Notas de Clase: Econometría II

Regresión Multivariada y otras herramientas del análisis de datos

Benjamín Oliva [1]
Jésica Tapia [2]
Omar Alfaro [3]

Draft Mayo 2023

[1] benjov@ciencias.unam.mx y https://github.com/benjov
[2] jesicatapia@gmail.com y https://github.com/
[3] omarxalpha@gmail.com

Documento siempre proceso de mejora.
Comentarios, siempre serán bienvenidos...

Índice general

1. El concepto de esperanza condicional y su relación con la econometría **3**
1.1. Introducción 3
1.2. Esperanza condicional y algunas de sus propiedades 5
1.3. Efectos Marginales 9
1.3.1. Ejericios 10
1.4. Introducción a Gráficos Acíclicos Dirigidos o Directed Acycle Graphical 10

2. Teoría de convergencia asintótica **15**
2.1. Convergencia de sucesiones 15
2.2. Convergencia en probabilidad y acotamientos en probabilidad 16
2.3. Convergencia en distribución 17
2.4. Teoremas límite para muestras aleatorias 18
2.5. Estimadores y estadísticas de prueba 18

3. Estimación de modelos lineales de una sola ecuación por el método de Mínimos Cuadrados Ordinarios **21**
3.1. El estimador de Mínimos Cuadrados Ordinarios (MCO) y el análisis clásico de regresión 21
3.2. Bondad de ajuste 25
3.3. Modelos lineales y algunas de sus propiedades 28
3.4. Propiedades asintóticas de los estimadores de MCO 30
3.5. Inferencia asintótica bajo MCO 32
3.6. Ejercicios 40

4. Estimación de modelos lineales de una sola ecuación por el método de Variables Instrumentales **45**

4.1. Introducción y motivación 45
4.2. Estimación por Mínimos Cuadrados en Dos Etapas – 2SLS . . 46
4.3. Método de Variables Instrumentales en casos con múltiples instrumentos 47
4.4. Ejercicios 50

5. Estimación de sistemas de ecuaciones por el por el método de Mínimos Cuadrados Ordinarios y Mínimos Cuadrados Generalizados 55
5.1. Introducción a Sistemas de Ecuaciones 55
5.2. Estimación de Sistemas de Ecuaciones Multivariantes 57
5.3. Mínimos Cuadrados Generalizados 59
5.4. Estimación de Sistemas de Ecuaciones con Variables Instrumentales 63
5.4.1. Estimación de un sistema por Mínimos Cuadrados en Dos Etapas (2SLS) 65
5.4.2. Estimación de un sistema por Mínimos Cuadrados en Tres Etapas (3SLS) 65
5.5. Ejercicios 66

6. Modelos de Datos Panel 67
6.1. Introducción y motivación 67
6.2. Modelos y métodos de estimación 68
6.2.1. Efectos Fijos 69
6.2.2. Efectos Aleatorios 71
6.2.3. Regresión Pool 72
6.2.4. Pruebas para seleccionar modelo 73
6.3. Conclusión 74

7. Métodos de estimación basados en verosimilitud 75
7.1. Introducción 75
7.2. Marco general de estimación por máxima verosimilitud 77
7.3. Pruebas de hipótesis 78

8. Estimación de modelos no lineales 79
8.1. Modelos de respuesta binaria 79
8.1.1. Planteamiento general 79
8.1.2. Algunos planteamientos adicionales 80

8.1.3. Modelos Logit y Probit 80
8.1.4. Estimación 82
8.2. Modelos de respuesta multinomial y ordenada 83
8.2.1. Modelos de respuesta multinomial 83
8.2.2. Modelos de respuesta ordenada: Logit y Probit Ordinal 85
8.3. Modelos de conteo y otras respuestas 89
8.3.1. Modelos de conteo del tipo Poisson 89

9. Modelos con datos truncados, censurados, con endogeneidad por selección de muestra y otros casos 93
9.1. Truncamiento 93
9.2. Datos con censura 100

10. Introducción a inferencia causal 103
10.1. Estimación de modelos para determinar efectos de tratamiento 103
10.1.1. Motivación del procedimiento Difference - in - Differences (DiD) 103
10.1.2. Estrategia de estimación 104
10.2. Control Sintético 106
10.2.1. Motivación 106
10.2.2. Planteamiento del método de estimación 107
10.2.3. ¿Por qué usar el control sintérico? 109

11. Estimación de modelos de duración 111

12. Introducción al Aprendizaje Estadístico 113
12.1. Motivación e introducción 113
12.2. Modelos linealles y el procedimiento de mínimos cuadrados . . 115
12.3. Método de regresiones restringidas 116
12.3.1. Regresión Ridge 116
12.3.2. Regresión Lasso (Least Absolute Shrinkage and Selection Operator) 117
12.4. Modelos lineales de clasificación 118
12.4.1. Evaluando los modelos de aprendizaje supervisado . . . 121
12.5. Aprendizaje no supervisado 122

Bibliografía 123

A. Algunos resultados importantes 125

Índice de figuras

1.1. Ejemplo 1 de un DAG 11
1.2. Ejemplo 2 de un DAG, confusor 11
1.3. Ejemplo 3 de un DAG, ganancias de la educación 12
1.4. Ejemplo 4 de un DAG, IV estimator 13

4.1. Instrumentación de la información de la demanda y la oferta, respectivamente 52

9.1. ¿Talento y la belleza están negativamente correlacionados? . . 94
9.2. Relación entre belleza y talento, retomado de [Cun21]. 95

12.1. División del conjunto de datos, retomado de Hastie, Tibshirani, y Friedman (2017, p. 222) [HTF17] 121

Índice de cuadros

3.1. Resultados de la regresión de salarios 42

10.1. Tabla XII de Snow (1854)–Casos de cólera en Londres por cada 10,000 habitantes según proveedor de agua, retomado de Cunningham(2021, p. 472) [Cun21] 105

12.1. Matriz de Confusión 121

Introducción, motivación y alcance del documento

Queremos discutir los documentos de la bibliografia como: Adams (2020) [Ada20], Cameron y Trivedi (2005) [CT05]; Cunningham (2021) [Cun21]; Greene (2012) [Gre12], Hastie, Tibshirani y Friedman (2017) [HTF17], James, Witten, Hastie y Tibshirani (2013) [Jam+13], y Wooldridge (2010) [Woo10]. Asimismo, discutir diversos papers relacionados.

Conocimientos previos

Preferentemente Econometría I y obligatoriamiente Estadísitica II, Introducción a la Econometría o sus equivalentes.

Recursos en línea y otros materiales

Direcciones del GitHub, recomendaciones de cursos en línea, materiales recomendados y demas recursos para el autoaprendizaje.

1

El concepto de esperanza condicional y su relación con la econometría

1.1. Introducción

El objetivo del análisis empírico en las ciencias sociales es determinar cuál es el efecto o cambio que una variable causa en otra. Por ejemplo, cuando pensamos en qué factores o variables determinan el nivel salarial promedio de las personas es posible que relacionemos el efecto que tienen los años de estudio y en particular el efecto que tiene un año adicional de estudio. Otro ejemplo puede presentarse cuando tratamos de entender el efecto que tienen las horas de estudio en las calificaciones finales de un grupo de estudiantes.

Dentro del análisis empírico se suele hacer uso de dos términos o conceptos:

1. *Ceteris paribus*
2. *Causalidad*

El primero se emplea en el análisis empírico para establecer que el efecto de una variable se sostiene siempre que asumamos que el resto de variables que pueden afectar a dicha variable que queremos explicar permanecen constantes. El segundo sirve para entender que en el análisis empírico la existencia de correlaciones no significa la presencia de causalidad. La causalidad, en la primera parte del curso, será la condición observada entre dos variables y

deriva de la construcción teórica, modelación o racionalización que hagamos de los fenómenos que queremos analizar. Más adelante modificaremos esta condición.

Tanto *ceteris paribus* como causalidad son conceptos contenidos en lo que en estadística se conoce como **esperanza condicional**. Supongamos dos variables, y y x, que tienen una distribución conjunta y para las cuales queremos estimar el efecto que tiene la segunda en la primera. Así, es posible que utilicemos una expresión de la esperanza condicional de y en x, misma que solemos representar como:

$$\mathbb{E}[y|x, \mathbf{C}] \tag{1.1}$$

Donde $\mathbf{C}$ representa un conjunto o vector de variables de control, en el sentido de que también explican la variabilidad de y por lo que no podemos omitirlas en un proceso de estimación de la esperanza condicional.

Dicho lo anterior, el análisis empírico que proponemos en este curso está basado en identificar correlaciones que pueden ser interpretadas como causalidad y que consiste en estimar a la ecuación (1.1) y determinar el efecto marginal o efecto parcial que tiene x en y, condicional en que todas las demás variables permanecen constantes, es decir, *ceteris paribus*.

En la ecuación (1.1) hemos asumido que $\mathbf{C}$ contiene la información disponible que sirve de control, por lo que es un vector que contiene sólo información que es observable. Sin embargo, no siempre es posible observar toda la información que sirve de control. Por ejemplo, al responder qué faactores determinan el salario de las personas podríamos pensar en múltiples factores cuantificables y observables como: edad, sexo, años de educación, etc. No obstante, también consideraríamos factores como las habilidades propias de cada individuo, las cuales no son necesariamente observables.

Adicionalmente, en el proceso de estimación de la ecuación (1.1) asumiremos que las muestras de datos que utilizaremos son muestras aleatorias de la población. En concreto asumiremos que la información es una muestra aleatoria independiente e idénticamente distribuida (iid) que proviene de una población dada.

Para finalizar esta sección introductoria plantearemos los tres casos que ejemplifican los tipos de datos que son comúnmente analizados en el análisis de regresión o econometría. El primero es un análisis de datos de sección cruzada en el que las ecuaciones representativas de la ecuación (1.1) son

como la siguiente:

$$ln(Salario_i) = \beta_0 + \beta_1 Educacion_i + \beta_2 Experiencia_i$$

Donde las variables de $Salario_i$, $Educacion_i$ y $Experiencia_i$ son observadas para cada uno de los individuos en la muestra es indexado por $i = 1, 2, 3, \ldots, N$.

El segundo ejemplo es respecto de datos de series de tiempo como la siguiente expresión:

$$ln(PIB_t) = \beta_0 + \beta_1 Empleo_t + \beta_2 Salarios_t$$

Donde las variables PIB_t, $Empleo_t$ y $Salarios_t$ son observadas para un individuo o entidad a lo largo de una muestra del tiempo indexada por $t = 1, 2, 3, \ldots, T$.

El tercer ejemplo resulta de la combinación de los dos anteriores. De esta forma podríamos observar a una muestra de un conjunto de individuos a lo largo del tiempo y analizar el comportamiento de una variable através de una ecuación como la siguiente:

$$ln(Salario_{it}) = \beta_0 + \beta_1 Educacion_{it} + \beta_2 Experiencia_{it}$$

Donde las variables $Salario_{it}$, $Educacion_{it}$ y $Experiencia_{it}$ se observan para los mismos individuos a lo largo del tiempo que se indexan con los pares (i, t), $i = 1, 23, , \ldots, N$ y $t = 1, 2, 3, \ldots, T$.

1.2. Esperanza condicional y algunas de sus propiedades

En primer lugar estableceremos un poco de notación. En análisis de regresión siempre partimos de una representación de una ecuación lineal como:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK} + \varepsilon_i$$

Donde cada una de las variables y_i, x_{ik} y ε_i, se observan para $i = 1, 2, 3, \ldots, N$ y $k = 1, 2, \ldots, K$, por lo que podremos utilizar una representa-

ción para cada i de la forma:

$$\begin{aligned} y_i &= \begin{pmatrix} x_{i1} & x_{i2} & \dots & x_{iK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \varepsilon_i \\ &= \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i \end{aligned} \tag{1.2}$$

En forma general la ecuación (1.2) se puede generalizar para el total de elementos en la muestra como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.3}$$

Donde:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \dots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

En la mayoría de los casos analizados vamos a asumir que la ecuación (1.2) tiene un término constante, por lo que $x_{i1} = 1$ para todo $i = 1, 2, \ldots, N$. Bajo este escenario, representaremos la matriz $\mathbf{X}$ con una columna compuesta del número 1 (uno) en todas sus entradas, tal que, el paramétro β_1 es un término constante en las ecuaciones (1.2) y (1.3). De esta forma la matriz anteriormente mostrada se puede ver como:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \ldots & x_{1K} \\ 1 & x_{22} & \ldots & x_{2K} \\ 1 & x_{32} & \ldots & x_{3K} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n2} & \ldots & x_{nK} \end{bmatrix}$$

De forma similar al caso de una regresión lineal que incluye un término constante, podemos construir matrices $\mathbf{X}$ para casos en que alguna de las variables explicativas es dicotómica –que toma valores de 0 y 1– y casos en los que las variables explicativas han sido interactuadas –multiplicada entre sí– o tranformadas mediante potencias o logaritmos.

Retomando las ecuaciones (1.2) y (1.3) y para facilitar la exposición no utilizaremos el índice i, plantearemos la siguiente definición de esperanza condicional.

Definición 1.1 *Sea y una variable aleatoria, la cual referiremos como la variable explicada o variable dependiente. Asimismo, sea $(x_1, x_2, \ldots, x_K)$ un vector aleatorio de dimensión $K \times 1$ de variables explicativas. Finalmente, asumamos que $\mathbb{E}[|y|] < \infty$. Entonces, existe una función $\mu : \mathbb{R}^{K+1} \to \mathbb{R}$ tal que:*

$$\mathbb{E}[y|x_1, x_2, \ldots, x_K] = \mu(x_1, x_2, \ldots, x_K)$$

La función $\mu(x_1, x_2, \ldots, x_k)$ determinará o describirá como el valor promedio de y cambia cuando observamos un cambio de alguno de los valores del conjunto $(x_1, x_2, \ldots, x_k)$. Por ejemplo, cuando analizamos el efecto de la educación, la expericia o el coeficiente IQ en el salario podríamos establecer:

$$\mathbb{E}[Salario|Educacion, Experiencia, IQ]$$

Esta expresión describe el salario promedio dadas las variables aleatorias mencionadas. Ahora bien, tomemos una definición del concepto de ecuación

de regresión y veamos como se relaciona con el concepto de esperanza condicional.

Definición 1.2 *Sea $f(x_1, x_2, \ldots, x_K, y)$ una función de densidad de probabilidad conjunta de $K+1$ variables aleatorias, la regresión multivariada consiste en determinar la función de densidad condicional de y dado $X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K$ y entonces evaluar la integral (o la suma en el caso discreto) dada por:*

$$\begin{aligned} \mu_{y|x_1,x_2,\ldots,x_K} &= \mathbb{E}[y|x_1, x_2, \ldots, x_K] \\ &= \int_{-\infty}^{\infty} y \cdot w(y|x_1, x_2, \ldots, x_k)dy \end{aligned}$$

Donde:

$$w(y|x_1, x_2, \ldots, x_k) = \frac{f(x_1, x_2, \ldots, x_k, y)}{g(x_1, x_2, \ldots, x_k)}$$

$$g(x_1, x_2, \ldots, x_k) = \int_D f(x_1, x_2, \ldots, x_k, y)dy$$

Visto de esta manera es claro que la Definición 1.1 es equivalente a la Defición 1.2. En nuestro caso particular para propósito de este curso asumiremos la siguiente equivalencia de la esperanza condicional:

$$\mu_{y|x_1,x_2,\ldots,x_K} = \mu(x_1, x_2, \ldots, x_K) = \mathbb{E}[y|x_1, x_2, \ldots, x_K]$$

Note que la expresión anterior implica, como mostraremos más adelante, que $\mathbb{E}[\varepsilon|x_1, x_2, \ldots, x_K x_1, x_2, \ldots, x_K] = \mathbb{E}[\varepsilon] = 0$. Veamos algunos ejemplos de las formas funcionales de la esperanza condicional $\mu(x_1, x_2, \ldots, x_K)$:

1. $\mathbb{E}[y|x_1, x_2, \ldots, x_K] = x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K$

2. $\mathbb{E}[y|x_1, x_2, \ldots, x_K] = x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K + x_{K+1}^2\beta_{K+1}$

3. $\mathbb{E}[y|x_1, x_2, \ldots, x_K] = e^{\beta_1 + ln(x_2)\beta_2 + \ldots + ln(x_K)\beta_K}$, con $y > 0$ y $x_k > 0$, para todo $k = 1, 2, \ldots, K$

1.3. Efectos Marginales

Una vez definida la esperanza condicional podemos, asu vez, definir el concepto de Efecto Parcial o Efecto Marginal de y –variable dependiente– cuando existe algún cambio de sólo una de las variables explicativas en el vector $(x_1, x_2, \ldots, x_K)$, manteniendo el resto de las variables constantes.

