelos de regresión con variables dicotómicas

En el análisis de regresión, la variable dependiente o regresada, está influida frecuentemente no sólo por variables de razón de escala (por ejemplo: ingreso, producción, precios, costos, estatura y temperatura), sino también por variables que son esencialmente cualitativas por naturaleza, o de escala nominal (por ejemplo, sexo, raza, color, religión, nacionalidad, región geográfica, trastornos políticos y afiliación a un partido).

Por ejemplo, manteniendo los demás factores constantes, se ha encontrado que las trabajadoras ganan menos que sus colegas masculinos y que las personas de color ganan menos que las blancas. Este patrón puede resultar de la discriminación sexual o racial, pero cualquiera que sea la razón, las variables cualitativas tales como sexo y raza sí influyen sobre la variable dependiente y es claro que deben ser incluidas dentro de las explicativas, o regresoras.

198

Modelos de regresión con variables dicotómicas

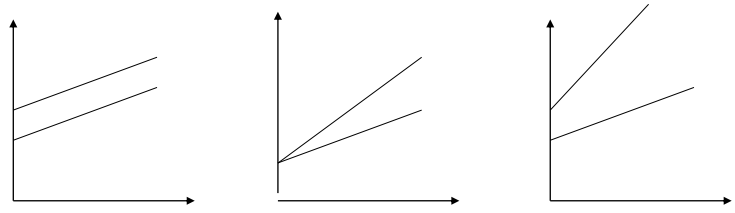
Puesto que tales variables usualmente indican la presencia o ausencia de una "cualidad" o atributo, tal como femenino o masculino, negro o blanco, católico o no católico, demócrata o republicano son variables de *escala nominal* esencialmente.
Se podrían "cuantificar" tales atributos mediante la elaboración de variables artificiales que tomaran los valores 0 y 1, donde 1 indicara la presencia (o la posesión) de ese atributo y 0 la ausencia de tal atributo. Por ejemplo, el 1 puede indicar que una persona es de sexo masculino y 0 puede designar una de sexo femenino; o el 1 puede indicar que una persona se ha graduado en la universidad y 0 que no lo ha hecho y así sucesivamente. Las variables que adquieren tales valores 0 y 1 se llaman variables dicótomas. Tales variables son, por tanto, esencialmente un recurso para clasificar datos en categorías mutuamente excluyentes, como masculino o femenino.

199

Modelos de regresión con variables dicotómicas

Las variables ficticias se pueden usar para establecer cambios en la ordenada en el origen, cambios en la pendiente y cambios tanto en la ordenada en el origen como en la pendiente.

$$Y = b_0 + b_1 X + b_2 D + u$$
$$Y = b_0 + b_1 X + b_2 XD + u$$
$$Y = b_0 + b_1 X + b_2 D + b_2 DX + u$$



200

Modelos de regresión con variables dicotómicas

Si una variable cualitativa tiene m categorías, sólo hay que agregar (m-1) variables dicotómicas

Ejemplo: Considérese el siguiente modelo:

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

donde *Y* = salario por hora en dólares
X = educación (años de escolaridad)
D2 = 1 si es mujer; 0 en otro caso
D3 = 1 si no es blanco y no hispano; 0 en otro caso

Modelos de regresión con variables dicotómicas

En este modelo el sexo y la raza son regresoras cualitativas y la escolaridad es cuantitativa. Está implícita en este modelo la suposición de que el efecto diferencial de la variable dicótoma sexo, *D2,* es constante en las dos categorías de raza y el efecto diferencial de la variable dicótoma raza, *D3,* también es constante en los dos sexos. Es decir, si el salario medio es mayor para los hombres que para las mujeres, se debe a que pertenezcan o no pertenezcan a la categoría de no hispanos ni blancos. De igual forma, si por ejemplo los no blancos ni hispanos tienen salarios medios menores, se debe a que son hombres o mujeres.

Modelos de regresión con variables dicotómicas

En muchas aplicaciones, dicha suposición puede ser insostenible. Una mujer no blanca ni hispana tal vez gane menor salario que un hombre de esa misma categoría. En otras palabras, quizá haya una interacción entre las dos variables cualitativas *D2* y *D3.* Por tanto, su efecto sobre la media quizá no sea simplemente aditivo, sino multiplicativo, como en el siguiente modelo:

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i$$

de donde

$$E(Y_i / D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i$$

Que es la función salario medio por hora para las trabajadoras no blancas ni hispanas.

203

Modelos de regresión con variables dicotómicas

Obsérvese que

α_2 = efecto diferencial de ser mujer

α_3 = efecto diferencial de ser no blanco ni hispano

α_4 = efecto diferencial de ser mujer no blanca ni hispana

lo cual muestra que el salario medio por hora de las mujeres no blancas ni hispanas es diferente (en una cantidad igual a α_4) del salario medio por hora de las mujeres blancas o hispanas. Si por ejemplo los tres coeficientes de las variables dicótomas son negativos, implicaría que las trabajadoras no blancas ni hispanas ganan un salario medio por hora mucho más bajo que las trabajadoras blancas o hispanas, si se compara con la categoría base, la cual en el ejemplo presente es la de hombres blancos o hispanos.

204

Modelos de regresión con variables dicotómicas

Ejemplo:
Ingresos promedio por hora en comparación con la escolaridad, sexo y raza

Los resultados de la regresión basados en el modelo

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

donde *Y* = salario por hora en dólares
X = educación (años de escolaridad)
D2 = 1 si es mujer; 0 en otro caso
D3 = 1 si no es blanco y no hispano; 0 en otro caso

son

$$\hat{Y}_i = -0.2610 - 2.3606 D_{2i} - 1.7327 D_{3i} + 0.8028 X_i$$
$$t = (-0.2357)^{**} \ (-5.4873)^{*} \ (-2.1803)^{*} \ (9.9094)^{*}$$
$$R^2 = 0.2032 \quad \text{n} = 528$$

205

Modelos de regresión con variables dicotómicas

donde * indica los valores *p* menores que el 5%, y ** señala los valores *p* mayores que 5%.
Los coeficientes diferenciales de la intersección son estadísticamente significativos y tienen los signos que se esperaban y la escolaridad tiene un gran efecto positivo sobre el salario por hora.
Como lo muestra la ecuación, ceteris paribus, los ingresos promedio por hora de las mujeres son inferiores por casi $2.36;además, los ingresos promedio por hora de los trabajadores no blancos ni hispanos también son menores por $1.73.

206

Modelos de regresión con variables dicotómicas

Ahora consideremos los resultados del modelo

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i$$

Que incluye la variable dicotómica de interacción.

$$\hat{Y}_i = -0.2610 - 2.3606 D_{2i} - 1.7327 D_{3i} + 2.1289 (D_{2i} D_{3i}) + 0.8028 X_i$$

$$t = (-0.2357)^{**} \quad (-5.4873)^{*} \quad (-2.1803)^{*} \quad (1.7420)^{**} \qquad (9.9095)^{**}$$

$$R^2 = 0.2032 \quad n = 528$$

donde * indica los valores *p* menores que el 5%, y ** señala los valores *p* mayores que 5%.

Como se observa las dos variables dicotómicas aditivas siguen siendo estadísticamente significativas, pero la variable dicotómica interactiva no está al nivel convencional del 5%

207

Si se considera estadísticamente significativa entonces se interpreta de la siguiente manera. Si se mantiene constante el nivel de educación y si se añaden los tres coeficientes de las variables dicotómicas entonces se obtendrá

-1,964=-2,3605-1,732+2,128

Lo cual significa que los salarios medios por hora de las trabajadoras no blancas ni hispanas es menor por casi $1,96, valor que está entre -2,3605 (diferencia sólo debida a sexo) y -1,7327 (diferencia sólo debida a la raza)

208

Problemas en el análisis de regresión

Multicolinealidad:
Se refiere al caso en el cual dos o más variables explicatorias en el modelo de regresión están altamente correlacionadas, haciendo difícil o imposible aislar sus efectos individuales sobre la variable dependiente. Con multicolinealidad, los coeficientes de MCO estimados pueden ser estadísticamente insignificantes (y aún tener el signo contrario) aunque R cuadrado puede ser alto. La multicolinealidad puede ser superada a veces o reducirse coleccionando más datos, usando información a priori, transformando la relación funcional, o reduciendo una de las variables altamente colineales.

209

Heteroscedasticidad

Si no se mantiene la suposición de MCO de que la varianza del término de error es constante para todos los valores de las variables independientes, enfrentamos el problema de la heterocedasticidad. Esto conduce a estimaciones sesgadas e ineficientes (es decir, con varianza mayor que la mínima) de los errores estándar (y así pruebas estadísticas incorrectas e intervalos de confianza también incorrectos)

210

Autocorrelación:

Cuando el término de error en un período está correlacionado positivamente con el término de error en el período anterior, enfrentamos el problema de autocorrelación (de primer orden positiva). Esto es común en análisis de series de tiempo.
La presencia de autocorrelación de primer orden se prueba utilizando la tabla del estadístico de Durbin- Watson a los niveles de significación del 5% o 1% para n observaciones y k, variables explicatorias

211

Estimación ponderada (MCP- WLS)

Los modelos de regresión lineal típicos asumen que la varianza es constante en la población objeto de estudio. Cuando éste no es el caso (por ejemplo cuando los casos con puntuaciones mayores en un atributo muestran más variabilidad que los casos con puntuaciones menores en ese atributo), la regresión lineal mediante mínimos cuadrados ordinarios (MCO, OLS) deja de proporcionar estimaciones óptimas para el modelo.