Definición 1.3 *Decimos que el Efecto Marginal de una variable x_k como la derivada parcial dada por:*

$$\begin{aligned} EMg_k &= \frac{\partial \mathbb{E}[y|x_1, x_2, \ldots, x_K]}{\partial x_k} \\ &\approx \frac{\Delta \mathbb{E}[y|x_1, x_2, \ldots, x_K]}{\Delta x_k} \end{aligned}$$

Cuando la variable x_k es discreta o, más concretamente, dicotómica, por ejemplo, tomando valores de 0 o 1, el Efecto Marginal estará definido como sigue:

Definición 1.4 *Decimos que el Efecto Marginal de una variable dicotómica $x_1 = \{0, 1\}$ (caso extendible a cualquiera de la variables explicativas) como la diferencia dada por:*

$$EMg_1 = \mathbb{E}[y|x_1 = 1, x_2, \ldots, x_K] - \mathbb{E}[y|x_1 = 0, x_2, \ldots, x_K]$$

Las Definiciones 1.3 y 1.4 pueden particularizarse y extenerse en casos como el siguiente. Asumamos que el valor esperado tiene la forma funcional siguiente:

$$\mathbb{E}[y|x_1, x_2, \ldots, x_K] = e^{\beta_1 + \beta_2 ln(x_2) + \ldots \beta_K ln(x_K)}$$

En este caso podríamos hacer la siguiente transformación:

$$ln(\mathbb{E}[y|x_1, x_2, \ldots, x_K]) = \beta_1 + \beta_2 ln(x_2) + \ldots \beta_K ln(x_K)$$

De esta manera podríamos aproximar el Efecto Marginal de $ln(x_2)$ como:

$$\begin{aligned} \frac{\partial ln(\mathbb{E}[y|x_1, x_2, \ldots, x_K])}{\partial ln(x_2)} &= \beta_2 \\ &= \frac{\frac{\partial \mathbb{E}[y|x_1, x_2, \ldots, x_K]}{\mathbb{E}[y|x_1, x_2, \ldots, x_K]}}{\frac{\partial x_2}{x_2}} \end{aligned} \tag{1.4}$$

Así resulta evidente que la ecuación (1.4) se puede interpretar como una elasticidad.

1.3.1. Ejericios

Para finalizar establezcamos una serie de ejercicios para continuar con la discusión del Efecto Marginal.

1. Sean y, x_1 y x_2 variables aleatorias y una forma funcional de la esperanza condicional de la forma:

$$\mathbb{E}[y|x_1, x_2] = \beta_1 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_1 x_2$$

   Al respecto, determine el Efecto Marginal de x_1 y x_2.

2. Sean y, x_1 y $D = \{0, 1\}$ variables aleatorias y una forma funcional de la esperanza condicional de la forma:

$$\mathbb{E}[y|x_1, D] = e^{\beta_1 + \beta_2 x_1 + \gamma D}$$

   Al respecto, determine el Efecto Marginal de x_1 y $D = \{0, 1\}$.

1.4. Introducción a Gráficos Acíclicos Dirigidos o Directed Acycle Graphical

La notación de los Gráficos Acíclicos Dirigidos (DAG) representa que la causalidad va en una dirección, así para mostrar causalidad en reversa es necesario creas múltiples nodos. La simúltaneidad, tal como en el caso de las curvas de oferta y demanda, no tiene una representación directa o inmediata con los GAGs.

Otra forma plantear a los DAG es que se trata de una explicación de un fenómeno en términos de contrafactuales.

Definición 1.5 *Un DAG es una representación gráfica de una cadena de efectos causales. Los efectos causales están en sí mismos basados en algún proceso subyacente no observable.*

Los efectos causales pueden observarse en dos vías:

1. Directa: $D \longrightarrow Y$

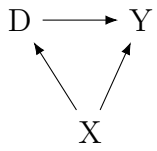


Figura 1.1: Ejemplo 1 de un DAG

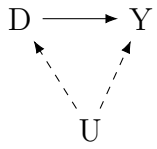


Figura 1.2: Ejemplo 2 de un DAG, confusor

2. Indirecta, a través de una tercer variable: $D \longrightarrow X \longrightarrow Y$

En este setido, la ausencia de $\longrightarrow$ indica que no hay relación entre las variables.

Un ejemplo sencillo de DAG es el siguiente:

El DAG mostrado en la Figura (1.1) ilustra que hay una ruta directa de D a Y, lo cual representa un efecto causal. Por su parte la ruta a D tiene una ruta trasera (backdoor path) a través de la ruta $D \longleftarrow X \longrightarrow Y$. En este caso consideremos que la ruta directa es un efecto causal, pero la backdoor path es no causal. En su caso este proceso crea una correlación espúrea entre D y Y.

Pensemos el proceso de ruta trasera como una situación en la que a veces cuando D toma diferentes valores, Y toma diferentes valores debido a que X toma diferentes valores. De esta forma, decimos que existe una correlación espuria entre D y Y.

Un segundo ejemplo de una DAG que nos permite ilustrar el concepto de confusor, el cual describe a una variable que no es observable, U, y que representamos su relación en el DAG con líneas punteadas de la forma que se muestra en la Figura (1.2).

De forma similar al caso anterior, existem dos formas para ir de D a Y. Existe la ruta directa de D a Y, lo cual representa un efecto causal. Por su parte la ruta a D tiene una ruta trasera a través de la ruta $D \longleftarrow U \longrightarrow Y$, pero con la diferencia de que la variable U es no observable.

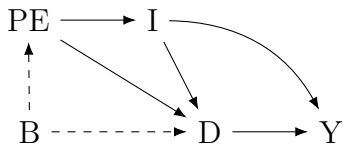


Figura 1.3: Ejemplo 3 de un DAG, ganancias de la educación

Veamos otro ejemplo. Una pregunta clásica en economía laboral es si la educación, en general, tiene el potencial de incrementar los ingresos laborales de las personas. La teoría económica indica que la educación incrementa el producto marginal del trabajo, de esta forma los trabajadores serán mejor pagados, ya que sus salarios son establecidos en función de su producto marginal en mercados competidos. Así, la teoría afirma que la educación incrementa los ingresos.

Particularicemos el ejemplo a un caso en el cual analizamos el efecto que tiene la educación secundaria en los ingresos. Partamos del hecho de que educarse a un nivel secundario no es un fenómeno aleatorio, puesto que existen múltiples factores que lo pueden explicar. Partamos de el siguiente DAG en el cual D será la variable de tratamiento o la variable que indica si una persona tiene formación de educación secundaria; Y es nuestra variable de interés o los ingresos; PE es la variable que indica el nivel educativo de los padres; I es una variable de ingreso familiar, y B son el background no observable de la persona (factores como la genética, habilidades individuales, ambiente familiar, etc.).

El DAG de la Figura 1.3 es en sí mismo una historia. Cada persona tiene un cierto background que normamente no se reporta en los conjuntos de datos. Medidas como el grado de inteligencia, la personalidad, la estabilidad emocional, capacidad de resilencia, dinámica familar y cualquier otro relacionado con factores ambientales o del entorno no se encuentran en los datos comúnmente. Por lo que se les denomina factores no observables.

Estos factores ambientales (B) están correlacionados o expresados en variables de padres e hijos. Este background causa que los padres elijan un nivel educativo o ruta de educación para los hijos y también afecta las decisiones individuales de los niños respecto de la ruta educativa que quieren seguir.

En este caso, también es posible notar que los DAG cuentan 2 historias. Nos dicen lo que está pasando, pero también nos dicen lo que no está pasando.

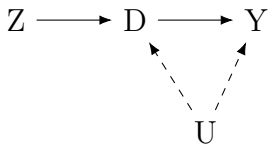


Figura 1.4: Ejemplo 4 de un DAG, IV estimator

Así, B no tiene un efecto directo en los ingresos laborales, excento a través de su efecto en la elección educativa. Sin embargo, en muchas ocasiones se suele criticar este tipo de planteamientos por parecer más un supuesto que un hecho real. En estos casos, la desición es del investigador.

Veamos un ejemplo más, el caso de las variables instrumentales. La Figura 1.4 ilustra que U son un conjunto de factores no observables que impieden indenticar la causalidad entre D y Y. De esta forma, bucamos un Z que no está correlacionado con U pero si con D con el objeto de hace una estimación auxiliar.

2

Teoría de convergencia asintótica

A continuación listamos y enunciamos una serie de definiciones relevantes para la teoría de grandes muestras. En su caso, en la medida de lo posible no haremos demostraciones de algunos de los resultados enunciados puesto que ello en sí mismo podría ser una parte signinificativa de un curso de probabilidad y estadística, lo que además implica tiempo de nuestro curso que podríamos requerir para otros temas.

2.1. Convergencia de sucesiones

Definición 2.1 *Sea una sucesión de números no aleatorios* $\{a_N : N = 1, 2, \ldots\}$*, decimos que* a_N *converge a un número* a *(es decir, tiene su límite en* a*) si* $\forall \varepsilon > 0$*,* $\exists N_\varepsilon$ *tal que si* $N > N_\varepsilon$*, entonces:*

$$|a_N - a| < \varepsilon$$

Es decir, $a_N \longrightarrow a$ *cuando* $N \longrightarrow \infty$*.*

Definición 2.2 *Sea una sucesión de números no aleatorios* $\{a_N : N = 1, 2, \ldots\}$*, decimos que la sucesión está acotada si y sólo si existe* $b < \infty$ *tal que* $|a_N| \leq b$*,* $\forall N = 1, 2, \ldots$*.*

Estas definiciones aplican de forma similar a vectores y a matrices cuando las analizamos y aplicamos elemento a elemento.

2.2. Convergencia en probabilidad y acotamientos en probabilidad

Definición 2.3 *Sea una sucesión de variables aleatorias* $\{x_N : N = 1, 2, \ldots\}$ *decimos que una sucesión de variables aleatorias es convergente en probabilidad a una constante a si* $\forall \varepsilon > 0$ *tenemos que:*

$$\lim_{N \longrightarrow \infty} P[|x_N - a| > \varepsilon] = 0$$

Alternativamente, una sucesión de variables aleatorias tiene un límite en probabilidad en una constante a si cumple la condición anterior, lo cual escribiremos como:

$$p \lim x_N = a$$

Definición 2.4 *Sea una sucesión de variables aleatorias* $\{x_N : N = 1, 2, \ldots\}$ *decimos que la sucesión está acotada en probabilidad si y sólo si para cada* $\varepsilon > 0$*, existe un* $b_\varepsilon < \infty$ *y un entero* N_ε *tal que* $\forall N \geq N_\varepsilon$ *sucede que:*

$$P[|x_N| \geq b_\varepsilon] < \epsilon$$

Todas las definiciones de esta sección aplican elemento a elemento a sucesiones de vectores o matrices de variables aleatorias. Por ejemplo, si $\{\mathbf{X}_N : N = 1, 2, \ldots\}$ es una sucesión de un vector de variables aleatorias o vector aleatorio de dimensión $K \times 1$, entonces decimos que $\mathbf{X}_N$ converge en probabilidad a $\mathbf{a}$ o $p \lim \mathbf{X}_N = \mathbf{a}$, donde $\mathbf{a}$ es un vector aleatorio de dimensión $K \times 1$, si y sólo si cada una de las entradas de la sucesión de vectores aleatorios comple que:

$$p \lim x_{Nj} = a_j$$

Para todo $j = 1, 2, \ldots, K$. Una forma equivalente de escribir este resultado es:

$$p \lim ||\mathbf{X}_N - \mathbf{a}|| = 0$$

Donde $||b|| \equiv (b'b)^{1/2}$, es decir, que el simbolo es una distancia euclideanea de de un vector de dimensión $K \times 1$.

De forma similar, supongamos que analizamos una sucesión de matrices de variables aleatorias o matrices aleatorias $\{\mathbf{Z}_N : N = 1, 2, \ldots\}$, cuya dimensión es de $M \times K$. Supongamos que existe una matriz de la misma dimensión $\mathbf{B}$, entonces podemos decir en forma equivalente que:

$$p \lim ||\mathbf{Z}_N - \mathbf{B}|| = 0$$

Donde $||A|| \equiv [tr(A'A)]^{1/2}$ y $tr(C)$ denota la traza de una matriz simetrica.

Teorema 2.1 *Sea* $g : \mathbb{R}^K \longrightarrow \mathbb{R}^J$ *una función continua en algún punto* $c \in \mathbb{R}^K$. *Sea* $\{\mathbf{X}_N : N = 1, 2, \ldots\}$ *una sucesión de vectores aleatorios de dimensión* $K \times 1$ *tal que* $p\lim \mathbf{X}_N = c$. *Entonces,*

$$\lim_{N \longrightarrow \infty} g(\mathbf{X}_N) = g(c)$$

En otras palabras:

$$p\lim g(\mathbf{X}_N) = g(p\lim \mathbf{X}_N)$$

2.3. Convergencia en distribución

Definición 2.5 *Una sucesión de variables aleatorias* $\{x_N : N = 1, 2, \ldots\}$ *converge en distribución a una variable aleatoria continua si y sólo si* $\forall \xi \in \mathbb{R}$*:*

$$\lim_{N \longrightarrow \infty} F_N(\xi) = F(\xi)$$

Donde F_N *es la función de distribución acumnulada (c.d.f.) de* x_N *y* F *es la función de distribución acumulada continua de* x. *En otras palabras, la expresión anterior se pude escribir como:*

$$x_N \xrightarrow{d} x$$

Definición 2.6 *Una sucesión de vectores aleatorios* $\{\mathbf{X}_N : N = 1, 2, \ldots\}$ *de dimensión* $K \times 1$ *converge en distribución a un vector aleatorio* $\mathbf{X}$ *de dimennsión* $K \times 1$ *si y sólo si para cualquier vector aleatorio* $\mathbf{C}$ *de dimennsión* $K \times 1$ *sucede que:* $\mathbf{C}'\mathbf{C} = 1$, *entonces:*

$$\mathbf{C}'\mathbf{X}_N \xrightarrow{d} \mathbf{C}'\mathbf{X}$$

o simplemente:

$$\mathbf{X}_N \xrightarrow{d} \mathbf{X}$$

Por ejemplo, cuando $\mathbf{X} \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, podemos decir que si:

$$\mathbf{C}'\mathbf{X}_N \xrightarrow{d} Normal(\mathbf{C}'\boldsymbol{\mu}, \mathbf{C}'\boldsymbol{\Sigma}\mathbf{C})$$

Para cada $\mathbf{C} \in \mathbb{R}^K$ tal que $\mathbf{C}'\mathbf{C} = 1$, en este caso escribiremos que:

$$\mathbf{X}_N \xrightarrow{d} Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

2.4. Teoremas límite para muestras aleatorias

Teorema 2.2 *Sea* $\{\mathbf{W}_N : N = 1, 2, \ldots\}$ *una sucesión de vectores aleatorios de dimensión* $K \times 1$ *que provienen de una población idénticamente distribuidos tal que:* $\mathbb{E}[|w_{Nk}|] < \infty$*, con* $k = 1, 2, \ldots K$*. Entonces, la sucesión satisface Ley Débil del los Grandes Números:*

$$N^{-1} \sum_{i=1}^{N} \mathbf{W}_i \xrightarrow{p} \boldsymbol{\mu}_w$$

Donde $\boldsymbol{\mu} \equiv \mathbf{E}[\mathbf{W}_i]$

Teorema 2.3 *Sea* $\{\mathbf{W}_N : N = 1, 2, \ldots\}$ *una sucesión de vectores aleatorios de dimensión* $K \times 1$ *independientes e idénticamente distribuidos tal que:* $\mathbb{E}[w_{Nk}^2] < \infty$*, con* $k = 1, 2, \ldots K$*, y* $\mathbb{E}[w_i] = 0$*. Entonces,* $\{\mathbf{W}_N : N = 1, 2, \ldots\}$ *satisface el Teorema del Límite Central en el siguiente sentido:*

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{W}_i \xrightarrow{d} Normal(\mathbf{0}, \mathbf{B})$$

Donde $\mathbf{B} = Var(\mathbf{W}_N) = \mathbb{E}[\mathbf{W}_N \mathbf{W}_N']$ *es una matriz positiva definida.*