212

Estimación ponderada (MCP- WLS)

Si las diferencias de variabilidad se pueden pronosticar a partir de otra variable, el procedimiento **Estimación ponderada** permite calcular los coeficientes de un modelo de regresión lineal mediante mínimos cuadrados ponderados (MCP, WLS), de forma que se les dé mayor ponderación a las observaciones más precisas (es decir, aquéllas con menos variabilidad) al determinar los coeficientes de regresión.

Ejemplo.
¿Cuáles son los efectos de la inflación y el paro sobre los cambios en el precio de las acciones? Debido a que los valores con mayor valor de cotización suelen mostrar más variabilidad que aquellos con menor valor de cotización, la estimación de mínimos cuadrados ordinarios no generará estimaciones que sean óptimas. El método de Estimación ponderada permite capturar el efecto del precio de cotización sobre la variabilidad de los cambios en el precio, al calcular el modelo lineal. 213

Estimación ponderada (MCP- WLS)

Consideraciones sobre los datos

Datos. Las variables dependiente e independientes deben ser cuantitativas. Las variables categóricas, como la religión, la edad o el lugar de residencia, han de recodificarse como variables binarias (dummy) . **La variable de ponderación deberá ser cuantitativa** y estar relacionada con la variabilidad de la variable dependiente

Supuestos. Para cada valor de la variable independiente, la distribución de la variable dependiente debe ser normal. La relación entre la variable dependiente y cada variable independiente debe ser lineal y todas las observaciones deben ser independientes. La varianza de la variable dependiente puede cambiar según los niveles de la variable o variables independientes, pero las diferencias se deben poder pronosticar en función de la variable de ponderación.

214

Regresión Logística

- **Consideraciones sobre los datos**
- **Fases fundamentales**
- **Fundamentos Función logística**
- **Cálculo de las probabilidades pronosticadas**
- **Interpretación de los coeficientes**
- **El problema de clasificación**
- **Estadísticos: Puntuación de Rao, Chi cuadrado, Wald**
- **Regresión logística versus análisis discriminante**

215

Regresión Logística

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de variables predictoras.
Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la **variable dependiente es dicotómica**.

Los coeficientes de regresión logística pueden utilizarse para estimar la razón de las ventajas (odds ratio) de cada variable independiente del modelo.
La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante.

216

Regresión logística: Consideraciones sobre los datos

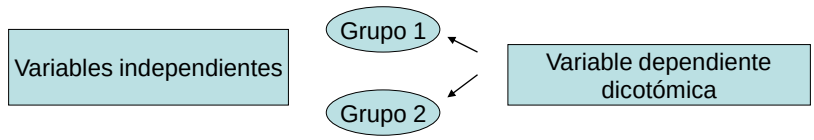
Datos. La variable dependiente debe ser dicotómica. Las variables independientes pueden estar a nivel de intervalo o ser categóricas; si son categóricas, deben ser variables dummy o estar codificadas como indicadores (existe una opción en el procedimiento para recodificar automáticamente las variables categóricas).

Supuestos. La regresión logística no se basa en supuestos distribucionales en el mismo sentido en que lo hace el análisis discriminante. Sin embargo, la solución puede ser más estable si los predictores tienen una distribución normal multivariante. Adicionalmente, al igual que con otras formas de regresión, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y a errores típicos inflados.
El procedimiento es más eficaz cuando la pertenencia a grupos es una variable categórica auténtica.

217

Regresión Logística

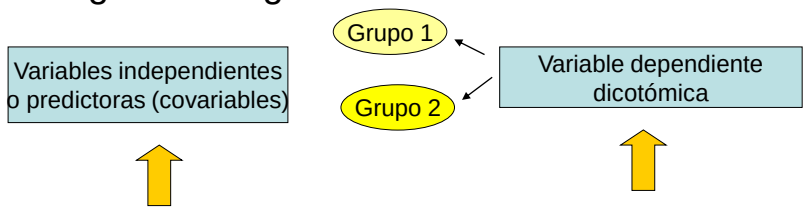
El *análisis de regresión logística* tiene como finalidad principal pronosticar la pertenencia a un grupo a partir de una serie de variables independientes.



Para llevar a cabo un análisis de regresión logística binaria es necesario disponer de una variable categórica que defina dos grupos:
Los clientes que devuelven un crédito y los que no
Los ciudadanos que votan y los que no
Los pacientes que tienen una determinada enfermedad y los que no

218

Regresión Logística



Las variables en las que se supone que se diferencian los grupos

La variable dicotómica es la variable *dependiente* del análisis, es decir, la variable cuyos valores se desea pronosticar.

El análisis de regresión logística genera una serie de pesos o coeficientes que:
(1) Informan sobre la capacidad individual de cada variable independiente para diferenciar entre los grupos.
(2) Permiten obtener pronósticos que sirven para clasificar a los sujetos

219

Fases fundamentales

Un análisis de regresión logística consta de cuatro fases fundamentales:

- La selección de las variables de análisis.
- La estimación de los pesos o coeficientes de las variables seleccionadas.
- La clasificación de los casos.
- El análisis de los residuos.

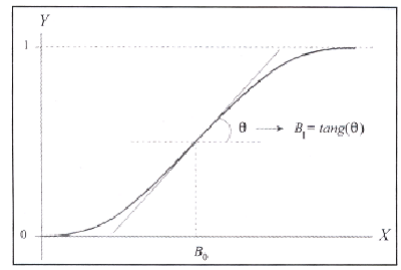
La *selección* de las variables puede realizarse a partir de criterios teóricos o puede obedecer a criterios estadísticos
La *estimación* de los pesos o coeficientes asociados a cada variable se realiza mediante una algoritmo iterativo de máxima verosimilitud.
La *clasificación* de los casos se realiza a partir de los pronósticos del modelo estimado.
El análisis de los residuos permite detectar posibles casos atípicos o predicciones anómalas.

220

Fundamentos Función Logística

$$Y = \frac{1}{1 + e^{-(B_0 + B_1 X)}}$$

Representación gráfica de una curva logística.



El coeficiente Bo representa la *posicion* de la curva sobre el eje horizontal o las abscisas (más hacia la izquierda o más hacia la derecha). Y el coeficiente B1 representa la *pendiente* de la curva medida en la zona de inflexión de la curva.

$$Y = \frac{1}{1 + e^{-(B_0 + B_1 + B_2 X_2 + \cdots + B_k X_k)}}$$

0<Y<1

El exponente del número *e* es una ecuación lineal múltiple en la que cada variable independiente recibe una ponderación proporcional a su capacidad para predecir *Y*.

221

Fundamentos Función Logística

Si dos sucesos son exclusivos entre sí (no se solapan) y exhaustivos (agotan el espacio muestral de posibles sucesos), la probabilidad de aparición de cualquiera de ellos es igual a 1 menos la probabilidad de aparición del otro.

Supongamos que la variable *Y* puede tomar sólo dos valores (0 y 1) Sea *P(Y* = 1) la probabilidad de que la variable *Y* tome el valor 1, entonces la probabilidad de que *Y* tome el valor 0 será:
P(Y = 0) = 1-*P* (*Y* = 1).

222

Ejemplo

Supongamos que interesa explicar y predecir si una persona ha votado o no en las últimas elecciones a partir de un conjunto de características socio-demográficas.
La variable que distingue a los sujetos que manifiestan haber votado de aquellos que manifiestan no haber votado es la variable *voto.*

¿Votó en 1992?

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí votó	1032	68,8	71,1	71,1
	No votó	420	28,0	28,9	100,0
	Total	1452	96,8	100,0	
Perdidos	Sistema	48	3,2		
Total		1500	100,0		

223

Ejemplo

Consideremos en primer lugar la variable lee como variable independiente (¿Lee el periódico?)

¿Lee el periódico?

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí lee	862	57,5	85,3	85,3
	No lee	148	9,9	14,7	100,0
	Total	1010	67,3	100,0	
Perdidos	Sistema	490	32,7		
Total		1500	100,0		

En el ejemplo propuesto, el fenómeno que interesa estudiar es la *abstención*, es decir, el hecho de que una persona no acuda a votar.
Una buena estimación de la probabilidad de este fenómeno es la frecuencia relativa de abstención observada en la muestra.

224

Ejemplo

Tabla de contingencia ¿Votó en 1992? * ¿Lee el periódico?

			¿Lee el periódico?		
			Sí lee	No lee	Total
¿Votó en 1992?	Sí votó	Recuento	624	68	692
		% de ¿Lee el periódico?	74,9%	47,6%	70,9%
	No votó	Recuento	209	75	284
		% de ¿Lee el periódico?	25,1%	52,4%	29,1%
Total		Recuento	833	143	976
		% de ¿Lee el periódico?	100,0%	100,0%	100,0%

En las frecuencias marginales de la tabla puede apreciarse que se ha abstenido de votar el 29,1% de los encuestados. La estimación de la probabilidad del suceso "No votó" será $P(Y = 1) = 0{,}291$. Por tanto, la probabilidad del suceso "Sí votó" será $P(Y = 0) = 0{,}709$.

225

Ejemplo

Se sabe que aproximadamente una tercera parte de los sujetos encuestados se abstiene y que, por tanto, cabe esperar que una de cada tres personas no acuda a las urnas; pero no se sabe nada acerca de las características de las personas que se abstienen.

La pregunta que interesa responder en este momento es:
¿es posible utilizar alguna otra variable, previa a la votación, que permita pronosticar adecuadamente la probabilidad de que un sujeto no vote?.
Es decir, ¿es posible construir un modelo de regresión que permita pronosticar la probabilidad de abstención a partir de una o varias variables independiente?.