2.5. Estimadores y estadísticas de prueba

Definición 2.7 *Sea* $\{\hat{\boldsymbol{\theta}}_N : N = 1, 2, \ldots\}$ *una sucesión de estimadores dados y acomodados en un vector* $P \times 1$*. Sea* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, donde* N *indexa el tamaño de la muestra. Si* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ *para cualquier valor de* $\boldsymbol{\theta}$*, entonces decimos que* $\hat{\boldsymbol{\theta}}_N$ *es un estimador consistente de* $\boldsymbol{\theta}$*.*

Definición 2.8 *Sea* $\{\hat{\boldsymbol{\theta}}_N : N = 1, 2, \ldots\}$ *una sucesión de estimadores dados y acomodados en un vector* $P \times 1$*. Sea* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, donde* N *indexa el tamaño de la muestra. Asumamos que:*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} Normal(\mathbf{0}, \mathbf{V})$$

Donde $\mathbf{V}$ *es una matriz semidefinida positiva de dimensión* $P \times P$*. Entonces, decimos que* $\hat{\boldsymbol{\theta}}_N$ *es* $\sqrt{N}$ *- asintóticamente distribuida de forma normal y* $\mathbf{V}$ *es la varianza asintótica de* $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$ *denotada como* $AVar\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) = \mathbf{V}$*.*

Teorema 2.4 *Supongamos que* $\mathbf{V}$ *es una matriz definida positiva. Entonces, para cualquier matriz no estocástica* $\mathbf{R}$ *de dimensión* $Q \times P$*, con* $Q \leq P$ *con* $Rango(\mathbf{R}) = Q$*, sucede que:*

$$\sqrt{N}\mathbf{R}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} Normal(\mathbf{0}, \mathbf{RVR}')$$

y que:

$$[\sqrt{N}\mathbf{R}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})]'[\mathbf{RVR}']^{-1}[\sqrt{N}\mathbf{R}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})] \xrightarrow{d} \chi^2_Q$$

3

Estimación de modelos lineales de una sola ecuación por el método de Mínimos Cuadrados Ordinarios

3.1. El estimador de Mínimos Cuadrados Ordinarios (MCO) y el análisis clásico de regresión

Si partimos de las ecuaciones (1.2) u (1.3), podemos establecer que el término de error estará dado por:

$$\varepsilon_i = y_i - \mathbf{X}_i'\boldsymbol{\beta} \tag{3.1}$$

Donde $i = 1, 2, \ldots, N$. De forma similar podremos decir que un estimador de éste término de error será aquel que resulte de:

$$e_i = y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}} \tag{3.2}$$

Donde $\hat{\boldsymbol{\beta}}$ es un vector de estimadores de los parámetros $\boldsymbol{\beta}$. De lo dicho hasta ahora es fácil ver que la siguiente ecuación es cierta $\forall i$:

$$y_i = \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + e_i \tag{3.3}$$

Intuitivamente, la ecuación (3.3) significa que siempre que poseamos una muestra de los elementos de la población, podremos explicar una parte de la variable dependiente, no su totalidad. En este sentido, el análisis de regresión consiste en un proceso de ajuste a la variable dependiente. Está es la idea que da origen al R^2 y otras medidas de bondad de ajuste, mismas que se analizan en textos convencionales de análisis de regresión.

El método de MCO, en consecuencia, resulta en encontrar la combinación de valores de los estimadores de los parámetros $\hat{\boldsymbol{\beta}}$ que permita minimizar la suma de los residuales (estimadores de los términos de erro $\boldsymbol{\varepsilon}$) al cuadrado dada por:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{X}_i' \hat{\boldsymbol{\beta}})^2 \tag{3.4}$$

Donde $\hat{\boldsymbol{\beta}}$ denota el vector de estimadores $\hat{\beta}_1, \ldots, \hat{\beta}_K$. En términos matriciales, dado que $(e_1, e_2, \ldots, e_n)'(e_1, e_2, \ldots, e_n) = \mathbf{e'e}$, el problema del método de MCO consiste en resolver el problema de óptimización:

$$\begin{aligned} Minimizar_{\hat{\boldsymbol{\beta}}} S(\hat{\boldsymbol{\beta}}) &= Minimizar_{\hat{\boldsymbol{\beta}}} \mathbf{e'e} \\ &= Minimizar_{\hat{\boldsymbol{\beta}}} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

Expandiendo la expresión $\mathbf{e'e}$ obtenemos:

$$\mathbf{e'e} = \mathbf{Y'Y} - 2\mathbf{Y'X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X'X}\hat{\boldsymbol{\beta}} \tag{3.5}$$

De esta forma obtenemos que las condiciones necesarias de un mínimo son:

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X'Y} + 2\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{0} \tag{3.6}$$

De ecuación anterior obtenemos para la solución del problema del mínimo a las ecuaciones siguientes conocidas como *ecuaciones normales* dadas por:

$$\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'Y} \tag{3.7}$$

Notemos que dichas ecuaciones normales son en realidad un sistema de ecuaciones de K variables o incógnitas. Por un lado, recordemos que $\mathbf{X}$ es una matriz de dimensión $N \times K$, con lo cual $\mathbf{X'}$ es de dimensión $K \times N$. Así, el producto $\mathbf{X'X}$ dará como resultado una matriz cuadrada de dimensión $K \times K$. Por otro lado, sabemos que $\mathbf{Y}$ es un vector de tamaño $N \times 1$, con

lo cual el producto $\mathbf{X}'\mathbf{Y}$ da como resultado un vector de dimención $K \times 1$. En conclusión, el sistema de ecuaciones normales consiste en K ecuaciones con K incógnitas $(\hat{\beta}_1, \ldots, \hat{\beta}_K)$. Ante este hecho, existen múltiples formas mediante las cuales se puede solucionar dicho sistema, sin embargo en nuesto caso seguiremos el siguiente procedimiento de operaciones matriciales.

Si las matriz $\mathbf{X}$ es de rango completo por columnas, entonces la inversa de la matriz $\mathbf{X}'\mathbf{X}$ existe. De esta forma, la solución esta dada por la siguiente expresión:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.8}$$

Esta expresión, a pesar de ser en apariencia compleja se puede ver como un conjunto de sumas. En general hemos supuesto que nuestra regresión a estimar esta descrita por la eccuación: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{iK}\beta_K + \varepsilon_i$, de esta forma tenemos K variables independientes es nuestra regresión.

Ahora bien, si denotamos a $\mathbf{X}_k$ como el vector columna formado por todas las observaciones de la muestra $(i = 1, 2, \ldots, N)$ para la variable k, podemos decir que la matriz $\mathbf{X}$ que contiene todas las variable independientes se forma por la concatenación de cada uno de los K vectores columna. Dicho esto, podemos ver que las matrices $\mathbf{X}$ y $\mathbf{X}'$ se pueden expresar como:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1K} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \ldots & x_{NK} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 & \ldots & \mathbf{X}_K \end{bmatrix}$$

$$\mathbf{X}' = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \ldots & x_{N1} \\ x_{12} & x_{22} & x_{32} & \ldots & x_{N2} \\ x_{13} & x_{23} & x_{33} & \ldots & x_{N3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1K} & x_{2K} & x_{3K} & \ldots & x_{NK} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \mathbf{X}'_3 \\ \vdots \\ \mathbf{X}'_K \end{bmatrix}$$

Si suponemos que nuestra regresión tiene una constante, la especificación sería: $y_i = \beta_1 + x_{i2}\beta_2 + \ldots + x_{iK}\beta_K + \varepsilon_i$, con unas matrices $\mathbf{X}$ y $\mathbf{X}'$ dadas:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & x_{13} & \ldots & x_{1K} \\ 1 & x_{22} & x_{23} & \ldots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N2} & x_{N3} & \ldots & x_{NK} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_N & \mathbf{X}_2 & \mathbf{X}_3 & \ldots & \mathbf{X}_K \end{bmatrix}$$

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{12} & x_{22} & x_{32} & \dots & x_{N2} \\ x_{13} & x_{23} & x_{33} & \dots & x_{N3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1K} & x_{2K} & x_{3K} & \dots & x_{NK} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'_N \\ \mathbf{X}'_2 \\ \mathbf{X}'_3 \\ \vdots \\ \mathbf{X}'_K \end{bmatrix}$$

Donde $\mathbf{1}_N$ es un vector columna compuesto de 1's (unos). Retomando la ecuación (3.8), desarrollemos cada uno de los casos anteriores, así obtenemos lo siguiente para el caso general:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \mathbf{X}'_3 \\ \vdots \\ \mathbf{X}'_K \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 & \dots & \mathbf{X}_K \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}'_1\mathbf{X_1} & \mathbf{X}'_1\mathbf{X}_2 & \mathbf{X}'_1\mathbf{X}_3 & \dots & \mathbf{X}'_1\mathbf{X}_K \\ \mathbf{X}'_2\mathbf{X}_1 & \mathbf{X}'_2\mathbf{X}_2 & \mathbf{X}'_2\mathbf{X}_3 & \dots & \mathbf{X}'_2\mathbf{X}_K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{X}'_K\mathbf{X}_1 & \mathbf{X}'_K\mathbf{X}_2 & \mathbf{X}'_K\mathbf{X}_3 & \dots & \mathbf{X}'_K\mathbf{X}_K \end{bmatrix}$$

Por lo tanto, obtenemos que:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum_{i=1}^N x_{i1}^2 & \sum_{i=1}^N x_{i1}x_{i2} & \sum_{i=1}^N x_{i1}x_{i3} & \dots & \sum_{i=1}^N x_{i1}x_{iK} \\ \sum_{i=1}^N x_{i2}x_{i1} & \sum_{i=1}^N x_{i2}^2 & \sum_{i=1}^N x_{i2}x_{i3} & \dots & \sum_{i=1}^N x_{i2}x_{iK} \\ \sum_{i=1}^N x_{i3}x_{i1} & \sum_{i=1}^N x_{i3}x_{i2} & \sum_{i=1}^N x_{i3}^2 & \dots & \sum_{i=1}^N x_{i3}x_{iK} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^N x_{iK}x_{i1} & \sum_{i=1}^N x_{iK}x_{i2} & \sum_{i=1}^N x_{iK}x_{i3} & \dots & \sum_{i=1}^N x_{iK}^2 \end{bmatrix}$$

Por otro lado, cuando supongamos que existe un término constante:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum_{i=1}^N x_{i2} & \sum_{i=1}^N x_{i3} & \dots & \sum_{i=1}^N x_{iK} \\ \sum_{i=1}^N x_{i2} & \sum_{i=1}^N x_{i2}^2 & \sum_{i=1}^N x_{i2}x_{i3} & \dots & \sum_{i=1}^N x_{i2}x_{iK} \\ \sum_{i=1}^N x_{i3} & \sum_{i=1}^N x_{i3}x_{i2} & \sum_{i=1}^N x_{i3}^2 & \dots & \sum_{i=1}^N x_{i3}x_{iK} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^N x_{iK} & \sum_{i=1}^N x_{iK}x_{i2} & \sum_{i=1}^N x_{iK}x_{i3} & \dots & \sum_{i=1}^N x_{iK}^2 \end{bmatrix}$$

Adicionalmente, el producto $\mathbf{X}'\mathbf{Y}$, en el caso general, se puede expresar como:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \mathbf{X}_3' \\ \vdots \\ \mathbf{X}_K' \end{bmatrix} \mathbf{Y} = \begin{bmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \\ \mathbf{X}_3'\mathbf{Y} \\ \vdots \\ \mathbf{X}_K'\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} x_{i1}y_i \\ \sum_{i=1}^{N} x_{i2}y_i \\ \sum_{i=1}^{N} x_{i3}y_i \\ \vdots \\ \sum_{i=1}^{N} x_{iK}y_i \end{bmatrix} \tag{3.9}$$

Si en el modelo suponemos la existencia de un término constante, dicho producto se expresa como:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \mathbf{1}_N' \\ \mathbf{X}_2' \\ \mathbf{X}_3' \\ \vdots \\ \mathbf{X}_K' \end{bmatrix} \mathbf{Y} = \begin{bmatrix} \mathbf{1}_N'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \\ \mathbf{X}_3'\mathbf{Y} \\ \vdots \\ \mathbf{X}_K'\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i2}y_i \\ \sum_{i=1}^{N} x_{i3}y_i \\ \vdots \\ \sum_{i=1}^{N} x_{iK}y_i \end{bmatrix}$$

Finalmente, para que esta solución dada para el procedimiento de MCO y mostrada en la ecuación (3.8) sea un mínimo, debemos buscar las condiciones de segundo orden:

$$\frac{\partial^2 S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X} \tag{3.10}$$

Donde la matriz $\mathbf{X}'\mathbf{X}$ debe ser positiva definida para que la solución de MCO sea un mínimo. Sea $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$ para algún vector $\mathbf{c}$ distinto de cero. Entonces:

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2, \text{ donde } \mathbf{v} = \mathbf{X}\mathbf{c}$$

Así, q es positivo. Si $\mathbf{v}$ fuera cero, entonces existe una combinación lineal de las columnas de $\mathbf{X}$ que da como resultado cero, lo cual contradice el supuesto de que $\mathbf{X}$ es de rango completo. En todos los casos, si $\mathbf{X}$ es de rango completo, entonces la solución del método de MCO, $\hat{\boldsymbol{\beta}}$, es la única que mínimiza la suma de los residuales al cuadrado.

3.2. Bondad de ajuste

Una vez planteada la solución de MCO plantearemos una medida para determinar en qué grado los datos estimados, definidos como $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, se

ajustan al valor real de $\mathbf{Y}$. La medida o métrica es el R^2, la cual contrasta el análisis de regresión respecto de hacer una simple estimación de una media para dar un pronóstico de $\mathbf{Y}$. El coeficente R^2 está montado en el supuesto de que el modelo incluye un término constate. Si el modelo no incluye una constante, no es posible hacer una interpretación del R^2, en los siguientes parráfos abundaremos al respecto.

Antes de iniciar el desarrollo del R^2, partamos de que la suma de residuales es igual a cero ($\sum_{i=1}^{n} e_i = 0$), si y sólo si, el modelo tiene un término constante. Un corololario de este hecho es que el valor promedio de los residuales es cero, sí y solo si, el modelo tiene un término constante, esto es, la única forma en que se cumpla la siguiente expresión es que la suma de residuales sea cero:

$$\begin{aligned}
\overline{e} &= \frac{\sum_{i=1}^{n} e_i}{n} \\
&= \frac{1}{n} \cdot \left[\begin{array}{cccccc} 1, & 1, & 1, & \ldots & 1 \end{array} \right] \left[\begin{array}{c} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{array} \right] \\
&= 0
\end{aligned}$$

Condición que resulta de las ecuaciones normales de MCO, recordemos que ellas se derivan de que $\mathbf{X'e} = \mathbf{0}$, y que el primer producto punto de vectores implica a la columna de la constante en la matriz $\mathbf{X}$ y que está definido por $\left[\begin{array}{cccccc} 1, & 1, & 1, & \ldots & 1 \end{array} \right]$.