226

Ecuación logística

Si existen variables capaces de predecir la abstención, entonces es posible incluirlas en un modelo de regresión y utilizarlas para corregir las estimaciones de proporción de votantes y no votantes.

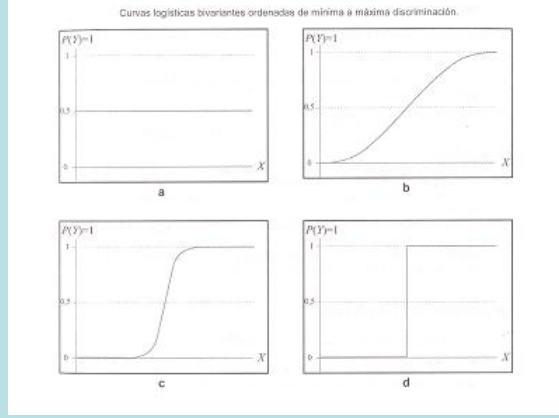
El problema que surge en una situación de estas características es que, al tener que pronosticar una probabilidad (es decir, un valor comprendido entre 0 y 1), un modelo de regresión lineal puede plantear serios problemas de predicción por no tener máximo ni mínimo teóricos en los pronósticos que arroja. Por esta razón es más apropiado recurrir a un modelo de tipo logístico. Considérese la siguiente ecuación logística:

$$P(Y=1) = \frac{1}{1+e^{-(B_0+B_1X)}}$$

227

Definida la ecuación que puede utilizarse, el objetivo consiste en encontrar una variable que discrimine bien entre los dos posibles valores de Y.

La figura muestra cuatro curvas logísticas correspondientes a cuatro posibles variables independientes o *predictoras*.



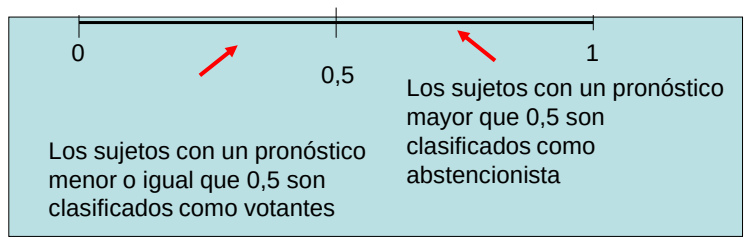
228

Las curvas se encuentran ordenadas por orden creciente de capacidad discriminativa. Puesto que el coeficiente que controla la pendiente de la curva es B1, una buena variable *predictora* será aquella que genere una curva con mucha pendiente (es decir, una variable que tenga asociado un coeficiente muy alto, en valor absoluto), mientras que una mala variable *predictora* será aquella que genere una curva sin pendiente o con muy poca pendiente (es decir, que tenga asociado un coeficiente B1 próximo a 0, en valor absoluto).

El objeto de análisis de regresión logística es encontrar las variables con mayor (en valor absoluto) coeficiente asociado.

229

Supongamos, por simplicidad, que para clasificar a un sujeto como votante o abstencionista se decide establecer como punto de corte el valor de probabilidad 0,5.



Una buena variable *predictora* (podría decirse óptima) será aquella que permita obtener pronósticos (probabilidades) iguales a 0 para el suceso Y=0 y pronósticos iguales a 1 para el suceso Y=1 .

230

Cálculo de las probabilidades pronosticadas

Utilizando los datos de la tabla (software SPSS) del ejemplo:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	LEE	1,192	,186	41,258	1	,000	3,293
	Constante	-1,094	,080	187,316	1	,000	,335

a. Variable(s) introducida(s) en el paso 1: LEE.

la ecuación de regresión logística toma la forma :

$$P(Y=1) = \frac{1}{1+e^{-(B_0+B_1X)}} = \frac{1}{1+e^{-(-1,094+1,192X)}}$$

Donde Y=0 Sí votó X=0 Sí lee
Y=1 No votó X=1 No lee

231

Cálculo de las probabilidades pronosticadas

En el caso de que un encuestado no lea el periódico , la probabilidad pronosticada por la ecuación de regresión logística para la categoría “No votó” vale:

$$P(Y=1 \mid X=1) = \frac{1}{1+e^{-[-1,094+1,192\ 1]}} = \frac{1}{1+e^{-0,098}} = 0,5245$$

Y en el caso de que un encuestado lea el periódico, la probabilidad pronosticada para la categoría “No votó” vale:

$$P(Y=1 \mid X=0) = \frac{1}{1+e^{-[-1,094+1,192\ 0]}} = \frac{1}{1+e^{1,094}} = 0,2509$$

232

Cálculo de las probabilidades pronosticadas

Por tanto, a partir de los pronósticos derivados de la ecuación de regresión logística, se puede afirmar que, entre los sujetos que manifiestan *no leer el periódico*, la probabilidad de *abstención en las elecciones* es mayor (aproximadamente el doble) que entre los sujetos que manifiestan *leer el periódico.*

Es muy importante tener en cuenta que los pronósticos obtenidos con la ecuación de regresión logística siempre se refieren a una de las dos categorías de la variable dependiente: aquella codificada con el valor mayor y que es la que el procedimiento ***Regresión logística*** *codifica internamente con el valor 1. En el ejemplo, la categoría "No votó".*

233

Interpretación de los coeficientes

¿Cómo interpretar los coeficientes de un modelo de regresión logística? Ya se ha dicho que $P(Y=0) = 1 - P(Y=1)$

Dividiendo la probabilidad de uno de los sucesos por su probabilidad complementaria y simplificando se obtiene el cuociente denominado la *ventaja* (*odds*) del suceso Y=1 frente al suceso Y=0 :

$$\frac{P(Y=1)}{P(Y=0)} = \frac{1/(1+e^{-(B_0+B_1X)})}{1-1/(1+e^{-(B_0+B_1X)})} = e^{B_0+B_1X}$$

La ventaja de un suceso es el cuociente entre la probabilidad de que el suceso ocurra y la probabilidad de que no ocurra

234

Interpretación de los coeficientes

Tomando el logaritmo neperiano de la ventaja se obtiene la transformación logit:

$$\ln\left(\frac{P\ Y=1}{P\ Y=0}\right) = B_0 + B_1 X$$

Este modelo se ajusta a un modelo de regresión lineal. Por tanto, el coeficiente de regresión de un modelo logístico puede interpretarse como **el cambio que se produce en la transformación *logit* (en el logaritmo de la *ventaja del suceso* *Y* = 1) por cada unidad de cambio que se produce en la variable independiente.**
Un coeficiente positivo debe interpretarse como un incremento en la probabilidad que el individuo tome el valor 1 debido a una variación unitaria en la variable, mientras que un valor negativo debe interpretarse como una disminución en la misma probabilidad

235

Interpretación de los coeficientes

Con los datos del ejemplo, la transformación *logit* del suceso "No votó" (*Y* = 1), cuando el encuestado "Lee el periódico" (*X = 0*) vale:

$$\ln\left(\frac{P\ Y=1 \mid X=0}{P\ Y=0 \mid X=0}\right) = B_0 = \ln\left(\frac{0,2509}{1-0,2509}\right) = \ln\ 0,335 = -1,094$$

Y la transformación *logit* del suceso "No votó" cuando el encuestado "No lee el periódico" (*X* = 1) vale:

$$\ln\left(\frac{P\ Y=1 \mid X=1}{P\ Y=0 \mid X=1}\right) = B_0 + B_1 = \ln\left(\frac{0,5245}{1+0,5245}\right) = \ln\ 1,103 = 0,098$$

Por tanto, la diferencia entre ambos logaritmos permite obtener el valor del coeficiente:

$$B_1 = 0,098 - -1,094 = 1,192$$

236

Interpretación de los coeficientes

Así, en el modelo de regresión logística, el coeficiente de regresión asociado a una variable independiente **representa el cambio producido en la transformación *logit* por unidad de cambio en la variable independiente.**

Es preferible interpretar directamente el cambio en las *ventajas* y no en los *logaritmos de las ventajas*. Volviendo a la expresión de la *ventaja*:

$$\frac{P\ Y=1}{P\ Y=0} = e^{B_0 + B_1 X} = e^{B_0}\, e^{B_1 X}$$

Se ve claramente que una *ventaja* se puede expresar en términos de potencias del número *e*. Por ello se suele informar del valor exponencial de los coeficientes de regresión.

237

Interpretación de los coeficientes

En los resultados de la regresión logística se incluye tanto el valor del coeficiente de regresión (*B*) como el de *Exp(B)* . En el ejemplo, la *ventaja* del suceso "No votó" cuando el encuestado "No lee el periódico" vale 1,103, mientras que la *ventaja* de ese mismo suceso cuando el encuestado "Sí lee el periódico" vale 0,335.