El R^2 descompone la variación total en dos tramos: la variación originada por la regresión y la variación originada por el término de error. Como primer paso definamos la variación total observada respecto de la media ($\overline{Y}$) como:

$$\sum_{i=1}^{n} (y_i - \overline{Y})^2 = \left[\begin{array}{ccccccc} y_1 - \overline{Y}, & y_2 - \overline{Y}, & y_3 - \overline{Y}, & \ldots & y_n - \overline{Y} \end{array} \right] \left[\begin{array}{c} y_1 - \overline{Y} \\ y_2 - \overline{Y} \\ y_3 - \overline{Y} \\ \vdots \\ y_n - \overline{Y} \end{array} \right]$$

$$
\begin{aligned}
&= (\mathbf{Y}-\overline{\mathbf{Y}})'(\mathbf{Y}-\overline{\mathbf{Y}}) \\
&= \left[\mathbf{Y}-\frac{1}{n}\begin{bmatrix}1&1&\dots&1\\1&1&\dots&1\\1&1&\dots&1\\\vdots&\vdots&\dots&\vdots\\1&1&\dots&1\end{bmatrix}\mathbf{Y}\right]'\left[\mathbf{Y}-\frac{1}{n}\begin{bmatrix}1&1&\dots&1\\1&1&\dots&1\\1&1&\dots&1\\\vdots&\vdots&\dots&\vdots\\1&1&\dots&1\end{bmatrix}\mathbf{Y}\right] \\
&= \mathbf{Y}'\left[\mathbb{I}_n-\frac{1}{n}\begin{bmatrix}1&1&\dots&1\\1&1&\dots&1\\1&1&\dots&1\\\vdots&\vdots&\dots&\vdots\\1&1&\dots&1\end{bmatrix}\right]'\left[\mathbb{I}_n-\frac{1}{n}\begin{bmatrix}1&1&\dots&1\\1&1&\dots&1\\1&1&\dots&1\\\vdots&\vdots&\dots&\vdots\\1&1&\dots&1\end{bmatrix}\right]\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{M}^0\mathbf{Y} \qquad (3.11)
\end{aligned}
$$

En la expresión (3.11) a $\mathbf{M}^0$ se puede interpretar como una matriz que resta a una matriz o vector su promedio, notemos que como resultado dará las desviaciones respecto de la media. Asimismo, $\mathbf{M}^0$ tiene un par de propiedades que son fácilmente demostrables: idémpotencia y simetría, es decir, i) $\mathbf{M}^0 = \mathbf{M}^{0'}$ y ii) $\mathbf{M}^0\mathbf{M}^0 = \mathbf{M}^0$. Dicho lo anterior, retomemos a (3.11) para mostrar que:

$$
\begin{aligned}
\mathbf{Y}'\mathbf{M}^0\mathbf{Y} &= \mathbf{Y}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{Y} \\
&= (\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e})'\mathbf{M}^{0'}\mathbf{M}^0(\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}) \\
&= (\hat{\boldsymbol{\beta}}'\mathbf{X}'+\mathbf{e}')\mathbf{M}^{0'}\mathbf{M}^0(\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}) \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{e}+\mathbf{e}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{e} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{e}+\mathbf{e}'\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{e} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{e} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{e} \qquad (3.12)
\end{aligned}
$$

No debe pasar desapercibido que el desarrollo algebraíco para llegar a la ecuación (3.12) sólo es posible si el promedio de los residuales es cero, es

decir:

$$\mathbf{M}^0\mathbf{e} = \mathbf{e} - \begin{bmatrix} \bar{e} \\ \bar{e} \\ \bar{e} \\ \vdots \\ \bar{e} \end{bmatrix} = \mathbf{e} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{e} \tag{3.13}$$

De otra forma, no se puede concluir la expresión (3.12). Finalmente, (3.12) la expresaremos así:

$$\mathbf{Y}'\mathbf{M}^0\mathbf{Y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}'\mathbf{e} \tag{3.14}$$

Así, como mencionamos anteriormente, (3.14) se puede interpretar como que la variación total respecto de la media se puede descomponer en dos variaciones, una que se origina de la regresión, $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}$, y otra de los residuales, $\mathbf{e}'\mathbf{e}$. De esta forma planteamos que R^2 es una metrica que cuantifica cuánto de la variación total es explicada por la regresión y cuánto es explicada por los residuales (es decir, por la información no observable):

$$\begin{aligned} R^2 &= \frac{Var.Regresion}{Var.Total} \\ &= \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \\ &= \frac{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \\ &= 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \end{aligned} \tag{3.15}$$

Existe una expresión más que se le denomina como un R^2 ajustado o $R^2 Adj.$, el cual castiga por un uso excesivo de variables independientes:

$$R^2 Adj. = 1 - \frac{n-1}{n-K}\frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \tag{3.16}$$

Así, es obvio que la siguiente relación siempre es cierta: $R^2 \geq R^2 Adj.$.

3.3. Modelos lineales y algunas de sus propiedades

Determinado una estimado de MCO dado por la ecuación (3.8), establezcamos el modelo poblacional que es lineal en sus parámetros y que está

dado por (notése que eliminanos los subindices i para facilitar la exposición–digamos, por economía de pizarrón):

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{3.17}$$

Donde y y $x_1, x_2, \ldots, x_K$ son variables aleatorias observables, ε es un término de error o perturbación no observable no observable, y $\beta_1, \beta_2, \ldots, \beta_K$ son los parámetros (constantes) que deseamos estimar.

La ecuación (3.17) es un modelo estructural estimable, el cual en ocasiones se obtiene directamente de una modelación teórica y otras veces deriva de un procedimiento algebraico para obtenerlo.

El término de error ε puede ser causa de una serie de fuentes relaciondas con: (1) información no observable, (2) variables relevantes omitidas y (3) errores de medición (que en el fondo implica la omisión de información).

Asumimos que para estimar consistentemente a $\boldsymbol{\beta}$ a través de una muestra aleatoria de la población, el término de error $\boldsymbol{\varepsilon}$ tiene una **media cero** que no está correlacionada con el resto de los regresosres, es decir:

$$\begin{aligned} \mathbb{E}[\varepsilon_i] &= 0 \\ Cov(x_k, \varepsilon_i) &= 0 \end{aligned} \tag{3.18}$$

Donde $k = 1, 2, \ldots, K$, e $i = 1, 2, \ldots, n$. Dada la ecuación de regresión y el supuesto anterior en la ecuación (3.18), entonces podemos ver este caso como una proyección lineal de y en el vector formado por $(x_1, x_2, \ldots, x_K)$ y descrito como: $x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K$.

Partiendo de que en la ecucación (3.18) establecimos que el análisis de regresión deberá cumplir con $\mathbb{E}[\varepsilon_i] = 0$, entonces podemos establecer que:

$$\mathbb{E}[\varepsilon | x_1, x_2, \ldots, x_K] = \mathbb{E}[\varepsilon | \mathbf{x}] = 0 \tag{3.19}$$

De esta forma podemos decir, utilizando la definición de función de regresión discutida en el capítulo previo tenemos que:

$$\mathbb{E}[y | \mathbf{x}] = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K \tag{3.20}$$

Dicho esto, podemos hacer algunas observaciones respecto del efecto de las variables utilizadas. Decimos que una variable explicativa x_k es endogena en la ecuación de regresión si esta no cumple con la ecuación (3.18), es decir, está correlacionada con el término de error, ε. Por el contrario, si x_k no está

correlacionado con ε, entonces se dice que x_k es una varaible exógena. De esta forma, la ecuación (3.19) se conoce como el supuesto de exogeneidad.

El supuesto de enxogenidad o el de endogenidad (depende de cómo lo queramos ver) tiene algunas aristas que debemos analizar.

1. La primera es el problema de variables omitidas. Supongamos que la ecuación de regresión está dada por una expresión como $\mathbb{E}[y|\mathbf{x}, q]$. No obstante, asumamos que q es una variable que no es posible observarse, de forma que nos vemos obligados a estimar una ecuación de regresión del tipo $\mathbb{E}[y|\mathbf{x}]$. En este caso q formará parte del término de error ε. Si q y x_k son variables que están correlacionadas, entonces, tendremos el problema de que x_k será endogena.

   Este tipo de situaciones, en las que existe correlación entre varaibles no observables y variables explicativas son conocidas como problemas de autoselección. En el capítulo de modelos de selección detallaremos más al respecto de este tipo de problemas.

2. La segunda es el problema de simultaniedad. que es un problema que aparece cuando una o más varaibles explicativas están determinadas simultáneamente con y. Así, si x_k esta parcialmente determinada por y, entonces, x_k y ε estarán correlacionados. En otro capítulo más adelante retomaremos este caso.

3.4. Propiedades asintóticas de los estimadores de MCO

Partamos de una forma funcional general de la ecuación de regresión dada por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Así para algún individuo y sin hacer uso del subíndice i, tendríamos:

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon \tag{3.21}$$

Donde $\mathbf{x}$ es un vector de regresores de dimensión $1 \times K$ y los parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)'$ es un vector de dimensión $K \times 1$. Dado que la mayoría de las veces asumiremos que el modelo considera un término constantes, diremos que $x_1 = 1$.

Asumamos que podemos obtener una muestra aleatoria de tamaño N de la población con el objeto de estimar $\boldsymbol{\beta}$ a partir de una familia de vectores $\{(\mathbf{x}, y_i) : i = 1, 2, \ldots, N\}$, que es una muestra aleatoria independiente e idénticmanete distribuida.

Podemos decir que para cada observación $i = 1, 2, \ldots, N$ tenemos que:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \tag{3.22}$$

De esta ecuación ahora discutamos algunas propiedades deseables que deberán complir los estimadores $\boldsymbol{\beta}$.

La primera es **consistencia**, la cual es una condición de ortogonalidad de la población, la cual implica:

$$\mathbb{E}[\mathbf{x}_i\varepsilon_i] = 0 \tag{3.23}$$

Donde $i = 1, 2, \ldots, N$. Si asumimos que el modelo tiene un término constante, entonces, esto implcaría que $\mathbb{E}[\varepsilon_i] = 0$. La segunda es una condición de **no multicolinealidad**:

$$Rank\mathbb{E}[\mathbf{X}'\mathbf{X}] = K \tag{3.24}$$

Así, buscamos que no exista dependencia lineal entre los elementos de $\mathbf{x}_i'\mathbf{x}_i$. Esto debería ponernos a pensar en qué pasa con casos en que incluimos una variable y su cuadrado, o cuando en en un modelo que incluye constante decidimos agregar variables dicotómicas con poca varibilidad (e.g. casí todos los individuos son de un tipo), entre otros casos.

Bajo este par de supuestos de las ecuaciones (3.23) y (3.23) se cumplirán los siguientes casos. Diremos que $\boldsymbol{\beta}$ está identificada, en términos de que se trata de parámetros lineales estimados a partir de una muestra aleatoria tales que $\boldsymbol{\beta}$ puede ser escrita en términos de los momentos poblacionales, es decir, partiendo de la estimación de estos dada por MCO tendríamos:

$$\hat{\boldsymbol{\beta}} = \mathbb{E}[\mathbf{x}_i'\mathbf{x}_i]^{-1}\mathbb{E}[\mathbf{x}_i'y_i] \tag{3.25}$$

Dado que está expresión es válida para cada elementos de la muestra y que ilustra la dependencia de que $\boldsymbol{\beta}$ sea identificable, podemos extender el

análisis para obtener:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \mathbb{E}[\mathbf{X}'\mathbf{X}]^{-1}\mathbb{E}[\mathbf{X}'\mathbf{Y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbb{E}[\mathbf{X}'\mathbf{Y}] \\
&= \boldsymbol{\beta} + \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] \\
&= \boldsymbol{\beta} + \left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i}{N}\right)^{-1}\left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon}}{N}\right) \\
&= \boldsymbol{\beta}
\end{aligned}$$

El último paso es válido Si tomamos el límite en probabilidad $p\lim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + A^{-1} \times 0 = \boldsymbol{\beta}$. Lo anterior implica que hemos demostrado el sigueinte teorema:

Teorema 3.1 *Bajo los supuestos de las ecuaciones (3.23) y (3.24) el estimador $\hat{\boldsymbol{\beta}}$ obtenido por MCO a partir de una muestra aleatoria es consistente para $\boldsymbol{\beta}$.*

Este teorema implica que $\mathbf{X}\boldsymbol{\beta}$ es una proyección ponderada de $\mathbf{X}$ sobre $\mathbf{Y}$. Más aún, implica que:

$$\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \boldsymbol{\beta} \tag{3.26}$$

3.5. Inferencia asintótica bajo MCO

Partamos de que:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i}{N}\right)^{-1}\left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon}}{N}\right) \\
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i}{N}\right)^{-1}\left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon}}{N^{1/2}N^{1/2}}\right) \\
\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i}{N}\right)^{-1}\left(\sqrt{N}\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon}\right)
\end{aligned}$$

Entonces por el Teorema del Límite Central, tenemos que:

$$\sqrt{N}\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon} \xrightarrow{d} Normal(\mathbf{0}, \Sigma) \tag{3.27}$$

Donde Σ es una matriz de dimensión $K \times K$ que resulta de:

$$\Sigma = \mathbb{E}[\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}]$$

En este momento introduciremos uno de los supuestos más importantes que es el de **homocedasticidad**, por lo que:

$$\Sigma = \mathbb{E}[\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}] = \sigma^2\mathbb{E}[\mathbf{X}'\mathbf{X}] \tag{3.28}$$

Donde $\sigma^2 = \mathbb{E}[\varepsilon_i]$ para todo $i = 1, 2, \ldots, N$. En general vamos a decir que $\mathbb{E}[\varepsilon_i|\mathbf{X}] = Var[\varepsilon_i|\mathbf{X}] = \sigma^2$. Dicho lo anterior vamos a enunciar un teorema:

Teorema 3.2 ***Normalidad asintótica de MCO***. *Considerando los supuestos descritos en las ecuaciones (3.23) y (3.28), entonces:*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim Normal(\mathbf{0}, \sigma^2(\mathbb{E}[\mathbf{X}'\mathbf{X}])^{-1}) \tag{3.29}$$

El teorema anterior se puede interpretar como:

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\sim Normal\left(\mathbf{0}, \sigma^2\left(\frac{\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i}{N}\right)^{-1}\right) \\
(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\sim Normal\left(\mathbf{0}, \sigma^2\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\right) \\
(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\sim Normal\left(\mathbf{0}, \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}\right) \\
\hat{\boldsymbol{\beta}} &\sim Normal\left(\boldsymbol{\beta}, \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}\right)
\end{aligned} \tag{3.30}$$

A partir de la expresión (3.30) contruiremos algunas pruebas. En estadística es común que porpongamos como estimador de σ^2 a:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K} \tag{3.31}$$

El cual es un estimador insesgado de la varianza. Además, queda pendiente en esta sección es demostrar que el estimador de MCO ($\hat{\boldsymbol{\beta}}$) alcanza la cota inferior Cramér-Rao. De esta forma podríamos afirmar que el estimador es de mínima varianza o, equivalentemente, es el más eficiente.

Dicho lo anterior podemos construir las dos pruebas: t y F. Previo a dicha construcción debemos recordar que una prueba t se construlle por la relación

de una función normal con media cero y variaza 1 (uno), y la raíz cuadrada de una función chi-cuadrado. Supongamos que una variable Z se distribuye de forma normal con media cero y variaza 1 (uno), y que la variable χ^2_m se distribulle de forma chi-cuadrado con m grados de libertad, entonces:

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{m}}} \sim t_m \tag{3.32}$$

Donde t_m tal y como se describe en (3.32) es una distribución t de Student con m grados de libertad. Por otro lado, cuando se tiene una variable Z con distribución normal con media cero y varianza $\sigma^2 = 1$, entonces lo siguiente es cierto:

$$Z^2 \sim \chi^2_1$$

Así, la suma de variables chi-cuadrado es una chi-cuadaro en los siguientes términos:

$$\sum_{i=1}^{m} Z^2 \sim \chi^2_m$$

Dada la distribución $\hat{\boldsymbol{\beta}}$, y que la varianza de cada uno de los elementos de $\hat{\boldsymbol{\beta}}$, digamos $\hat{\beta}_k$, es: $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}$. En este punto denotaremos con $\hat{\beta}_k$ al elemento k-ésimo dentro de $\hat{\boldsymbol{\beta}}$. Adicionalmente, con $(\mathbf{X}'\mathbf{X})^{-1}_{kk}$ al elemento en la fila k y la columna k en la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. Derivado de lo anterior, lo siguiente es cierto:

$$Z = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim \mathcal{N}(0,1)$$

Considerando la distribución de ε_i y de que e_i es un estimador de ε_i, podemos afirmar que:

$$Z = \frac{e_i - 0}{\sigma} = \frac{e_i}{\sigma} \sim \mathcal{N}(0,1)$$

De esta forma:

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} \sim \chi^2_n$$

Dada la distribución t de Student en la ecuación (3.32) y los elementos antes mencionados podemos llegar a las siguientes conclusiones:

$$\begin{aligned} t &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}}{\sqrt{\sum_{i=1}^{n} \frac{e_i^2}{\sigma^2}/(n-K)}} \\ &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}}{\sqrt{\sum_{i=1}^{n} \frac{e_i^2}{(n-K)\sigma^2}}} \\ &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \\ &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}}{\frac{\sqrt{\hat{\sigma}^2}}{\sqrt{\sigma^2}}} \\ &= \frac{(\hat{\beta}_k - \beta_k)\sqrt{\sigma^2}}{\sqrt{\hat{\sigma}^2}\sqrt{\sigma^2}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \\ &= \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \end{aligned}$$

De esta forma:

$$t = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{n-k} \tag{3.33}$$

Construida esta expresión (3.33) podemos establecer el siguiente intervalo de confianza para cada uno de nuestros estimadores $\hat{\beta}_k$:

$$-t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}} < \hat{\beta}_k - \beta_k < t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}$$

$$\hat{\beta}_k - t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}} < \beta_k < \hat{\beta}_k + t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}$$

Donde $t_{\alpha/2,n-K}$ es el valor crítico de tablas t-Student. Así, podemos enunciar que el estimador de MCO de la ecuación (3.8) se sujetará a la siguiente

prueba de hipotésis general:

$$\begin{aligned} H_0 : \beta_k &= r \\ H_a : \beta_k &\neq r \end{aligned}$$

Donde $k = 1, 2, \ldots, K$ y r es un valor respecto del cual se desea comparar a $\hat{\beta}_k$. Dicha prueba se distribuye como una t_{n-K} y su especificación más común en el análisis de regresión es la siguiente:

$$\begin{aligned} H_0 : \beta_k &= 0 \\ H_a : \beta_k &\neq 0 \end{aligned}$$

Lo que en términos de una prueba t es la siguiente:

$$t = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{n-k} \tag{3.34}$$

Dicha prueba la podemos interpretar como sí es posible que el estimador $\hat{\beta}_k$ sea igual a 0, o en térmninos económicos, si existe evidencia estadística de que el efecto de la variable x_k en y es nulo.