Si se expresa el cambio proporcional de la *ventaja* en términos de un cociente (como una razón) se obtiene 1,103/0,335 = 3,293, que es justamente el valor de Exp(B). A este cambio proporcional se le denomina *razón de las ventajas (odds ratio en inglés)*, dado que es el resultado de dividir dos *ventajas*. **Y se interpreta en términos del cambio proporcional (ya sea aumento o disminución) que se produce en la *ventaja* del suceso o evento de interés ("No vota" en el ejemplo) por cada unidad de cambio que se produce en la variable independiente (VI).**

238

El problema de la clasificación

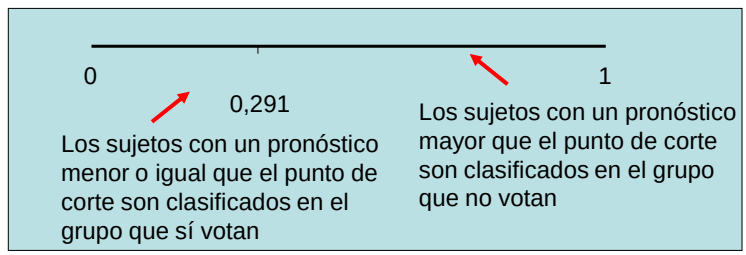
Una ecuación de regresión logística raramente arroja pronósticos con valores 0 y 1, es decir, raramente genera una curva en forma de escalón Lo habitual es encontrar que la probabilidades pronosticadas adoptan valores comprendidos entre 0 y 1.

Este es el motivo que obliga a tener que establecer un punto de corte para poder tomar la decisión de clasificar a los sujetos en uno u otro grupo a partir de las probabilidades pronosticadas.

Cuando la variable independiente es dicotómica, como en el ejemplo, establecer el punto de corte es una tarea bastante sencilla. Puesto que una variable dicotómica sólo adopta dos valores (en el ejemplo: 0 = “Si lee el periódico” y 1 = “No lee el periódico”), sólo es posible obtener dos pronósticos. (0,2509 y 0,5245).

239

También se ha visto que la probabilidad del suceso “No votó” vale 0,291 Así parece razonable pensar que el punto del corte debería encontrarse entre las dos probabilidades pronosticadas; ese punto de corte bien podría ser, por ejemplo 0,291.



240

En los modelos con más de una variable independiente se incrementa el número de valores distintos que es posible pronosticar

Existen dos caminos alternativos para determinar el punto de corte óptimo, es decir, para encontrar cuál es el valor (la probabilidad) a partir del cual se consigue diferenciar al máximo a los sujetos de uno y otro grupo y, consecuentemente, para efectuar la mejor clasificación posible.

El primero de estos caminos consiste en generar múltiples tablas de *clasificación* variando en cada una de ellas el punto de corte hasta optimizar el porcentaje de casos correctamente clasificados.

El segundo camino para determinar el punto de corte óptimo consiste en utilizar la *curva COR*.

Tablas de clasificación con distintos valores de corte

Tabla de clasificación[a]

Observado			Pronosticado		
			¿Votó en 1992?		Porcentaje correcto
			Sí votó	No votó	
Paso 1	¿Votó en 1992?	Sí votó	624	68	90,2
		No votó	209	75	26,4
	Porcentaje global				71,6

a. El valor de corte es ,500

Tabla de clasificación[a]

Observado			Pronosticado		
			¿Votó en 1992?		Porcentaje correcto
			Sí votó	No votó	
Paso 1	¿Votó en 1992?	Sí votó	624	68	90,2
		No votó	209	75	26,4
	Porcentaje global				71,6

a. El valor de corte es ,300

Tabla de clasificación[a]

Observado			Pronosticado		
			¿Votó en 1992?		Porcentaje correcto
			Sí votó	No votó	
Paso 1	¿Votó en 1992?	Sí votó	624	68	90,2
		No votó	209	75	26,4
	Porcentaje global				71,6

a. El valor de corte es ,260

Tabla de clasificación[a]

Observado			Pronosticado		
			¿Votó en 1992?		Porcentaje correcto
			Sí votó	No votó	
Paso 1	¿Votó en 1992?	Sí votó	0	692	,0
		No votó	0	284	100,0
	Porcentaje global				29,1

a. El valor de corte es ,250

Regresión logística múltiple

Nos interesa interpretar un análisis de regresión logística utilizando más de una variable independiente.

Variable dependiente

Voto

Variables independientes

Leer
Edad
Hijos
Educ
Ingfam91

Consideremos la misma variable dependiente que en el primer ejemplo (*voto*) y, además de la variable independiente allí utilizada (*leer*), otras cuatro nuevas: *edad* (Edad del encuestado), *hijos* (Número de hijos), *educ* (Años de escolarización) e *ingfam91* (Ingresos familiares en 1991)

Estadístico de puntuación de Rao

La tabla siguiente contiene los valores del ***estadístico de puntuación*** **de Rao.** Este estadístico mide la contribución *individual* de cada variable a la mejora del ajuste global del modelo. El nivel crítico (*Sig*) asociado a cada estadístico indica qué variables contribuyen significativamente al ajuste. Puede verse que, exceptuando la variable *hijos*, todas las variables incluidas en el análisis son significativas; por tanto, buenas candidatas para formar parte del modelo de regresión. La última línea, *Estadísticos globales*, contiene una valoración global de todas las variables independientes tomadas juntas.

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	LEE	45,137	1	,000
		EDAD	20,956	1	,000
		EDUC	60,910	1	,000
		INGFAM91	53,935	1	,000
		HIJOS	,188	1	,664
	Estadísticos globales		136,954	5	,000

245

El estadístico *chi-cuadrado*

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	143,754	5	,000
	Bloque	143,754	5	,000
	Modelo	143,754	5	,000

El estadístico *chi-cuadrado* permite contrastar la hipótesis de que el incremento obtenido en el ajuste global del modelo es nulo. Este estadístico sirve para determinar si, al introducir las cinco variables independientes en el modelo, se consigue un incremento significativo del ajuste global. Este incremento se valora tomando como punto de referencia el modelo nulo. Puesto que el modelo se construye en un único paso (pues se está utilizando el método *introducir*; ver siguiente apartado), todas las secciones de tabla informan del mismo valor; la mejora respecto al modelo nulo, es decir, respecto al modelo del paso 0 (*Chi-cuadrado* = 143,754). En el ejemplo, esta mejora es significativa:

246

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	970,392[a]	,143	,205

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

La tabla ofrece un resumen del modelo. Contiene tres estadísticos que permiten valorar el ajuste global del modelo en el paso 1, es decir, del modelo que incluye todas las variables.

Los parámetros están estimados a través del método de máxima verosimilitud (-2LL), de modo que sus valores tenderán a 0 si la verosimilitud tiende a máxima y al revés si ésta es baja.En el ejemplo se observa un un bajo ajuste del modelo a los datos. Este resultado queda corroborado con el estadístico de Cox y Snell, que se interpreta de la misma forma que el coeficiente de determinación de un modelo de regresión lineal.

247

Matriz de confusión

Tabla de clasificación[a]

			Pronosticado		
			¿Votó en 1992?		Porcentaje correcto
Observado			Sí votó	No votó	
Paso 1	¿Votó en 1992?	Sí votó	615	54	91,9
		No votó	189	76	28,7
	Porcentaje global				74,0

a. El valor de corte es ,500

La tabla muestra la *matriz de confusión* con los resultados de la clasificación.
Aunque no es posible mejorar el porcentaje global de clasificación correcta sin incluir nuevas variables independientes, si es posible equilibrar la tasa de aciertos en los dos grupos manipulando el punto de corte utilizado en la clasificación.

248

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	LEE	,671	,207	10,468	1	,001	1,956
	EDAD	-,034	,006	36,188	1	,000	,967
	EDUC	-,192	,032	35,276	1	,000	,825
	INGFAM91	-,056	,016	12,722	1	,000	,945
	HIJOS	,018	,051	,125	1	,723	1,018
	Constante	3,636	,526	47,706	1	,000	37,957

a. Variable(s) introducida(s) en el paso 1: LEE, EDAD, EDUC, INGFAM91, HIJOS.

La tabla muestra las estimaciones de los coeficientes (B) del modelo y los datos necesarios para valorar su significación e interpretarlos.
La significación de cada coeficiente se evalúa a partir del **estadístico de Wald. Este estadístico permite contrastar la hipótesis nula de que el coeficiente vale cero en la población y se obtiene elevando al cuadrado el cociente entre el valor estimado del coeficiente (*B*) y su error típico (*error tip*).**

249

Estadístico de Wald

$$Estadístico \quad de \quad Wald = \left(\frac{Coeficiente}{E.estándar} \right)^2$$

Es un estadístico similar a una t^2. Cuando el nivel crítico (*Sig.*) asociado al estadístico de Wald es menor que 0,05, se puede rechazar la hipótesis nula y, por tanto, concluir que la correspondiente variable independiente está significativamente relacionada con la variable dependiente.
Un inconveniente de este estadístico es que es demasiado sensible al tamaño de los coeficientes; en general, cuando el valor de un coeficiente es muy grande (en valor absoluto) el estadístico de Wald es poco fiable. En estos casos es preferible evaluar la significación de las variables utilizando un método por pasos

250

Razón de las ventajas

La columna de la *razón de las ventajas, Exp(B)*, permite cuantificar en qué grado aumenta la abstención cuando los sujetos no leen el periódico (y se mantienen constantes las restantes variables). Puesto que el punto de comparación es el valor 1 y el *Exp(B)* de la variable *leer* vale 1,956, se puede concluir que la *ventaja* de la abstención entre los sujetos que no leen el periódico es aproximadamente el doble que entre los que sí lo leen.

El signo negativo del resto de los coeficientes indica que el incremento en cualquiera de las demás variables disminuye la probabilidad de que un sujeto no vote: la abstención es menos probable a medida que aumentan la edad, los ingresos familiares y los años de escolarización.

251

Análisis de regresión logística por pasos

Cuando, se dispone de más de una variable independiente, existen varios métodos para seleccionar la variable o variables que deben formar parte del modelo final.

El método de *introducción forzosa* hace que el modelo de regresión incluya todas las variables independientes seleccionadas.