Como se mostró anteriormente, sólo se hemos hablado de pruebas de hipotésis cuándo se desea comprobar si alguno de los estimadores $\hat{\beta}_k$ de MCO es estadísticamente igual a algún valor determinado. No obstante, en ciertas condiciones puede ser interesante cuestionar si en conjunto todos los estimadores cumplen una cierta restricción. A este tipo de situaciones se les suele conocer como pruebas de hipótesis globales. Sin pérdida de generalidad podemos enunciar a una prueba global como:

$$\begin{aligned} H_0 : \mathbf{R}\boldsymbol{\beta} &= \mathbf{r} \\ H_a : \mathbf{R}\boldsymbol{\beta} &\neq \mathbf{r} \end{aligned}$$

Donde, $\boldsymbol{\beta}$ es el vector de parámetros del MCO de la ecuación (3.8), $\mathbf{R}$ y $\mathbf{r}$ son de la siguiente manera:

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & \ldots & R_{1K} \\ R_{21} & R_{22} & \ldots & R_{2K} \\ R_{31} & R_{32} & \ldots & R_{3K} \\ \vdots & \vdots & \ldots & \vdots \\ R_{J1} & R_{J2} & \ldots & R_{JK} \end{bmatrix}$$

En $\mathbf{R}$ cada una de las R_{jk} toma el valor de 1 o 0, según sea la estructura de la restricción. Con $j = 1, 2, \ldots, J$. En el caso de $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_J \end{bmatrix}$$

Donde cada uno de los r_j representa el valor de la restricción que es evaluada. No debe pasar desapercibido que en este caso existen J restricciones ($J \leq K$), mismas que se evaluan de forma conjunta en una sola prueba. Esa es la ventaja de la prueba que se enuncia a continuación.

Definamos a F de Fisher como la razón de dos pruebas t de Student, la primera para los valores de una combinación lineal de $\hat{\boldsymbol{\beta}}$, y la segunda para el estimador de la varianza $\hat{\sigma}^2$:

$$\begin{aligned} \mathbf{F} &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/J}{\hat{\sigma}^2} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}Var(\hat{\boldsymbol{\beta}})\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \end{aligned}$$

Por lo tanto la estadística F de prueba será:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \sim F_{J,n-K} \tag{3.35}$$

Así, la prueba de hipótesis que es la más común en el análisis de regresión y que se le conoce como prueba global. Dicha prueba asume una forma de la matriz $\mathbf{R}$ definida así:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Y un vector $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

De esta forma la hipotésis nula de una prueba global se puede escribir como:

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_2 = \ldots = \beta_K = 0 \\ H_a : &\quad NoH_0 \end{aligned}$$

Esta prueba se le conoce como prueba global, ya que prueba si en conjunto todas las variables independientes tienen un efecto nulo en $\mathbf{Y}$.

Pero ahora discutamos qué sucede cuando no se cumplen los supuestos. El primer caso es cuando no se cumple la ecuación (3.23). La respuesta es que en ese caso MCO no generará un estimador consistente. La solución a ese problema es la implementación del Método de Variables Instrumentales que discutiremos más adelante.

Por otro lado, en los casos en que existe un término de error heterocedástico, es decir, que no cumple con la ecuación (3.30), podemos corregir el problema mediante la matriz de White, la cual asume que la varianza asintórica de $\hat{\boldsymbol{\beta}}$ estará dada por:

$$\widehat{AVar(\hat{\boldsymbol{\beta}})} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \tag{3.36}$$

Donde $\mathbf{W}$ es una matriz diagonal de dimensión $N \times N$ y cada uno de los elementos de la diagonal resultan de $w_i = e_i^2$, considerando que: $e_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$. Esta matriz por la que se pondera la varianza en la ecuación (3.36) fue introducida en la econometría por White en 1980.

La heterocedasticidad modelada y corregida de esta manera se conoce como errores estándar robustos. Una vez que podemos determinar los errores estándar, tendremos la posibilidad de establecer pruebas como la t. No obstante, cuando el supuesto de la ecuación (3.28) no se cumple diremos que la prueba F usual será erronéa y no será valida para probar restricciones múltiples.

En este caso nos referimos a casos como:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = r$$

Donde $\mathbf{R}$ es una matriz de restricciones de dimensión $Q \times K$, con $Q \leq K$, r es un vector de dimensión $Q \times 1$. Así en casos de heterocedasticidad utilizaremos una prueba de Wald dada por:

$$\mathbf{W} = (\mathbf{R}\hat{\boldsymbol{\beta}} - r)'(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - r)$$

Donde $\hat{\mathbf{V}} = \widehat{AVar(\hat{\boldsymbol{\beta}})}$, por lo que la distribución de $\mathbf{W}$ será como:

$$\mathbf{W} \sim \chi^2_Q$$

Lo que implica que la formulación $F_{Q,N-K}$ no es de aplicación.

Otro tipo de pruebas que pueden realizarse son aquellas basadas en Multiplicadores de Lagrange. Supongamos un modelo lineal particionado de la siguiente manera:

$$y_i = \mathbf{x}_{i1}\boldsymbol{\beta}_1 + \mathbf{x}_{i2}\boldsymbol{\beta}_2 + \varepsilon \tag{3.37}$$

Donde se cumplen los supuestos dados en las ecuaciones (3.23), (3.24) y (3.28), y $\mathbf{x}_1$ y $\mathbf{x}_2$ son vectores de dimensión $1 \times K_1$ y $1 \times K_2$, respectivamente.

Supongamos que desemos probar la hipótesis de que:

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

Sea $\widetilde{\boldsymbol{\beta}}_1$ esl estimador de $\boldsymbol{\beta}_1$ bajo la hipótesis nula anterior, digamos que este estimador corresponde al modelo restringido. Definamos los residuales:

$$e_i = y_i - \mathbf{x}_{i1}\widetilde{\boldsymbol{\beta}}_1$$

Bajo la hiótesis nula, $\mathbf{x}_{i2}$ debería ser una varible no correlacionada con e_i. En este caso la prueba estadística es obtenida como sigue:

1. Corremos una regresión de e_i en $\mathbf{x}_{i1}$ y $\mathbf{x}_{i2}$

2. Asumiendo que $\mathbf{x}_{i1}$ contiene término constante, digamos que R^2_E denota al R^2 asociado a la regresión.

3. Entonces propongamos que la estadística de Multiplicadores de Lagrange es:

   $$LM = N \cdot R^2_E$$

   Para llgar a la expresión anterior se requiere de solucionar un problema de optimización.

4. Bajo la H_0, tenemos que:

$$LM \sim \chi^2_{K_2}$$

Donde K_2 es el número de restricciones que hemos probado. De esta forma, si $N \cdot R^2_E$ es suficientemente grande, entonces e_i estaba significativamente correlacionado con $\mathbf{x}_{i2}$, por lo que la hipótesis debió ser rechazada.

En este punto debe quedar claro que R^2 requiere que $\sum_{i=1}^{N} e_i = 0$, lo que implica que el modelo tiene un término constante.

Veamos un ejemplo ilustrativo. Sea la siguiente ecuación:

$$\begin{aligned} ln(Salario) &= \beta_1 + \beta_2 Experiencia + \beta_3 Experiencia^2 \\ &\quad + \beta_4 Educacion + \beta_5 Edad + \beta_6 NumHijos + \varepsilon \end{aligned}$$

Supongamos que la hipótesis que queremos probar es:

$$H_0 : \beta_4 = \beta_5 = 0$$

3.6. Ejercicios

Ahora proponemos los siguientes ejercicios:

1. Utilice el material hasta ahora expuesto en el curos para analizar y derterminar si las siguientes afirmaciones son FALSAS o VERDADERAS.

   a) Cuando se realiza un proceso de estimación por el método de MCO, la suma de los erroes resultantes siempre es igual a cero, es decir,

   $$\sum_{i=1}^{n} e_i = 0$$

   b) Suponga que realiza una estimación por el método de mínimos cuadrados de la ecuación dada por $y_i = \beta_2 x_{i2} + \beta_3 x_{i3} + ... + + \beta_K x_{iK} + \varepsilon_i$, para $i = 1, 2, \ldots, N$. Asimismo, suponga que como resultado de su estimación encunetra que la estadística R^2 es muy próxima al valor de 0.5, entonces ¿se puede decir que las variables dadas por $x_{i2}, x_{i3}, ..., x_{iK}$ explican aproximadamente la mitad de la variabilidad de la variable y_i?

c) Suponga que estima una ecuación de salario dada por:

$$salario_i = \beta_1 + \beta_2 educa_i + \beta_3 educa_i^2 + \beta_4 DSexo_i + \varepsilon_i$$

Donde todas los estimadores $\hat{\beta}_k$, $k = 1, 2, 3, 4$ son estadísticamente significativos; $i = 1, 2, \ldots, N$; $salario_i$ es el ingresos mesual por sueldos y salarios de una persona; $educa_i$ es la educación de una persona en años, y $DSexo_i$ es una variable dummy que toma el valor de 1 si el individuo es hombre y 0 en cualquier otro caso. Asuma que los resultados indican que: $\hat{\beta}_2 > 0$ y $\hat{\beta}_3 > 0$. Entonces, ¿podemos afirmar qué existe evidencia estadística de la presencia de redimientos marginales decrecientes de la educación en los salarios?

d) Asuma que estima una regresión de los salarios de las personas en una cierta industria y encuenta que sus resultados son:

$$\begin{array}{rcccc} \hat{y}_i & = & 5.15 & + & 1.17\ x_i \\ & & (2.5) & & (0.3) \end{array}$$

Donde y_i es el salario mensual en miles de pesos de las personas, y x_i es una variable dummy que toma el valor de 1 cuando el individuo i es hombre y 0 cuando es mujer. Entre paréntesis se indica los errores estándar de los estimadores (conocidos como término respectivo a $\hat{\sigma^2}(\mathbf{X'X})^{-1}_{kk}$).

Entonces, existe evidencia estadística y significativa para afirmar que la mejor estimación al salario de los hombres es 5.15 mil pesos y para las mujeres es de 6.31 mil pesos.

2. Suponga que ha decidido estimar una regresión del logarittmo natural del salario en función de un conjunto de variables para un conjunto de individuos en una industria; de tal forma que obtiene los resultados mostrados en el Cuadro 3.1.

   En donde la variables son como se describe a continuación:

   a) Masculino: es una variable binaria que toma el valor de 1 (uno) cuando el individuo es del sexo masculino y 0 (cero) en cualquier otro caso.

Cuadro 3.1: Resultados de la regresión de salarios

Variable	Coeficiente	Error Estándar	Estadística t
Constante	4.216	0.078	
Masculino	0.154	0.005	
Primaria	0.224	0.018	
Secundaria	0.833	0.063	
Preparatoria	0.902	0.063	
Univesidad	0.550	0.065	8.462
ln(Experiencia)	0.907		22.535
Primaria×Masculino	-0.97		-1.242
Secundaria×Masculino	-0.67		-2-272
Preparatoria×Masculino	-0.72		-2.317
Univesidad×Masculino	-0.146	0.076	
ln(Experiencia)×Masculino	0.041		4.891
$R^2 = 0.6032$	F = 89.69	n = 1,000	$\mathbf{e}'\mathbf{e} = 415.37$

b) Primaria, Secundaria, Preparatoria y Universidad: son un conjunto de variables categóricas (Dummy) que indican el nivel máximo de estudios del individo, es decir, toman el valor de 1 si el individuo cumple con la categoria y cero si su nivel máximo de estudios es algún otro novel. Por lo tanto hemos dejado fuera la dummy que indica nula educación.

c) Experiencia: es una variable que indica el número de años de experiencia del individuo.

Se le solicita:

a) Completar el Cuadro 3.1, así como indicar la hipótesis nula y alternativa asociada. También deberá determinar cuales de las variables incluidas en la regresión resultan estadísticamente significativas.

b) Utilizando la información de Cuadro 3.1 determine la hipótesis nula y alternativa para saber si el logarítmo natural de la experiencia es estadísticamente igual a 1. realice la prueba t asociada y concluya si se puede rechazar la hipótesis nula.

c) Utilizando la información de Cuadro 3.1 determine y explique cuál es el Efecto Marginal de la variable Universidad.

3. Cómo se modifica el resultado expresado en la ecuación (3.15) cuando el R^2 no es respecto de la media $\bar{y}$, es decir, cuando la variación total es como se muestra en la ecuación (3.38):

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - 0)^2 \tag{3.38}$$

4. Un embotellador de bebidas gaseosas analiza las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Le interesa predecir el tiempo necesario para que el representante de ruta atienda las máquinas expendedoras en una tienda. Esta actividad de servicio consiste en abastecer la máquina con productos embotellados, y algo de mantenimiento o limpieza. El ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan el tiempo de entrega, y, son la cantidad de cajas de producto abastecido, x_1, y la distancia caminada por el representante, x_2. El ingeniero ha reunido una muestra aleatoria de algunas observaciones de tiempo de entrega.

   Suponga que desea estimar la ecuación de regresión asumiendo una media condicional dada por:

$$\mu_{y_i|x_{i1},x_{i2},x_{i3}} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} \tag{3.39}$$

   Derivado de la operación matricial de la muestra suponga que obtiene los siguientes resultados:

$$\begin{bmatrix} \sum y_i^2 & \sum y_i x_{i1} & \sum y_i x_{i2} & \sum y_i x_{i3} \\ \sum x_{i1} y_i & \sum x_{i1}^2 & \sum x_{i1} x_{i2} & \sum x_{i1} x_{i3} \\ \sum x_{i2} y_i & \sum x_{i2} x_{i1} & \sum x_{i2}^2 & \sum x_{i2} x_{i3} \\ \sum x_{i3} y_i & \sum x_{i3} x_{i1} & \sum x_{i3} x_{i2} & \sum x_{i3}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 18310.63 & 559.60 & 7375.44 & 324972.70 \\ 559.60 & 25.00 & 219.00 & 9932.00 \\ 7375.44 & 219.00 & 3055.00 & 129099.00 \\ 324972.69 & 9932.00 & 129099.00 & 6402888.00 \end{bmatrix} \tag{3.40}$$

Al respecto:

a) Determine los coeficientes $\boldsymbol{\beta}$ de regresión lineal

b) Determine el valor de la varianza estimada de los errores: $\hat{\sigma}^2$

c) Pruebe la hipótesis nula de que $\beta_2 = 0$ con un $\alpha = 0.05$

d) Pruebe la hipótesis nula de que $\beta_3 = 0$ con un $\alpha = 0.05$

e) Escriba una conclusión de las dos pruebas de hipótesis planteadas

f) Determine el valor del R^2 asociado al caso, para ello recuerde que el procedimiento empleado para determinar la ecuación (3.15)

4

Estimación de modelos lineales de una sola ecuación por el método de Variables Instrumentales

4.1. Introducción y motivación

El método de Variables Instrumentales (VI o IV, por sus siglas en inglés) es el segundo método después de MCO más utilizado en el análisis de regresión. Principalmente, la implementación de IV bajo un procedimiento conocido como Mínimos Cuadrados Ordinarios en 2 Etapas (en inglés Two-Stage Least Squares (2SLS)). Para iniciar partamos de un modelo lineal:

$$y = \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \tag{4.1}$$

Donde asumimos que $\mathbb{E}[\varepsilon] = 0$ y $Cov[x_j, \varepsilon] = 0$, para $j = 1, 2, \ldots, K-1$. Así, asumiremos que la variable x_K es la única que está correlacionada con el término de error de forma que $Cov[x_K, \varepsilon] \neq 0$. En otras palabras, diremos que sólo las variables $x_1, x_2, \ldots, x_{K-1}$ son exógenas y que potencialmente x_K es una variable endógena en la ecuación (4.1).

El método de IV provee una solución general al problema de endogenidad de una o múltiples variables explicativas. Para utilizar el procedimiento necesitamos encontrar una variable observable z, que no éste incluida en la ecuación (4.1) y que además cumpla con las siguientes dos condiciones:

1. z no debe estar correlacionada con ε, es decir:

$$Cov[z, \varepsilon] \neq 0$$

   Es decir, que el siguiente vector de variables explicativas son exógenas: $\{x_1, x_2, \ldots, x_{K-1}, z\}$.

2. La proyección lineal de x_K en todas las variables exógenas estará dada por:

$$x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta z + \xi_K$$

   Donde $\mathbb{E}[\xi_K] = 0$, y ξ_K no está correlacionado con $\{x_1, x_2, \ldots, x_{K-1}, z\}$ y $\theta \neq 0$

Si $Cov(z, \varepsilon) = 0$ y $x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta z + \xi_K$, entonces z es una variable instrumental para x_K. A la proyección $x_K = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta z + \xi_K$ la denominaremos como la forma reducida. Si introducimos la forma reducida en la ecuación (4.1) podríamos plantear una expresión como:

$$y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_{K-1} x_{K-1} + \lambda z + \nu$$

Donde $\nu = \varepsilon + \beta_K \xi_K$ es la forma reducida del término de error, $\alpha_j = \beta_j + \beta_K \delta_j$, $\alpha_0 = \beta_K \delta_0$ y $\lambda = \beta_K \theta$.

4.2. Estimación por Mínimos Cuadrados en Dos Etapas – 2SLS

Ahora plateamos la situación al problema de estimación. Para verlo, partamos:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

Asumamos:

$$\begin{aligned} \mathbf{x}_i &= (1, x_{i2}, \ldots, x_{iK}) \\ \mathbf{z}_i &= (1, x_{i2}, \ldots, x_{iK-1}, z) \end{aligned}$$

Donde $\mathbf{z}_i$ tiene todas las variables exógenas, por lo que cumple con:

$$\mathbb{E}[\mathbf{z}_i' \varepsilon] = \mathbf{0}$$

De esta forma, podríamos multiplicar la ecuación (4.2) por $\mathbf{z}_i$ y tomamos valor esperado, obtendríamos:

$$\begin{aligned} \mathbf{z}_i' y_i &= \mathbf{z}_i' \mathbf{x}_i \boldsymbol{\beta} + \mathbf{z}_i' \varepsilon_i \\ \mathbb{E}[\mathbf{z}' y] &= \mathbb{E}[\mathbf{z}' \mathbf{x}] \boldsymbol{\beta} + \mathbb{E}[\mathbf{z}' \varepsilon] \end{aligned}$$

Donde $\mathbf{z}'\mathbf{x}$ es de rango $K \times K$ y $\mathbf{z}'y$ es de $K \times 1$. Por lo tanto, la solución poblacional estará dada por:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbb{E}[\mathbf{z}'\mathbf{x}]\right)^{-1} \mathbb{E}[\mathbf{z}' y] \\ &= \left(N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' y_i \right) \\ &= (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{Y} \end{aligned}$$

Por lo tanto el estimador de Variables Instrumentales estará dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{Y} \tag{4.2}$$

Donde $\mathbf{Z}$ y $\mathbf{X}$, ambas, son dimensión $N \times K$. Ahora veamos un ejemplo para el caso de estimación por VI en el caso de variables omitidas puesto que la variable es no observable.

Supongamos una ecuación del logaritmo de los salarios para una muestra dada, la cual estima de la forma:

$$ln(Salario) = \beta_1 + \beta_2 Experiencia + \beta_3 Experiencia^2 + \beta_4 Educacion + \varepsilon$$

Donde ε está correlacionado con la variable de *Educacion*, ya que en la estimación hemos omitido las variables relacionadas con la habilidad, así como otros factores, tales como calidad de la educación y los antecedentes basados en condiciones familiares.

Suponngamos que ha sido posible identntificar la educación de la madre, la cual no está correlacionada con el tñermino ε, pero si lo está con los factores omitidos.

4.3. Método de Variables Instrumentales en casos con múltiples instrumentos

Consideremos un modelo como el de la ecuación (4.1), donnde x_K puede estar correlacionada con el ε. Asumamos que tenemos más de una variable

instrumental para x_K, las cuales:

$$\{z_1, z_2, \ldots, z_M\}$$

Cada una de las variables cumple con $Cov(z_h, \varepsilon) = 0$, $h = 1, 2, \ldots, M$. Así, cada una de las variables z_h es exógena, por lo que caulquier combinación lineal de:

$$\{x_1, x_2, \ldots, x_{K-1}, z_1, z_2, \ldots, z_M\}$$

será exógena o no estará correlacionada con ε. Digamos que el vector que incluye las varialbes anteriores será definido como por unoo de dimensión $1 \times L$, donde $L = K + M$:

$$\mathbf{z} = (1, x_1, x_2, \ldots, x_{K-1}, z_1, z_2, \ldots, z_M)$$

Entonces, podemos escribir la forma reducida como:

$$x_k = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \ldots + \theta_M z_M + \xi_K$$

Dado que hemos asumido que ξ_K no está correlacionado con el resto de los regresores, entonces podemos establecer que:

$$x_K^* = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \ldots + \theta_M z_M$$

es la parte de x_K que está correlacionda con ε.

Partiendo de lo anterior podemos escribir una ecuación para $\hat{x}_K$ de la forma:

$$\hat{x}_{iK} = \hat{\delta}_0 + \hat{\delta}_1 x_{i1} + \ldots + \hat{\delta}_{iK-1} x_{K-1} + \hat{\theta}_1 z_{i1} + \ldots + \hat{\theta}_M z_{iM}$$

Entonces podemos definir al vector:

$$\hat{\mathbf{x}}_K = (x_{i1}, \ldots, x_{iK-1}, \hat{x}_{iK})$$

Usando lo anterior podemos utilizar el estimado de IV para determinar:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_K \mathbf{x}_K \right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_K y \right) \\ &= (\hat{\mathbf{X}}' \mathbf{X})^{-1} \hat{\mathbf{X}}' \mathbf{Y} \end{aligned} \tag{4.3}$$

Por otro lado, podemos establecer el siguiente vector:

$$\mathbf{z}_i = (1, x_{i1}, \ldots, x_{iK-1}, z_{i1}, \ldots, z_{iM})$$

Contruyendo de forma simimar a otras matrices a $\mathbf{Z}$ apilado la información de cada uno de los individuos. De esta forma podremos constriuir $\hat{\mathbf{X}}$ mediante el uso de un estimador de MCO:

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{Z}\hat{\boldsymbol{\gamma}} \\ &= \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \end{aligned}$$

De lo anterior tendríamos que:

$$\hat{\mathbf{X}}' = \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

Usando las expresiones anteriores en la ecuación (4.3), obtenemos que:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{Y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Y} \end{aligned} \tag{4.4}$$

De forma similar al caso de MCO proponemos el siguiente estimador de la varianza del término de error:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \frac{e_i^2}{N-K} \tag{4.5}$$

Utilizando la ecuación (4.4) establecemos que la varianza de $\hat{\boldsymbol{\beta}}$ estará dada por:

$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

De esta forma todas las pruebas de hipótesis planteadas para MCO son aplicables.

Finalmente, una prueba que se requiere es la determinar si los intrumentos no están correlacionados con el término de error, por lo que se requiere de estimar:

$$e_i = \delta_0 + \delta_1 x_1 + \ldots + \delta_{K-1} + \theta z_1 + \ldots + \theta_M z_M$$

De esta expresión requerimos mostrar que todos los parámetros son no significativos ya sea de forma individual o en conjunto. En este punto es posible incorporar una prueba de LM como la explicada en la sección anterior.

4.4. Ejercicios

En esta sección platearemos un caso típico del método de VI. Para lo cual le pedimos seguir la discusión y realizar las actividades que se le piden. Consideremos un modelo de equilibrio de mercado (la discusión original se encuentra en el trabajo de *Working (1927). What do statistical demand curves show?. Quaterly Journal of Economics, 41 pp. 212 - 235*). Supongamos que tenemos el siguiente modelo de demanda o modelo en su forma estructural:

$$\begin{aligned} q^d &= \alpha_0 + \alpha_1 p_i + u_i \\ q^s &= \beta_0 + \beta_1 p_i + v_i \end{aligned}$$

Donde q^d es la cantidad demandada para el bien en cuestión, q^s es la cantidad ofertada, p_i es el precio de dicho bien. Los términos de error u_i y v_i representan otros factores observables que afectan o desplazan a la cantidad ofertada y a la cantidad demandada. Podemos asumir que dichos elementos u_i y v_i son factores no observables. De esta forma $\mathbb{E}[u_i] = 0$ y que $\mathbb{E}[v_i] = 0$. Adicionalment, suponga que que $Cov[u_i, v_i] = 0$ y que $q^d = q^s = q$, de tal forma que el mercado está en equilibrio, por lo que tenemos:

$$q = \alpha_0 + \alpha_1 p_i + u_i \tag{4.6}$$

$$q = \beta_0 + \beta_1 p_i + v_i \tag{4.7}$$

Ambas ecuaciones tienen problemas de endogeneidad, ya que el precio p_i es endógeno en cada una de las ecuaciones, ya que se determina simúltaneamente.

Utilizando las ecuaciones (4.7) y (4.7) obtenga la ecuación de demanda:

$$q_i = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1} \tag{4.8}$$

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1} \tag{4.9}$$

De esta forma hemos obtendrá dos expresiones (4.8), ecuación de demanda, y (4.9), ecuación de oferta, que representan la solución para el sistema de ecuaciones. No obstante, se puede apreciar que cada una de las variables tiene influencia del término de error de cada una de las ecuaciones, por lo que NO se pueden separar enteramente y se sule decir que tenemos un problema de endogeneidad.

A partir de la ecuación (4.9), determine la covarianza entre p_i y cada uno de los factores de cambio de la demanda (u_i) y de la oferta (v_i). Muestre que es igual a:

$$Cov(p_i, u_i) = -\frac{Var(u_i)}{\alpha_1 - \beta_1} \tag{4.10}$$

$$Cov(p_i, v_i) = \frac{Var(v_i)}{\alpha_1 - \beta_1} \tag{4.11}$$

De esta forma, el precio está correlacionado positivamente con los factores que modifican la demanda y negativamente con los facores que modifican la oferta, ello en razón de que la pendiente de la demanda es negativa ($\alpha_1 < 0$) y la de la oferta es positiva ($\beta_1 > 0$). En este caso, decimos que el problema de endogenidad se origina por el proceso de equilibrio en el mercado.

La razón de la endogeniedad referida es que de la curva de demanda o de la curva de oferta no es posible inferir si los cambios en el nivel de precios o en las cantidades en el mercado son por causa de desplazamientos de la curva de oferta o de la curva de demanda. Esto significa que sólo es posible estimar la curva de oferta o de demanda si es posible observar alguno de los factores que determina a la demanda o la oferta.

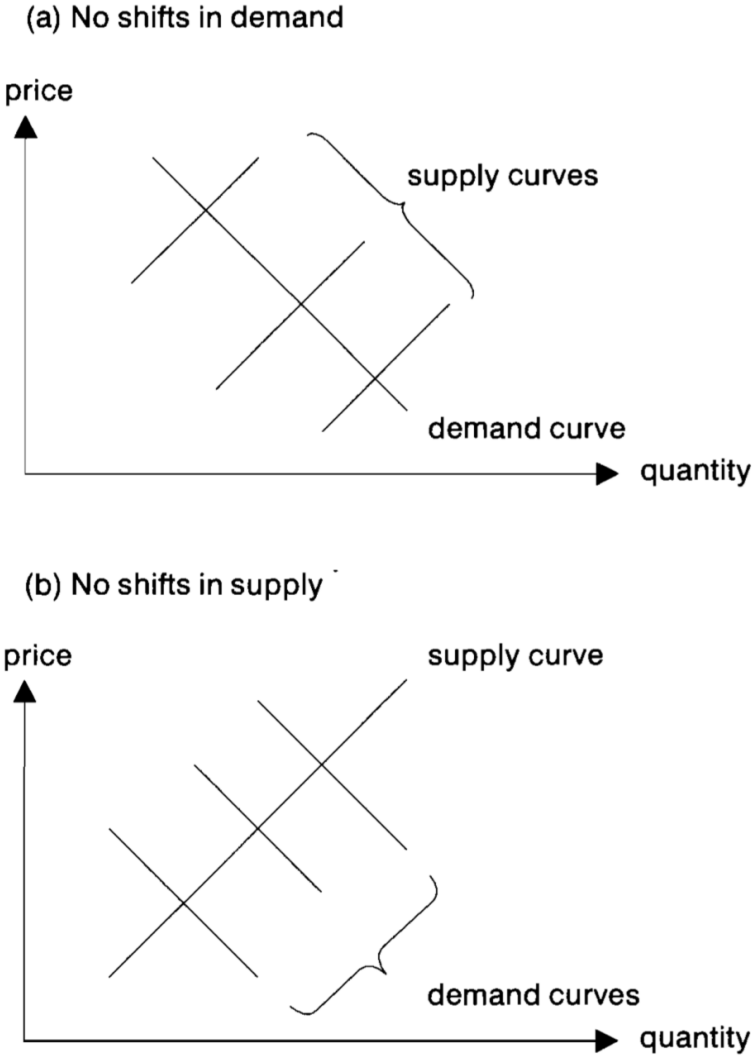
Supongamos que el efecto del termino de error en la cantidad ofertada, v_i, puede ser dividido en: los fatores observables que están asociados a la demanda x_i y un factor no observable ζ_i que no está correlacionado con x_i. Es decir, que tenemos una formulación de la ecuación de oferta similar a:

$$q_i = \beta_0 + \beta_1 p_i + \beta_2 x_i + \zeta_i \tag{4.12}$$

De esta forma podemos decir que existe una posibilidad de que se elimine la endogeidad, esto es, que se pueda instrumentar mediante el uso de variables asociadas a la demanda (x_i). Lo anterio, para efecto de separar del término de error la información que puede ser endogena al sistema, tal y como se muestra en la ecuación (4.12). La Figura 4.1 ilustra está discusión para cada uno de los casos: de la demanda y de la oferta.

Esta discusión es una aplicación del Método de Variables Instrumentales (IV). Para el cual requerimos un regresor que es endogeno y que requerimos sea instrumentado. Al instrumento se le denomina variable instrumental. En este caso, es la variable x_i quién sera nuestra variable instrumental. Retomemos nuestro sistema de ecuaciones de demanda y oferta. Si realizamos el

Figura 4.1: Instrumentación de la información de la demanda y la oferta, respectivamente



mismo procedimiento de solución del sistema de ecuaciones pero considerando a la ecuación (4.12).

Dicho lo anterior, muestre que obtendríamos un sistema que quedará como sigue:

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} + \frac{\zeta_i - u_i}{\alpha_1 - \beta_1} \tag{4.13}$$

$$q_i = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1} + \frac{\alpha_1\zeta_i - \beta_1 u_i}{\alpha_1 - \beta_1} \tag{4.14}$$

En este punto, si fue posible instrumentar correctamente la variable endogena, muestre que deberá pasar que $Cov(x_i, \zeta_i) = 0$. Es decir, es esperado que no exista información endogena. Así aseguraríamos que $Cov(x_i, u_i) = 0$.

5

Estimación de sistemas de ecuaciones por el por el método de Mínimos Cuadrados Ordinarios y Mínimos Cuadrados Generalizados

5.1. Introducción a Sistemas de Ecuaciones

En esta sección analizaremos una extensión de la estimación por MCO de una sola ecuación al caso de múltiples ecuaciones para el caso de los Sistemas Ecuaciones Aparentemente No Relacionnados (SUR). Adicionalmente, si el tiempo nos lo permite, incorporaremos el análisis de Mínimos Cuadrados Generalizados Factibles, en el cual propondremos una solución por el caso en el que existe heterocedásticidad, y para sistemas de ecuaciones con problemas de endogenidad. Algunas de las carácteristicas de está metodología aplican al caso de los modelos de datos panel, en estricto sentido los modelos de datos panel son un caso particular de los sistemas de ecuaciones, como lo veremos más adelante.