Los métodos de *selección por pasos* permiten utilizar criterios estadísticos para, de forma automática, incluir en el modelo las variables que son significativas y dejar fuera las que no lo son.

Los **métodos de *selección por bloques*** permiten al usuario manipular la inclusión y/o exclusión de variables mediante la combinación secuenciada de distintos procedimientos, pudiendo generar modelos jerárquicos.

252

Regresión logística multinomial

La opción Regresión logística multinomial resulta útil en aquellas situaciones en las que desee poder clasificar a los sujetos según los valores de un conjunto de variables predictoras. Este tipo de regresión es similar a la regresión logística, pero más general, ya que la variable dependiente no está restringida a dos categorías.

Ejemplo. Para conseguir una producción y distribución de películas más eficaz, los estudios de cine necesitan predecir qué tipo de películas es más probable que vayan a ver los aficionados. Mediante una regresión logística multinomial, el estudio puede determinar la influencia que la edad, el sexo y las relaciones de pareja de cada persona tienen sobre el tipo de película que prefieren. De esta manera, el estudio puede orientar la campaña publicitaria de una película concreta al grupo de la población que tenga más probabilidades de ir a verla.

253

Regresión logística versus análisis discriminante

La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante.

El análisis discriminante resulta útil para las situaciones en las que se desea construir un modelo predictivo para pronosticar el grupo de pertenencia de un caso a partir de las características observadas de cada caso. El procedimiento genera una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basada en combinaciones lineales de las variables predictoras que proporcionan la mejor discriminación posible entre los grupos. Las funciones se generan a partir de una muestra de casos para los que se conoce el grupo de pertenencia; posteriormente, las funciones pueden ser aplicadas a nuevos casos que dispongan de medidas para las variables predictoras pero de los que se desconozca el grupo de pertenencia.

254

Análisis discriminante

Datos. La variable de agrupación debe tener un número limitado de categorías distintas, codificadas como números enteros. Las variables independientes que sean nominales deben ser recodificadas a variables dummy o de contraste.

Supuestos. Los casos deben ser independientes. Las variables predictoras deben tener una distribución normal multivariada y las matrices de varianzas-covarianzas intra-grupos deben ser iguales en todos los grupos. Se asume que la pertenencia al grupo es mutuamente exclusiva (es decir, ningún caso pertenece a más de un grupo) y exhaustiva de modo colectivo (es decir, todos los casos son miembros de un grupo). El procedimiento es más efectivo cuando la pertenencia al grupo es una variable verdaderamente categórica; si la pertenencia al grupo se basa en los valores de una variable continua (por ejemplo, un cociente de inteligencia alto respecto a uno bajo), deberá considerar el uso de la regresión lineal para aprovechar la información más rica ofrecida por la propia variable continua.

255

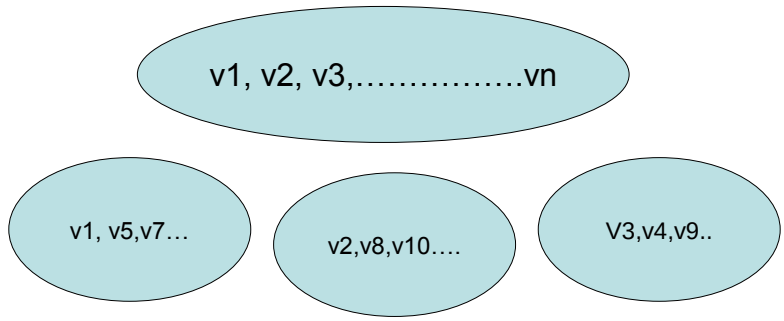
Análisis Factorial

- **Datos y supuestos**
- **Técnica del análisis factorial**
- **Fases del análisis factorial**
- **Matriz de correlaciones**
  - **Extracción de factores**
  - **Métodos de Rotación**
  - **Puntuaciones factoriales**

256

Análisis Factorial

El análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables.



Esos grupos homogéneos se forman con las variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros.

257

Datos y supuestos

Objetivo

Buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos.

Variables

En el análisis factorial no existe variable dependiente. Todas las variables del análisis tienen el mismo rango: todas ellas son independientes en el sentido de que no existe a priori una dependencia conceptual de unas variables sobre otras.
Las variables deberían ser cuantitativas a nivel de intervalo o de razón. Los datos categóricos (como la religión o el país de origen) no son adecuados para el análisis factorial.

258

Datos y supuestos

Supuestos

Los datos han de tener una distribución normal bivariada para cada pareja de variables, y las observaciones deben ser independientes.

Ejemplo

¿Qué actitudes subyacentes hacen que las personas respondan a las preguntas de una encuesta política de la manera en que lo hacen? Con el análisis factorial, se puede investigar el número de factores subyacentes y, en muchos casos, se puede identificar lo que los factores representan conceptualmente. Adicionalmente, se pueden calcular las puntuaciones factoriales para cada encuestado, que pueden utilizarse en análisis subsiguientes.

259

Técnica del análisis factorial.

Cada variable aparece como combinación lineal de una serie de factores

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + ... + A_{ik}F_k + U_i$$

donde *F* son los factores comunes a todas las variables y *U* es el factor único referido a la parte de la variable *i* que no puede ser explicada por los factores comunes. Las *A,* son los coeficientes de cada uno de los factores. Los factores únicos se asume que están incorrelacionados con el resto de factores únicos y con los factores comunes.
Cada factor es una combinación lineal de las variables originales

$$F_j = W_{j1}X_1 + W_{j2}X_2 + W_{j3}X_3 + ... + W_{ip}X_p$$

Wj son los coeficientes de las puntuaciones factoriales
P es el número de variables

260

Fases del Análisis factorial

El análisis factorial consta de cuatro fases características:

- El cálculo de una matriz capaz de expresar la variabilidad conjunta de todas las variable.
- La extracción del número óptimo de factores.
- La rotación de la solución para facilitar su interpretación.
- La estimación de las puntuaciones de los sujetos en las nuevas dimensiones.

Para ejecutar correctamente un análisis factorial es necesario tomar algunas decisiones en cada una de estas fases.

261

Ejemplo

Analicemos, por ejemplo, la pregunta del cuestionario cuyos datos recoge el archivo trabajo.sav y referida a la evaluación por parte de los encuestados de la importancia que según su opinión pueden tener cada una de las causas que se enumeran, en el alto índice de paro en un país.

B13: La crisis económica.
B14: La política de empleo del gobierno.
B15: La mala gestión de los empresarios.
B16: La comodidad de la gente, que sólo quiere buenos trabajos.
B17: La falta de preparación del trabajador.
B18: Las pocas ganas de trabajar de la gente.
B19: El no saber buscar trabajo.
B20: Que hay mucho pluriempleo.
B21: Que el trabajo que hay no se reparte bien socialmente.

262

El modelo matemático que subyace a esta técnica es similar al de la regresión simple y en él cada variable aparece como combinación lineal de una serie de factores que no son en este momento observables. Por ejemplo, B13 (la crisis económica) puede aparecer expresada como:

B13 = a(sujeto) + b(externos al sujeto) + c(entorno) + U_{B13}

donde sujeto, externos al sujeto y entorno no son variables independientes sino grupos de variables desconocidas por nosotros a priori, que pueden ser los factores subyacentes y que hemos denominado «sujeto» como factor que puede englobar las variables referidas a causas del paro inherentes al propio sujeto, «externas al sujeto», en donde estarían como causantes del paro el gobierno y los empresarios, por ejemplo, y el «entorno» en donde bien podrían estar la crisis económica y el reparto del trabajo.

263

Descriptivos

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
Crisis	3,93	,882	1009
Política de empleo	3,91	,933	1009
Empresarios	3,53	1,005	1009
Comodidad	3,02	1,133	1009
Preparación	2,92	1,086	1009
Ganas de trabajar	2,85	1,203	1009
Búsqueda	2,77	1,099	1009
Pluriempleo	3,57	1,005	1009
Reparto	3,87	,877	1009

Figura 1

264

Matriz de correlaciones

Matriz de correlaciones[a]

		Crisis	Política de empleo	Empresarios	Comodidad	Preparación	Ganas de trabajar	Búsqueda	Pluriempleo	Reparto
Correlación	Crisis	1,000	,397	,185	-,120	-,003	-,157	-,101	,019	,084
	Política de empleo	,397	1,000	,202	-,077	-,050	-,104	-,078	,054	,103
	Empresarios	,185	,202	1,000	,028	-,010	-,024	,044	,101	,161
	Comodidad	-,120	-,077	,028	1,000	,336	,559	,387	,214	,043
	Preparación	-,003	-,050	-,010	,336	1,000	,425	,345	,115	,045
	Ganas de trabajar	-,157	-,104	-,024	,559	,425	1,000	,451	,195	,071
	Búsqueda	-,101	-,078	,044	,387	,345	,451	1,000	,231	,134
	Pluriempleo	,019	,054	,101	,214	,115	,195	,231	1,000	,376
	Reparto	,084	,103	,161	,043	,045	,071	,134	,376	1,000
Sig. (Unilateral)	Crisis		,000	,000	,000	,457	,000	,001	,273	,004
	Política de empleo	,000		,000	,007	,057	,000	,006	,044	,001
	Empresarios	,000	,000		,185	,373	,221	,082	,001	,000
	Comodidad	,000	,007	,185		,000	,000	,000	,000	,087
	Preparación	,457	,057	,373	,000		,000	,000	,000	,077
	Ganas de trabajar	,000	,000	,221	,000	,000		,000	,000	,012
	Búsqueda	,001	,006	,082	,000	,000	,000		,000	,000
	Pluriempleo	,273	,044	,001	,000	,000	,000	,000		,000
	Reparto	,004	,001	,000	,087	,077	,012	,000	,000	

a. Determinante = ,240

Figura 2

Es importante que todas las variables tengan al menos un coeficiente de correlación significativo en la matriz.