Supongamos un conjunto de G ecuaciones descritas como:

$$\begin{aligned} y_1 &= \mathbf{x}_1\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2 &= \mathbf{x}_2\boldsymbol{\beta}_2 + \varepsilon_2 \\ &\vdots \\ y_G &= \mathbf{x}_G\boldsymbol{\beta}_G + \varepsilon_G \end{aligned} \tag{5.1}$$

Donde $\mathbf{x}_g$ es un vector de dimensión $1 \times K_g$, y $\boldsymbol{\beta}_g$ es un vector de dimensión $K_g \times 1$, para $g = 1, 2, \ldots G$.

En muchos casos $\mathbf{x}_g$ podría ser el mismo para cada una de las G ecuaciones y, en consecuencia, $\boldsymbol{\beta}_g$ necesariamente tiene las misma dimensión. No obstante, el modelo debe permitir que los elementos en $\mathbf{x}_g$ cambien a lo largo de las ecuaciones.

Pongamos como ejemeplo un caso en el que queremos estimar funciones de demanda de las familias de un país, por ejemplo:

$$\begin{aligned} Vivienda &= \beta_{11} + \beta_{12}PrecioCasas + \beta_{13}PrecioAlimentos \\ &\quad + \beta_{14}PrecioRopa + \beta_{15}Ingreso + \beta_{16}IntHogar + \varepsilon \\ Alimentos &= \beta_{11} + \beta_{12}PrecioCasas + \beta_{13}PrecioAlimentos \\ &\quad + \beta_{14}PrecioRopa + \beta_{15}Ingreso + \beta_{16}IntHogar + \varepsilon \\ Vestido &= \beta_{11} + \beta_{12}PrecioCasas + \beta_{13}PrecioAlimentos \\ &\quad + \beta_{14}PrecioRopa + \beta_{15}Ingreso + \beta_{16}IntHogar + \varepsilon \end{aligned}$$

En general hemmos omitido el subíndice $i = 1, 2, \ldots, N$, puesto que la forma adecuada de cada ecuación sería:

$$y_{ig} = \mathbf{x}_{ig} + \varepsilon_{ig}$$

En los casos como el modelo anterior en la ecuación (5.1) asumimos que no existe endogeneidad (por razones como omisión de variables, errores estándar o simúltaneidad), en particular hemos asumido que existe exogenidad estricta:

$$\mathbb{E}[\varepsilon_g|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g] = 0 \text{ para } g = 1, 2, \ldots, G$$

De forma general podemos decir que:

$$\mathbb{E}[y_g|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g] = \mathbb{E}[y_g|\mathbf{x}] = \mathbf{x}_g\boldsymbol{\beta}_g, \text{ para } g = 1, 2, \ldots, G$$

5.2. Estimación de Sistemas de Ecuaciones Multivariantes

Ahora generalicemos el problema. Asumamos que tenemos observaciones de corte transversal independienntes e idénticamente distribuidas:

$$\{\mathbf{X}_i, \mathbf{Y}_i; i = 1, 2, \ldots, N\}$$

Donde $\mathbf{X}_i$ es una matriz $G \times K$ y $\mathbf{Y}_i$ es un vector $G \times 1$. Así, la matriz $\mathbf{X}_i$ continene todas las variables explicativas que aparecen en el sistema, para cada uno de los individuos indexados i:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{5.2}$$

Donde $\boldsymbol{\beta}$ es un vector $K \times 1$ y $\boldsymbol{\varepsilon}_i$ es un vector $G \times 1$. De esta forma en general tendríamos:

$$\mathbf{Y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iG} \end{bmatrix}$$

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iG} \end{bmatrix}$$

Donde $i = 1, 2, \ldots, N$.

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_G \end{bmatrix}$$

Donde:

$$\boldsymbol{\beta}_g = \begin{bmatrix} \boldsymbol{\beta}_{1g} \\ \boldsymbol{\beta}_{2g} \\ \vdots \\ \boldsymbol{\beta}_{K_g g} \end{bmatrix}$$

Y $g = 1, 2, \ldots, G$. por lo que $\boldsymbol{\beta}$ es de dimensión $(K_1 + K_2 + \ldots + K_G) \times 1$, considerando que $K = K_1 + K_2 + \ldots + K_G$.

Par el caso particular de un modelo de ecuaciones simúltneas aparentemente no relacionadas que no compartan variables explicativas:

$$\mathbf{x}_{ig} = [x_{i1}, x_{i2}, \ldots, x_{iK_g}]$$

Así, podemos proponer que la matriz $\mathbf{X}_i$ con dimensión $G \times (K_1 + K_2 + \ldots + K_G)$es de la forma:

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{i2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{x}_{iG} \end{bmatrix}$$

Dado el modelo de la ecuación (5.2) establecemos las condiciones de ortogonalidad para $\boldsymbol{\beta}$ como el supuesto:

$$\mathbb{E}[\mathbf{X}_i' \boldsymbol{\varepsilon}_i] = \mathbf{0} \tag{5.3}$$

Dado este supuesto obtenemos que:

$$\begin{aligned} \mathbb{E}[\boldsymbol{\varepsilon}_i] &= \mathbf{0} \\ \mathbb{E}[\varepsilon_{ig}|\mathbf{X}_i] &= 0 \text{ para } g = 1, 2, \ldots, G \end{aligned}$$

Bajo este supuesto de ortogonalidad entonces:

$$\begin{aligned} \mathbb{E}[\mathbf{X}_i' \boldsymbol{\varepsilon}_i] &= \mathbb{E}[\mathbf{X}_i'(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})] \\ \mathbb{E}[\mathbf{X}_i' \mathbf{Y}_i] &= \mathbb{E}[\mathbf{X}_i' \mathbf{X}_i] \boldsymbol{\beta} \end{aligned}$$

Donde la expresión anterior implica que $\mathbf{X}_i' \mathbf{Y}_i$ es de dimensión $X \times 1$ y $\mathbf{X}_i' \mathbf{X}_i$ es de dimensión $K \times K$ y no es una matriz singular.

De esta forma, el estimador $\boldsymbol{\beta}$ de sistemas de ecuaciones será:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbb{E}[\mathbf{X}_i' \mathbf{X}_i]^{-1} \mathbb{E}[\mathbf{X}_i' \mathbf{Y}_i] \\ &= \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Y}_i \right) \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \end{aligned}$$

Donde $\mathbf{X}$ es una matriz $NG \times K$ y $\mathbf{Y}$ es un vector de $NG \times 1$. En este punto podemos enunciar el siguiente teorema:

Teorema 5.1 ***Consistencia del Sistema OLS****. Bajo los supuestos anteriores* $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$

De forma similar al caso de MCO requerimos establecer una distribución para $\hat{\boldsymbol{\beta}}$, partamos de:

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \sqrt{N}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \boldsymbol{\beta}) \\
&= \sqrt{N}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}) \\
&= \sqrt{N}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}_i \\
&= \left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon}_i\right)
\end{aligned}$$

De esta forma, por el Teorema del Límite Central:

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\varepsilon}_i \xrightarrow{d} N(0, \Sigma)$$

Donde $\Sigma = \mathbb{E}[\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}] = Var\left(\mathbf{X}'\boldsymbol{\varepsilon}\right)$. Así, si el modelo es homocedástico:

$$\Sigma = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{5.4}$$

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{5.5}$$

Donde:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{N - K}$$

Si el modelo tiene heterocedásticidad la varianza asintótica de $\hat{\boldsymbol{\beta}}$ estará dada por:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

Finalmente, podemos establecer las mismas pruebas de hipótesis que en el caso de MCO.

5.3. Mínimos Cuadrados Generalizados

Como vimos en la sección anterior, podemos generalizar un modelo de regresión lineal commola ecuación:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Donde el número de elementos o individuos en la muestra es N (segmentada de la forma que sea, en el caso de un sistema de ecuaciones) y el número de variables en $\mathbf{X}$ es K (segmenta de la forma que sea). Asimismo, del cual asumimos que se cumplen:

$$\begin{aligned} \mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] &= \mathbf{0} \\ \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] &= \Sigma = \sigma^2\Omega \end{aligned}$$

En general, Σ puede contener elementos que nos indiquen un problema de heterodedásticidad y autocorrelación. Sin embargo, para nuestro caso utilizaremos como supuesto que el modelo sólo tienen heterocedásticidad, de esta forma enfrentaremos un caso en el que la matriz estará dada por:

$$\begin{aligned} \Sigma &= \sigma^2\Omega \\ &= \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_n \end{bmatrix} \end{aligned}$$

No debe pasar desapercibido que en el caso de que en el modelo exista autocorrelación es posible que enfrentemos una matriz como:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_1 & \dots & \rho_{n-1} \\ \rho_1 & \sigma_2^2 & \dots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \dots & \sigma_n^2 \end{bmatrix}$$

En estos casos es cuando necesitamos un procedimiento del tipo de Mínimos Cuadrados Generalizados. El método consiste en un algorimos que asume que:

$$\begin{aligned} \Sigma^{-1} &= \mathbf{P}'\mathbf{P} \\ &= \Sigma' \\ \Sigma &= \mathbf{P}\mathbf{P}' \end{aligned}$$

Si lo anterior es posible, entonces podemos establecer un nuevo modelo lineal como:

$$\begin{aligned} \mathbf{PY} &= \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \\ \mathbf{Y}_* &= \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_* \end{aligned} \tag{5.6}$$

Así, los supuestos necesarios para resolver el problema de regresión en la ecuación (5.6) son:

$$\begin{aligned} \mathbb{E}[\boldsymbol{\varepsilon}_*|\mathbf{X}_*] &= \mathbf{0} \\ \mathbb{E}[\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*'|\mathbf{X}_*] &= \sigma^2\mathbb{I}_n \end{aligned}$$

La explicación para este último resultado es que:

$$\begin{aligned} \mathbb{E}[\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*'|\mathbf{X}_*] &= \mathbb{E}[\mathbf{P}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{P}'|\mathbf{X}] \\ &= \mathbf{P}\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{P}' \\ &= \mathbf{P}\Sigma\mathbf{P}' \\ &= \sigma^2\mathbb{I}_n \end{aligned}$$

Bajo esta condición, veríamos que el estimador de Mínimos Cuadrados estará dado por:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{Y}_* \\ &= (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{Y} \\ &= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y} \end{aligned}$$

Definamos al estimador de Mínimos Cuadros Generalizados Factibles al siguiente:

$$\hat{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{Y}$$

Entonces el procedimiento consiste en determinar o estimar a $\hat{\Sigma}$, para lo cual debemos seguir el siguiente algoritmo:

1. Asumamos una forma funcional de σ_i^2:

$$\sigma_i^2 = \sigma^2 f(\mathbf{z}_i)$$

   Donde $\mathbf{z}_i$ es un conjunto de variables que pueden explicar la heterocedásticidad.

2. Estimamos $\hat{\sigma}_i^2$, asumiendo que $\sigma_i^2 = e_i^2$, que provienen de una regresión de mínimos cuadrados ordinarios:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

3. Determinamos cada elemento de la diagonal de la matriz $\hat{\Sigma}$ será como se describe en el inciso 2.

4. Estimamos:

$$\hat{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{Y}$$

En este punto la única incognita es cómo determinar si el modelo tiene heterocedásticidad. Para lo cual plateamos una hipótesis dada por:

$$\begin{aligned} H_0 &: \sigma_i^2 = \sigma^2 \ , \forall i \\ H_1 &: \text{No } H_0 \end{aligned}$$

Para probar la hipótesis anterior vamos a utilizar una prueba de Breusch-Pagan / Godfrey que está basada en una prueba LM. Esta prueba asume que $\sigma_i^2 = \sigma^2 f(\mathbf{z}_i)$ y propone como estadñsitica a:

$$LM = 1/2 \cdot \begin{bmatrix} \frac{e_1^2}{\frac{\mathbf{e}'\mathbf{e}}{n}} - 1 \\ \frac{e_2^2}{\frac{\mathbf{e}'\mathbf{e}}{n}} - 1 \\ \vdots \\ \frac{e_n^2}{\frac{\mathbf{e}'\mathbf{e}}{n}} - 1 \end{bmatrix}' \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \begin{bmatrix} \frac{e_1^2}{\frac{\mathbf{e}'\mathbf{e}}{n}} - 1 \\ \frac{e_2^2}{\frac{\mathbf{e}'\mathbf{e}}{n}} - 1 \\ \vdots \\ \frac{e_n^2}{\frac{\mathbf{e}'\mathbf{e}}{n}} - 1 \end{bmatrix} \tag{5.7}$$

Donde la ecuación (5.7) se distribuye como:

$$LM \sim \chi^2_{[\text{Número de variables en } \mathbf{Z}]}$$

Notemos que es una convención utilizar una forma funcional de la varianza a:

$$\sigma_i^2 = \sigma^2 \cdot e^{\mathbf{z}_i\gamma}$$

5.4. Estimación de Sistemas de Ecuaciones con Variables Instrumentales

El enfoque moderno para la estimación del sistema con variables instrumentales se basa en el principio del método generalizado de momentos (GMM). Una clase relacionada de estimadores cae bajo el título de variables instrumentales generalizadas (GIV). También veremos que el estimador GIV, debido a su dependencia de una transformación similar a GLS, es inconsistente en algunos casos importantes cuando GMM permanece consistente.

Partamos de un modelo lineal:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{5.8}$$

Donde $\mathbf{X}_i$ es una matriz $G \times K$ y $\mathbf{Y}_i$ es un vector $G \times 1$. Así, la matriz $\mathbf{X}_i$ continene todas las variables explicativas que aparecen en el sistema, para cada uno de los individuos indexados i.

De forma similar al caso de Variables Instrumentales, podemos asumir que:

$$\mathbb{E}[\mathbf{Z}_i'\boldsymbol{\varepsilon}_i] = \mathbf{0}$$

Donde $\mathbf{Z}_i$ es una matriz $G \times L$ de variables instrumentales y $L \geq K$. También asumiremos que la matriz siguiente es de rango completo:

$$rank\mathbb{E}[\mathbf{Z}_i'\mathbf{X}_i'] = K$$

El proceso de estimación que plantearemos aquí es el del Método Generalizado de Momentos (GMM, por sus siglas en inglés). De esta forma proponemos que la condición del momento poblacional que plantearemos es:

$$\mathbb{E}[\mathbf{Z}_i'(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})] = \mathbf{0} \tag{5.9}$$

De esta forma nuestra propuesta de solución será aquella dada por elegir el estimador $\hat{\boldsymbol{\beta}}$ que resulta de:

$$N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{5.10}$$

La ecuación (5.10) es un conjunto de L ecuaciones con K incógnitas. Para analizar este caso primero consideremos el caso en el que $L = K$, es

decir, estamos en un caso de Variables Instrumentales pero para un sistema de ecuaciones. En este caso sabemos que la matriz siguiente que resulta de despejar $\hat{\boldsymbol{\beta}}$ en (5.10) es de dimensión $K \times K$ y es no singular:

$$\hat{\boldsymbol{\beta}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{X}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Y}_i \right) \tag{5.11}$$

Lo cual podemos reescribir en forma matricial como:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{Y} \tag{5.12}$$

Donde, por se un sistema de ecuaciones, $\mathbf{Z}$ es una matriz de $N \times G \times L$ obtenida por el apilamiento de cada $\mathbf{Z}_i$, para $i = 1, 2, \ldots, N$, $\mathbf{X}$ es una matriz de $N \times G \times K$ obtenida por el apilamiento de cada $\mathbf{X}_i$, para $i = 1, 2, \ldots, N$ y $\mathbf{Y}$ es un vector de $N \times G \times 1$ obtenido por el apilamiento de cada $\mathbf{Y}_i$, para $i = 1, 2, \ldots, N$. De esta forma la ecuación (5.12) es un estimado de Variables Instrumentales para un sistema de ecuaciones.

¿Qué sucede cuando $L > K$? En esos casos tenemos más columnas en la matriz de variables instrumentales $\mathbf{Z}$ que las las necesarias para poder identificar el sistema. Es decir, elegir $\hat{\boldsymbol{\beta}}$ es más complicado. En estos casos proponemos obtener un estimador generado mediante la creación de una matriz de pesos ("weighting matrix") en la forma cuadrática siguiente.

Sea $\hat{\mathbf{W}}$ una matriz simétrica definida positiva, definamos el estimador de GMM de $\hat{\boldsymbol{\beta}}$ como aquel que resulta de resolver el problema:

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} \mathbf{Z}_i' (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) \right]' \hat{\mathbf{W}} \left[\sum_{i=1}^{N} \mathbf{Z}_i' (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) \right] \tag{5.13}$$

Dado que la ecuación (5.13) es una forma cuadrática función de $\boldsymbol{\beta}$, la solución es una solución cerrada.