265

El índice KMO

Inversa de la matriz de correlaciones

	Crisis	Política de empleo	Empresarios	Comodidad	Preparación	Ganas de trabajar	Búsqueda	Pluriempleo	Reparto
Crisis	1,239	-,441	-,134	,061	-,115	,138	,059	-,008	-,049
Política de empleo	-,441	1,223	-,153	,014	,034	,024	,048	-,043	-,059
Empresarios	-,134	-,153	1,086	-,063	,032	,054	-,058	-,038	-,128
Comodidad	,061	,014	-,063	1,548	-,148	-,674	-,207	-,160	,080
Preparación	-,115	,034	,032	-,148	1,290	-,376	-,231	,005	,005
Ganas de trabajar	,138	,024	,054	-,674	-,376	1,740	-,361	-,067	-,027
Búsqueda	,059	,048	-,058	-,207	-,231	-,361	1,380	-,144	-,086
Pluriempleo	-,008	-,043	-,038	-,160	,005	-,067	-,144	1,247	-,427
Reparto	-,049	-,059	-,128	,080	,005	-,027	-,086	-,427	1,201

Figura 3

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,712
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	1434,418
	gl	36
	Sig.	,000

Figura 4

En la tabla tenemos la inversa de la matriz de correlaciones, los «KMO» (Kaiser-Meyer-Olkin) y el test de Bartlett. Este último, es decir, el test de Bartlett, se utiliza para verificar si la matriz de correlaciones es una matriz de identidad, es decir, si todos los coeficientes de la diagonal son iguales a la unidad y los externos a la diagonal iguales a O.

266

El índice KMO

Este estadístico se obtiene a partir de la transforma-ción X^2 del determinante de la matriz de correlaciones y cuanto mayor sea y por
tanto menor el grado de significación, más improbable que la matriz sea una matriz de identidad. En el ejemplo, con un valor 1434,418 y un grado de significación p = 0,000 resulta evidente que no se trata de una matriz de identidad.

En el supuesto de que no se pudiese rechazar esta hipótesis, se desaconseja proceder a realizar un análisis factorial con los datos.

267

El índice KMO

El índice KMO nos compara los coeficientes de correlación de Pearson obtenidos en la Figura 2 con los coeficientes de correlación parcial entre variables. Se obtiene

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2}$$

rij el coeficiente de correlación de Pearson entre las variables i y j y *aij* es el coeficiente de correlación parcial entre las variables i y j.

268

El índice KMO

Si la suma de los coeficientes de correlación parcial al cuadrado es muy pequeña, KMO será un índice muy próximo a la unidad y por tanto el análisis factorial un procedimiento adecuado. En cambio, valores pequeños en este índice nos dan a entender todo lo contrario. De hecho para Kaiser :

1≥ KMO > 0,90 son considerados excelentes.
0,90 ≥ KMO > 0,80 son considerados buenos.
0,80 ≥ KMO > 0,70 son considerados aceptables.
0,70 ≥ KMO > 0,60 son considerados mediocres o regulares.
0,60 ≥ KMO > 0,50 son considerados malos.
KMO < 0,50 son considerados inaceptables o muy malos.

En el ejemplo este valor es de 0,712 y por tanto se puede considerar como aceptable y continuar con el análisis factorial.

Matrices anti-imagen

Matrices anti-imagen

		Crisis	Política de empleo	Empresarios	Comodidad	Preparación	Ganas de trabajar	Búsqueda	Pluriempleo	Reparto
Covarianza anti-imagen	Crisis	,807	-,291	-,100	,032	-,072	,064	,034	-,005	-,033
	Política de empleo	-,291	,818	-,115	,007	,022	,011	,028	-,028	-,040
	Empresarios	-,100	-,115	,921	-,038	,023	,029	-,039	-,028	-,098
	Comodidad	,032	,007	-,038	,646	-,074	-,250	-,097	-,083	,043
	Preparación	-,072	,022	,023	-,074	,775	-,168	-,130	,003	,003
	Ganas de trabajar	,064	,011	,029	-,250	-,168	,575	-,151	-,031	-,013
	Búsqueda	,034	,028	-,039	-,097	-,130	-,151	,725	-,083	-,052
	Pluriempleo	-,005	-,028	-,028	-,083	,003	-,031	-,083	,802	-,285
	Reparto	-,033	-,040	-,098	,043	,003	-,013	-,052	-,285	,833
Correlación anti-imagen	Crisis	,601[a]	-,358	-,116	,044	-,091	,094	,045	-,006	-,040
	Política de empleo	-,358	,609[a]	-,132	,010	,027	,016	,037	-,034	-,048
	Empresarios	-,116	-,132	,690[a]	-,049	,027	,039	-,047	-,033	-,112
	Comodidad	,044	,010	-,049	,744[a]	-,104	-,411	-,142	-,115	,058
	Preparación	-,091	,027	,027	-,104	,791[a]	-,251	-,173	,004	,004
	Ganas de trabajar	,094	,016	,039	-,411	-,251	,721[a]	-,233	-,046	-,019
	Búsqueda	,045	,037	-,047	-,142	-,173	-,233	,816[a]	-,109	-,067
	Pluriempleo	-,006	-,034	-,033	-,115	,004	-,046	-,109	,669[a]	-,349
	Reparto	-,040	-,048	-,112	,058	,004	-,019	-,067	-,349	,591[a]

a. Medida de adecuación muestral

Figura 5

En la Figura 5 tenemos las matrices anti-imagen de covariancias y correlaciones entre todas las variables del ejemplo. Serán los negativos de los coeficientes de correlación parcial entre cada par de variables, neutralizando el efecto de todas las restantes. Interesan por tanto coeficientes cuanto más pequeños, mejor.

En la diagonal de esta última tenemos los coeficientes MSA (Measures of Sam-pling Adequacy) que vienen a ser los KMO pero en este caso para cada variable por separado. La interpretación de sus valores es idéntica a la realizada para los KMO.
En resumen, tenemos:

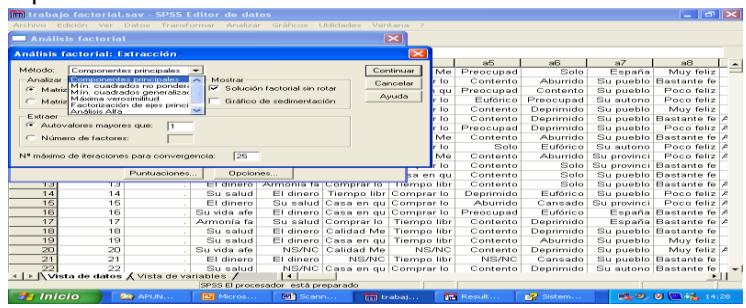
- Coeficientes de correlación de Pearson que en la mayoría de los casos son altamente significativos.
- El determinante de la matriz de correlaciones (0,240) relativamente bajo.
- El índice KMO = 0,712 bastante aceptable.
- El resultado del test de Bartlett con un $\chi 2$ = 1434,418 Y p = 0,000.
- Valores muy bajos en la matrices anti-imagen,
- MSA bastante altos en la diagonal de la matriz de correlaciones anti-imagen.

Todo ello nos lleva a concluir que el análisis factorial que sigue a continuación resulta a priori pertinente y puede proporcionamos conclusiones satisfactorias.

271

Extracción de factores

Método. Permite especificar el método de extracción factorial. Los métodos disponibles son: Componentes principales, Mínimos cuadrados no ponderados, Mínimos cuadrados generalizados, Máxima verosimilitud, factorización de Ejes principales, factorización Alfa y factorización Imagen.
En SPSS el sistema coge por defecto el método de componentes principales que es el que hemos utilizado en esta primera parte del ejemplo.



272

Componentes principales (PC)

Consiste básicamente en llevar a cabo una combinación lineal de todas las variables de modo que el primer componente principal sea una combinación que explique la mayor proporción de variancia de la muestra, el segundo la segunda mayor y que a su vez esté incorrelacionado con el primero, y así sucesivamente hasta tantos componentes como variables.

El método de extracción de componentes principales, es el que actúa por defecto, asume que es posible explicar el 100% de la varianza observada y, por ello, todas las comunalidades iniciales son iguales a la unidad (que es justamente la varianza de una variable en puntuaciones típicas).

273

Comunalidades

Comunalidades

	Inicial	Extracción
Crisis	1,000	,644
Política de empleo	1,000	,620
Empresarios	1,000	,329
Comodidad	1,000	,592
Preparación	1,000	,523
Ganas de trabajar	1,000	,684
Búsqueda	1,000	,519
Pluriempleo	1,000	,646
Reparto	1,000	,693

Método de extracción: Análisis de Componentes principales.

La comunalidad de una variable es la proporción de su varianza que puede ser explicada por el modelo factorial obtenido.

Figura 6

En la Figura 6 tenemos las **comunalidades iniciales** de la solución de componentes principales. Estos resultados se obtienen si en el subcuadro de diálogo «Descriptives» de la Figura 2 y dentro de «Statistics» seleccionamos «Initial Solution». Si utilizamos tantos componentes principales como variables, cada variable puede ser explicada por ella misma y por tanto toda la variabilidad de cada variable, que expresada en unidades de desviación estandarizadas es igual a la unidad, explicada a su vez por los factores comunes. Esta es la razón por la que en la Figura 6 la comunalidad inicial es igual a la unidad para todas las variables.

274

Valores propios

La decisión respecto al número de factores que deseamos para representar los datos puede adoptarse desde una doble vía que es la que aparece en el subcuadro de diálogo «Extraction. Por defecto el sistema extraerá tantos factores como haya en la solución inicial con valores propios (eigenvalues) superiores a la unidad. En la Figura 7 vemos que hay tres factores con valores propios superiores a 1 y que en definitiva será el número que extraerá el sistema. Evidentemente, podemos cambiar el valor por defecto correspondiente al «eigenvalue». La segunda posibilidad corresponde al botón de radio «Number of factors» y consiste sencillamente en fijar un número entero determinado de factores, siempre inferior, lógicamente, al número de variables. **Los autovalores (o valores propios) expresan la cantidad de la varianza total que está explicada por cada factor; y los porcentajes de varianza explicada asociados a cada factor se obtienen dividiendo su correspondiente autovalor por la suma de los autovalores (la cual coincide con el número de variables**

275

Matriz Varianza total explicada

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,449	27,211	27,211	2,449	27,211	27,211	2,274	25,265	25,265
2	1,684	18,714	45,925	1,684	18,714	45,925	1,553	17,252	42,518
3	1,116	12,395	58,320	1,116	12,395	58,320	1,422	15,802	58,320
4	,848	9,426	67,747						
5	,705	7,834	75,580						
6	,616	6,842	82,422						
7	,597	6,629	89,051						
8	,568	6,314	95,365						
9	,417	4,635	100,000						

Método de extracción: Análisis de Componentes principales.

Figura 7

La Figura 7 recoge, en porcentajes individuales y acumulados, la proporción de variancia total explicada por cada factor, tanto para la solución no rotada como para la rotada. En concreto, qué porcentaje supone 2,449 sobre el total de variabilidad (nueve en el ejemplo) de toda la muestra. Los tres factores incluidos en el modelo son capaces de explicar exactamente un 58,32 por 100 de la variabilidad total, lo que puede interpretarse como un porcentaje aceptable.

276

Gráfico de sedimentación

Gráfico de sedimentación

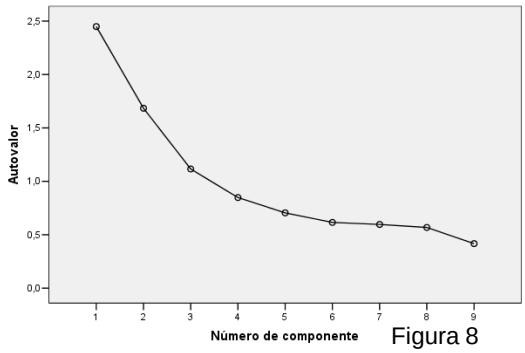


Figura 8

El *gráfico de sedimentación* sirve para determinar el número óptimo de factores. Consiste simplemente en una representación gráfica del tamaño de los autovalores. Según se ha señalado ya, los autovalores indican la cantidad de varianza que está explicada por cada componente principal

Tanto la tabla de porcentajes de varianza explicada como el gráfico de sedimentación muestran los autovalores ordenados de mayor a menor: el primer autovalor es el mayor de los posibles, el segundo autovalor es el mayor de los restantes, y así sucesivamente. Si un autovalor se aproxima a cero, esto significa que el factor correspondiente a ese autovalor es incapaz de explicar una cantidad relevante de la varianza total. Por tanto, un factor al que corresponde un autovalor próximo a cero se consi-dera un factor residual y carente de sentido en el análisis.

277

Matriz de componentes

Matriz de componentes[a]

	Componente		
	1	2	3
Crisis	-,229	,640	,426
Política de empleo	-,185	,668	,373
Empresarios	,020	,569	,070
Comodidad	,748	-,036	,177
Preparación	,629	,002	,357
Ganas de trabajar	,804	-,085	,174
Búsqueda	,718	,045	,040
Pluriempleo	,436	,437	-,515
Reparto	,236	,551	-,577

Método de extracción: Análisis de componentes principales.
a. 3 componentes extraídos

Figura 9

En la Figura 9 tenemos los coeficientes utilizados para expresar cada variable estandarizada en términos de los tres factores del modelo.

Estos coeficientes se conocen también con el nombre de pesos factoriales, cargas, ponderaciones factoriales o saturaciones factoriales ya que nos indican la carga de cada variable en cada factor, de modo que los factores con unos pesos factoriales más elevados en términos absolutos nos indican una relación estrecha con las variables.

278

Matriz de componentes

El ideal desde el punto de vista del análisis factorial es encontrar un modelo en el que todas las variables saturen en algún factor, es decir, pesos factoriales altos en uno y bajos en el resto. Por ejemplo la variable B18 (Ganas de trabajar como posible explicación del alto índice de desempleo en el país) es una variable con una elevada carga factorial en el primero de los fac-tores y mucho más pequeña en los dos restantes. Podríamos expresar la variable B18 como:

B18 = 0,80408 *F1* - 0,08519 *F2* + 0,17407 F3

donde *F1, F2 y* F3 son los tres factores del modelo.

Matriz de componentes

Matriz de componentes[a]

	Componente		
	1	2	3
Ganas de trabajar	,804	-,085	,174
Comodidad	,748	-,036	,177
Búsqueda	,718	,045	,040
Preparación	,629	,002	,357
Política de empleo	-,185	,668	,373
Crisis	-,229	,640	,426
Empresarios	,020	,569	,070
Reparto	,236	,551	-,577
Pluriempleo	,436	,437	-,515

Método de extracción: Análisis de componentes principales.
a. 3 componentes extraídos

Matriz de componentes[a]

	Componente		
	1	2	3
Ganas de trabajar	,804		,174
Comodidad	,748		,177
Búsqueda	,718		
Preparación	,629		,357
Política de empleo	-,185	,668	,373
Crisis	-,229	,640	,426
Empresarios		,569	
Reparto	,236	,551	-,577
Pluriempleo	,436	,437	-,515

Método de extracción: Análisis de componentes principales.
a. 3 componentes extraídos

En la Figura 9 aparecen ordenadas las variables tal y como están en la base de datos. La segunda tabla de esta figura es la que corres-ponde a la opción que hemos seleccionado en el subcuadro de diálogo Options al seleccionar «Sorted by size» el sistema ordena las variables en la matriz de mayor a menor peso o carga factorial y siempre comenzando por el primer factor, posteriormente el segundo, y así sucesivamente. Finalmente la opción del mismo subcuadro de diálogo «Suppress absolute values less than» nos permite una lectura todavía más clara de la matriz puesto que permite eliminar de la misma aquellos coeficientes con valores inferiores a uno dado (0,10 por defecto).

Matriz de componentes

Para determinar en qué medida los tres factores son capaces de explicar las variables originales, podemos sumar la proporción de variancia de la variable explicada por cada uno de ellos (es decir, los coeficientes al cuadrado) y de este modo obtener las comunalidades que aparecen en la diagonal de la Figura 10. Cojamos de nuevo la variable B18 (Ganas de trabajar) y calculemos este sumatoria:

$B\ 18 = 0{,}80408^2 + 0{,}08519^2 + 0{,}17407^2 = 0{,}68410$

Casi el 70 por 100 de la variabilidad de *B18* es explicada por los tres factores
del modelo, en tanto que por ejemplo en la variable *BI5* (Empresarios) los mismos únicamente explican en torno al 33 por 100. Reiteramos que esta proporción de la variabilidad de cada variable explicada por los factores del modelo es lo que se conoce con el nombre de comunalidad de la variable. Obviamente su valor oscila entre 0 y 1 y la parte de variancia no explicada por el modelo factorial, es decir, 1-comunalidad, es lo que se conoce con el nombre de factor único o unicidad.

281

Rotación

La finalidad de la rotación es la de ayudamos a interpretar.
En el subcuadro de diálogo “Rotación” existen varios procedimientos.
VARIMAX, EQUAMAX y QUARTlMAX son procedimientos ortogonales es decir que los factores se mantienen incorrelacionados y los ejes forman ángulos rectos.
El PROMAX y el DIRECT OBLIMIN pertenecen al grupo de los denominados oblicuos o no ortogonales.
La rotación no afecta a la comunalidad y al porcentaje de variancia explicada por el modelo, aunque sí puede cambiar la de cada factor.

282

Métodos de Rotación

Varimax. Método de rotación ortogonal que minimiza el número de variables que tie-nen saturaciones altas en cada factor. Simplifica la interpretación de los factores opti-mizando la solución por columna.
Quartimax. Método de rotación ortogonal que minimiza el número de factores nece-sarios para explicar cada variable. Simplifica la interpretación de las variables obser-vadas optimizando la interpretación por filas.
Equamax. Método de rotación que es combinación del método Varimax, que simpli-fica los factores, y del método Quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.

283

Métodos de Rotación

Oblimin directo. Método para la rotación oblicua (no ortogonal). Cuando *delta* es igual a cero (el valor por defecto), las soluciones son las más oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el va-lor por defecto de delta, puede introducirse un número menor o igual que 0,8.
Delta. El valor de delta permite controlar el grado de oblicuidad que pueden lle-gar a alcanzar los factores de la solución.
Promax. Rotación oblicua que permite que los factores estén correlacionados. Puede calcularse más rápidamente que una rotación *oblimin directa,* por lo que es útil para grandes conjuntos de datos.
Kappa. Parámetro que controla el cálculo de la rotación Promax. El valor por de-fecto es 4. Este valor es adecuado para la mayoría de los análisis.

284

Matriz factorial

En resumen, todos los métodos tratan de obtener una matriz factorial que se aproxime al principio de estructura simple. Según este principio, la matriz factorial debe reunir las siguientes características:

- Cada factor debe tener unos pocos pesos altos y el resto próximos a 0.
- Cada variable no debe estar saturada mas que en un solo factor.
- No deben existir factores con la misma distribución.

El método utilizado en todos los casos ha sido el de componentes principales. Todos ellos coinciden a grandes rasgos en la siguiente asignación:

285

Matriz factorial

Factor 1 Variables:

B18: Pocas ganas de trabajar de la gente.
B16: La comodidad de la gente, que sólo quiere buenos trabajos.
*B*19: El no saber buscar trabajo.
B17: La falta de preparación del trabajador.

Factor 2 Variables:

B 14: La política de empleo del gobierno.
B 13: La crisis económica.
B15: La mala gestión de los empresarios.

Factor 3 Variables:
B21: Que el trabajo que hay no se reparte bien socialmente.
B22: Que hay mucho pluriempleo.

286

Matriz de pesos factoriales

Matriz de componentes rotados[a]

	Componente		
	1	2	3
Ganas de trabajar	,818	-,111	,055
Comodidad	,765	-,058	,062
Preparación	,712	,085	-,095
Búsqueda	,688	-,056	,205
Crisis	-,090	,795	-,059
Política de empleo	-,067	,784	,010
Empresarios	,027	,512	,257
Reparto	,014	,129	,822
Pluriempleo	,226	,033	,771

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Quartimax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Matriz de componentes rotados[a]

	Componente		
	1	2	3
Ganas de trabajar	,813	-,122	,088
Comodidad	,761	-,069	,094
Preparación	,716	,077	-,063
Búsqueda	,678	-,068	,233
Crisis	-,077	,797	-,051
Política de empleo	-,057	,785	,019
Empresarios	,023	,508	,265
Reparto	-,019	,117	,824
Pluriempleo	,194	,019	,780

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Equamax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Matriz de componentes rotados[a]

	Componente		
	1	2	3
Ganas de trabajar	,815	-,118	,076
Comodidad	,762	-,065	,083
Preparación	,715	,080	-,074
Búsqueda	,682	-,064	,224
Crisis	-,081	,796	-,053
Política de empleo	-,061	,785	,017
Empresarios	,024	,509	,263
Reparto	-,008	,120	,823
Pluriempleo	,205	,023	,777

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Matriz de configuración[a]

	Componente		
	1	2	3
Ganas de trabajar	,813	-,085	-,014
Comodidad	,762	-,035	-,021
Preparación	,740	,122	,146
Búsqueda	,664	-,048	-,172
Crisis	-,026	,806	,106
Política de empleo	-,013	,790	,036
Empresarios	,028	,496	-,229
Reparto	-,091	,056	-,834
Pluriempleo	,124	-,028	-,775

Método de extracción: Análisis de componentes principales.

Metodo de rotación: Normalización Oblimin con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

287

Matriz de pesos factoriales

Matriz de configuración[a]

	Componente		
	1	2	3
Ganas de trabajar	,815	-,068	-,002
Comodidad	,765	-,019	,009
Preparación	,751	,136	-,154
Búsqueda	,663	-,034	,161
Crisis	-,006	,804	-,080
Política de empleo	,005	,789	-,010
Empresarios	,034	,497	,246
Reparto	-,111	,057	,841
Pluriempleo	,105	-,022	,776

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Promax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Analizando someramente estos resultados, bien podría tratarse de tres factores claramente diferenciados y referidos:

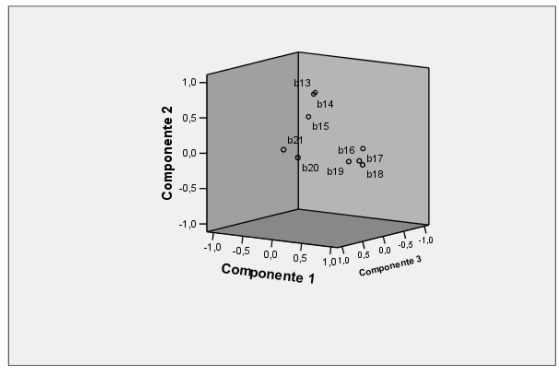
1) Al *trabajador*
2) *Gobierno y empresarios*
3) Reparto o redistribución del *trabajo*

288

Gráfico de componentes en espacio rotado

Gráficamente podemos ver estos mismos resultados en la Figura que corresponde al gráfico tridimensional de la solución rotada VARIMAX y componentes principales.

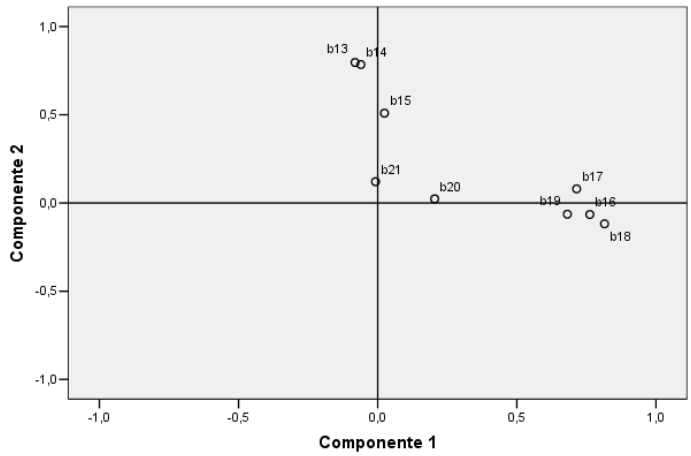
Gráfico de componentes en espacio rotado



289

Gráfico de componentes en espacio rotado

Gráfico de componentes en espacio rotado



290

Gráfico de componentes en espacio rotado

Los valores de cada variable en las coordenadas corresponden a los pesos factoriales de las mismas en los ejes de cada factor. Pueden ser valores comprendidos entre -1 y 1,(cuanto mayor sea esta coordenada, más contribuye a la formación del eje, a la inercia del mismo).
En realidad los planos factoriales están situados en el interior de un círculo de radio la unidad, y en ese sentido lo ideal es que los puntos «variables del estudio» no estén concentrados en torno al origen del espacio bidimensional en este caso (0,0) sino próximos al borde del círculo o de los ejes factoriales.

291

Gráfico de componentes en espacio rotado

En concreto y referido a los resultados del ejemplo en el gráfico de las dos primeras dimensiones:

Las variables:
B17: La falta de preparación del trabajador.
B16: La comodidad de la gente, que sólo quiere buenos trabajos.
B18: Las pocas ganas de trabajar de la gente.
B 19: El no saber buscar trabajo.
BI3: La crisis económica.
B14: La política de empleo del gobierno
Son las variables que están mejor representadas sobre el plano.

En peor posición están las variables:
B13: La mala gestión de los empresarios.
B20: Que hay mucho pluriempleo.
B21: Que el trabajo que hay no se reparte bien socialmente.

B 16, *B* 17, *B* 18 y *B*19 están altamente correlacionadas entre si y a su vez correlacionadas positivamente con el factor1 (están situadas
B 13 y *B*14 lo mismo pero para el factor 2. Es negativa en cambio la relación con el primer factor.

292

Matriz de componentes rotados

El primer factor contrapone variables inherentes al propio trabajador con variables referidas a la redistribución del trabajo. En el factor 2 son políticas de empleo y crisis económica versus reparto.

Matriz de componentes rotados[a]

	Componente		
	1	2	3
Ganas de trabajar	,815	-,118	
Comodidad	,762		
Preparación	,715		
Búsqueda	,682		,224
Crisis		,796	
Política de empleo		,785	
Empresarios		,509	,263
Reparto		,120	,823
Pluriempleo	,205		,777

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Siguiendo con la idea de identificar del mejor modo posible las variables que en cualquier caso tienen pesos factoriales más elevados o saturan más en cada uno de los factores , el sistema nos posibilita eliminar de la matriz de pesos factoriales y en las columnas de los diversos factores, los pesos de aquellas variables con un valor inferior a uno determinado y que por defecto es 0,10.

293

Puntuaciones factoriales

Puesto que la finalidad última del análisis factorial es reducir un gran número de variables a un pequeño número de factores, es a veces aconsejable estimar las pun-tuaciones factoriales de cada sujeto.

Matriz de coeficientes para el cálculo de las puntuaciones en las componentes

	Componente		
	1	2	3
Crisis	-,093	,380	,382
Política de empleo	-,076	,397	,335
Empresarios	,008	,338	,063
Comodidad	,305	-,021	,159
Preparación	,257	,001	,320
Ganas de trabajar	,328	-,051	,156
Búsqueda	,293	,027	,036
Pluriempleo	,178	,259	-,462
Reparto	,096	,327	-,518

Método de extracción: Análisis de componentes principales.

Puntuaciones de componentes.

294

Nota: Aunque en la práctica el análisis factorial (AF) y el método de componentes principales (PC) se utilizan indistintamente y dan resultados similares, conviene señalar que así como en el análisis de componentes principales el objetivo consiste en encontrar una serie de componentes que expliquen el máximo de variancia *total* de las variables originales, el objetivo del análisis factorial es encontrar una serie de *factores* que expliquen el máximo de variancia *común* de las variables originales.

295