[FALTA procedimiento para demostrar lo siguiente]

De esta forma la solución estará dada por:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}\mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}\mathbf{Z}'\mathbf{Y} \right) \tag{5.14}$$

5.4.1. Estimación de un sistema por Mínimos Cuadrados en Dos Etapas (2SLS)

Una forma de seleccionar a $\hat{\mathbf{W}}$ es la siguiente:

$$\hat{\mathbf{W}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} = \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right)^{-1} \tag{5.15}$$

Donde la ecuación (5.15) es un estimador consistente de $\mathbb{E}[\mathbf{Z}_i'\mathbf{Z}_i]^{-1}$. Remplazando la ecuación (5.15) en la ecuación (5.14) podemos obtener:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{Z} \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z} \left(\frac{\mathbf{Z}'\mathbf{Z}}{N} \right)^{-1} \mathbf{Z}'\mathbf{Y} \right) \\
&= \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{Y} \right) \\
&= \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \left(\mathbf{Z}'\mathbf{Z} \right) \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{Y} \right) \\
&= \left(\hat{\mathbf{X}}'\hat{\mathbf{X}} \right)^{-1} \hat{\mathbf{X}}'\mathbf{Y}
\end{aligned} \tag{5.16}$$

A la ecuación (5.16) la conocemos como el estimador de 2SLS para un sistema de ecuaciones.

5.4.2. Estimación de un sistema por Mínimos Cuadrados en Tres Etapas (3SLS)

En algunas momentos el estimador GMM se usa con una forma particular de poderadores. Así podemos definir el estimador de Mínimos Cuadrados en Tres Etapas, 3SLS, sean los residuales de una estimación inicial, usualmente, de un 2SLS:

$$\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$$

Definamos la matrix $G \times G$ como

$$\hat{\boldsymbol{\Sigma}} = N^{-1} \sum_{i=1}^{N} \mathbf{e}_i \mathbf{e}_i' \tag{5.17}$$

Usando los mismos argumentos de Mínimos Cuadrados Generalizados Factibles, podemos proponer que:

$$\hat{\mathbf{W}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \hat{\mathbf{\Sigma}} \mathbf{Z}_i \right)^{-1} = \left(\frac{\mathbf{Z}'(\mathbf{I}_N \otimes \hat{\mathbf{\Sigma}})\mathbf{Z}}{N} \right)^{-1} \tag{5.18}$$

Donde $\mathbf{I}_N$ es la matriz identidad. Sustituyendo la ecuación (5.18) en la ecuación (5.14) podemos obtener:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{Z} \left(\frac{\mathbf{Z}'(\mathbf{I}_N \otimes \hat{\mathbf{\Sigma}})\mathbf{Z}}{N} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z} \left(\frac{\mathbf{Z}'(\mathbf{I}_N \otimes \hat{\mathbf{\Sigma}})\mathbf{Z}}{N} \right)^{-1} \mathbf{Z}'\mathbf{Y} \right) \\ &= \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'(\mathbf{I}_N \otimes \hat{\mathbf{\Sigma}})\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'(\mathbf{I}_N \otimes \hat{\mathbf{\Sigma}})\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{Y} \right) \end{aligned} \tag{5.19}$$

A la ecuación (5.19) la conocemos como el estimador de 3SLS para un sistema de ecuaciones.

5.5. Ejercicios

[PENDIENTE, abril 2020]

6

Modelos de Datos Panel

6.1. Introducción y motivación

Datos panel es un caso particular del probema de sistemas de ecuaciones, no obstante, le dedicaremos una sección de estas notas. Un conjunto de datos longitudinales o conjunto de datos panel tiene múltiples observaciones para un número dado de unidades de sección cruzada. Como su nombre lo sugiere un panel tiene dos dimensiones:

1. La primera para las observaciones de sección cruzada
2. La segunda para otro tipo de observaciones, por ejemplo, tiempo u otra característica

En adelante utilizaremos 'grupo' para identificar a la unidad de sección cruzada.

La motivación principal para el concepto de datos panel es el componenete de error del modelo o **sesgo por omisión de información relevante** contenida en el término de error. Esta información puede ser observable o no observable. Condición que da lugar a dos modelo tradicionales de datos panel.

El primer tipo de error está relacionado con información que no cambia a través de la segunda dimensión (por ejemplo, a través del tiempo) y que está capturada por el término de error. Este tipo de información relevante omitida no se puede solucionar por el método de variables instrumentales, puesto que no existen instrumentos como tal. A este planteamiento se le conoce como **efecto fijo**.

De forma similar existen un **efecto aleatorio** que que está contenido en el término de error y no es capturado por el ninguno de los regresores empleados. Es decir, es información del término de error que no está correlacionada con los regresores.

Así, la principal diferencia entre ambos plateamientos, como mostraré más adelante es si la información omitida está correlacionada con lo regresores o no.

6.2. Modelos y métodos de estimación

Al igual que en casos anteriores supondremos que la muestra de datos es una muestra aleatoria. De esta forma tendremos que para cada una de las unidades o grupos de individuos $i = 1, 2, \ldots, n$ tendremos que M observaciones que corresponden con la segunda dimensión de nuestro panel de datos:

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{iM} \end{bmatrix}$$

$$\mathbf{x}_i = \begin{bmatrix} x_{i11} & x_{i12} & x_{i13} & \ldots & x_{i1K} \\ x_{i21} & x_{i22} & x_{i23} & \ldots & x_{i2K} \\ x_{i31} & x_{i32} & x_{i33} & \ldots & x_{i3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{iM1} & x_{iM2} & x_{iM3} & \ldots & x_{iMK} \end{bmatrix}$$

Así, para cada grupo tendremos M ecuaciones expresadas como:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{iM} \end{bmatrix} = \begin{bmatrix} x_{i11} & x_{i12} & x_{i13} & \ldots & x_{i1K} \\ x_{i21} & x_{i22} & x_{i23} & \ldots & x_{i2K} \\ x_{i31} & x_{i32} & x_{i33} & \ldots & x_{i3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{iM1} & x_{iM2} & x_{iM3} & \ldots & x_{iMK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \vdots \\ \varepsilon_{iM} \end{bmatrix}$$

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{6.1}$$

Donde:

$$\mathbb{E}[\boldsymbol{\varepsilon}_i \cdot \mathbf{x}_i] = 0$$

Así, al igual que en el caso del plateamiento del modelo lineal clásico, en datos panel asumiremos una serie de supuestos:

1. Diremos que el modelo será lineal en sus parámetros: $\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$

2. Asumiremos que la matriz $\mathbf{x}_i$ es de rango completo $M \times K$

3. Esperamos que las variables incluidas como regresores sean exógenas, es decir, $\mathbb{E}[\boldsymbol{\varepsilon}_i|\mathbf{x}_j] = 0$ para $i, j = 1, 2, \ldots, n$

4. El término de error es homocedástico y sin autocorrelación

Dicho lo anterior, para la estimación y modelación de datos panel existen tres alternativas que veremos en este curso: estimación considerando efectos fijos, estimación con efectos aleatorios y estimaciones pool ('pooled regression').

6.2.1. Efectos Fijos

En este caso, diremos que $\mathbf{x}_i$ es una matriz de regresores exogénos. Ahora bien, supondremos que la pertubación es un componenete dado por:

$$\varepsilon_{im} = \alpha_i + \eta_{im}$$

Donde $m = 1, 2, \ldots, M$. Al primer componente α_i se le conoce como el componente de los errores que es común a lo largo de la segunda dimensión, a este se le conoce como efecto individual, heterogeneidad individual o efecto fijo.

Así, la ecuación (6.1) se verá modificada para quedar como:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{iM} \end{bmatrix} = \begin{bmatrix} x_{i11} & x_{i12} & x_{i13} & \ldots & x_{i1K} \\ x_{i21} & x_{i22} & x_{i23} & \ldots & x_{i2K} \\ x_{i31} & x_{i32} & x_{i33} & \ldots & x_{i3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{iM1} & x_{iM2} & x_{iM3} & \ldots & x_{iMK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \alpha_i \\ \alpha_i \\ \alpha_i \\ \vdots \\ \alpha_i \end{bmatrix} + \begin{bmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{i3} \\ \vdots \\ \eta_{iM} \end{bmatrix}$$

$$\mathbf{x}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\alpha}_i + \boldsymbol{\eta}_i \tag{6.2}$$

Donde

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\varepsilon}_i \cdot \mathbf{x}_i] &= 0 \\
\mathbb{E}[\alpha_i \cdot x_{im}] &= 0 \\
\text{Obs} &: \text{si la matriz de regresores tiene constante, entonces } m \neq 1 \\
\mathbb{E}[\eta_{im} \cdot x_{im}] &= 0
\end{aligned}$$

Para estimar este modelo se requiere de acomodar la información para cada uno de los grupos de forma que al final quedaría una expresión como:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ \vdots \\ y_{1M} \\ y_{2M} \\ \vdots \\ y_{2M} \\ \vdots \\ y_{n1} \\ y_{n2} \\ y_{n3} \\ \vdots \\ y_{nM} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 & x_{111} & x_{112} & x_{113} & \dots & x_{11K} \\ 1 & 0 & \dots & 0 & x_{121} & x_{122} & x_{123} & \dots & x_{12K} \\ 1 & 0 & \dots & 0 & x_{131} & x_{132} & x_{133} & \dots & x_{13K} \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 & x_{1M1} & x_{1M2} & x_{1M3} & \dots & x_{1MK} \\ 0 & 1 & \dots & 0 & x_{211} & x_{212} & x_{213} & \dots & x_{21K} \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 & x_{2M1} & x_{2M2} & x_{2M3} & \dots & x_{2MK} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & x_{n11} & x_{n12} & x_{n13} & \dots & x_{n1K} \\ 0 & 0 & \dots & 1 & x_{n21} & x_{n22} & x_{n23} & \dots & x_{n2K} \\ 0 & 0 & \dots & 1 & x_{n31} & x_{n32} & x_{n33} & \dots & x_{n3K} \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & x_{nM1} & x_{nM2} & x_{nM3} & \dots & x_{nMK} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix}$$

$$+ \begin{bmatrix} \eta_{11} \\ \eta_{12} \\ \eta_{13} \\ \vdots \\ \eta_{1M} \\ \vdots \\ \eta_{n1} \\ \eta_{n2} \\ \eta_{n3} \\ \vdots \\ \eta_{nM} \end{bmatrix}$$

6.2.2. Efectos Aleatorios

En este caso, igualmente que en el caso anterior, diremos que $\mathbf{x}_i$ es una matriz de regresores exogénos. Ahora bien, supondremos que la pertubación es un componenete dado por:

$$\varepsilon_{im} = \nu_i + \eta_{im}$$

Donde $m = 1, 2, \ldots, M$ y ν_i es una perturbación aleatoria que captura la información estocástica del individuo y que no está correlacionada con los regresores. A este componente ν_i se le conoce como el componente de los errores que es aleatorio a lo largo de la segunda dimensión, a este se le conoce como efecto aleatorio.

Así, la ecuación (6.1) se verá modificada para quedar como:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{iM} \end{bmatrix} = \begin{bmatrix} x_{i11} & x_{i12} & x_{i13} & \ldots & x_{i1K} \\ x_{i21} & x_{i22} & x_{i23} & \ldots & x_{i2K} \\ x_{i31} & x_{i32} & x_{i33} & \ldots & x_{i3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{iM1} & x_{iM2} & x_{iM3} & \ldots & x_{iMK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \nu i \\ \nu i \\ \nu i \\ \vdots \\ \nu i \end{bmatrix} + \begin{bmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{i3} \\ \vdots \\ \eta_{iM} \end{bmatrix}$$

$$\mathbf{x}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\nu}_i + \boldsymbol{\eta}_i \tag{6.3}$$

Donde

$$\begin{aligned} \mathbb{E}[\boldsymbol{\varepsilon}_i \cdot \mathbf{x}_i] &= 0 \\ \mathbb{E}[\nu_i \cdot x_{im}] &= 0 \\ \mathbb{E}[\eta_{im} \cdot x_{im}] &= 0 \end{aligned}$$

Para estimar este modelo se requiere de acomodar la información para cada uno de los grupos de forma similar al modelo de efectos fijos para que

al final quedaría una expresión como:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ \vdots \\ y_{1M} \\ y_{2M} \\ \vdots \\ y_{2M} \\ \vdots \\ y_{n1} \\ y_{n2} \\ y_{n3} \\ \vdots \\ y_{nM} \end{bmatrix} = \begin{bmatrix} x_{111} & x_{112} & x_{113} & \dots & x_{11K} \\ x_{121} & x_{122} & x_{123} & \dots & x_{12K} \\ x_{131} & x_{132} & x_{133} & \dots & x_{13K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1M1} & x_{1M2} & x_{1M3} & \dots & x_{1MK} \\ x_{211} & x_{212} & x_{213} & \dots & x_{21K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{2M1} & x_{2M2} & x_{2M3} & \dots & x_{2MK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n11} & x_{n12} & x_{n13} & \dots & x_{n1K} \\ x_{n21} & x_{n22} & x_{n23} & \dots & x_{n2K} \\ x_{n31} & x_{n32} & x_{n33} & \dots & x_{n3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{nM1} & x_{nM2} & x_{nM3} & \dots & x_{nMK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \nu_1 + \eta_{11} \\ \nu_1 + \eta_{12} \\ \nu_1 + \eta_{13} \\ \vdots \\ \nu_1 + \eta_{1M} \\ \nu_2 + \eta_{21} \\ \vdots \\ \nu_2 + \eta_{2M} \\ \vdots \\ \nu_n + \eta_{n1} \\ \nu_n + \eta_{n2} \\ \nu_n + \eta_{n3} \\ \vdots \\ \nu_n + \eta_{nM} \end{bmatrix}$$

6.2.3. Regresión Pool

Si $\mathbf{x}_i$ contine sólo una constante, entonces estimamos por mínimos cuadrados un estimador consistente y estará dado por los estimadores del vector $\boldsymbol{\beta}$. Asimismo, en este asumiremos que el término de error no se puede separar de fomar que sólo quedará como $\boldsymbol{\varepsilon}_{im}$

Para estimar este modelo se requiere de acomodar la información para cada uno de los grupos de forma similar al modelo de efectos aleatorios para

que al final quedaría una expresión como:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ \vdots \\ y_{1M} \\ y_{2M} \\ \vdots \\ y_{2M} \\ \vdots \\ y_{n1} \\ y_{n2} \\ y_{n3} \\ \vdots \\ y_{nM} \end{bmatrix} = \begin{bmatrix} x_{111} & x_{112} & x_{113} & \dots & x_{11K} \\ x_{121} & x_{122} & x_{123} & \dots & x_{12K} \\ x_{131} & x_{132} & x_{133} & \dots & x_{13K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1M1} & x_{1M2} & x_{1M3} & \dots & x_{1MK} \\ x_{211} & x_{212} & x_{213} & \dots & x_{21K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{2M1} & x_{2M2} & x_{2M3} & \dots & x_{2MK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n11} & x_{n12} & x_{n13} & \dots & x_{n1K} \\ x_{n21} & x_{n22} & x_{n23} & \dots & x_{n2K} \\ x_{n31} & x_{n32} & x_{n33} & \dots & x_{n3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{nM1} & x_{nM2} & x_{nM3} & \dots & x_{nMK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \vdots \\ \varepsilon_{1M} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2M} \\ \vdots \\ \varepsilon_{n1} \\ \varepsilon_{n2} \\ \varepsilon_{n3} \\ \vdots \\ \varepsilon_{nM} \end{bmatrix}$$

6.2.4. Pruebas para seleccionar modelo

Para datos panel existen dos pruebas. La primera es la de Hausman la cual nos permite analizar si la estimación por el efecto fijo no es distinta de la estimación por efectos aleatorios. De esta forma tenemos que:

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_{Fix} = \boldsymbol{\beta}_{Random} \\ H_1 &: \boldsymbol{\beta}_{Fix} \neq \boldsymbol{\beta}_{Random} \end{aligned}$$

Equivalentemente:

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_{Fix} - \boldsymbol{\beta}_{Random} = \mathbf{q} = 0 \\ H_1 &: \boldsymbol{\beta}_{Fix} - \boldsymbol{\beta}_{Random} \neq 0 \end{aligned}$$

Donde $\boldsymbol{\beta}_{Fix}$ es el estimador de parámetros asumiendo efectos fijos y $\boldsymbol{\beta}_{Random}$ es el estimador asumiendo efectos aleatorios.

De esta forma la estadística plateada estará dada por:

$$\begin{aligned} H &= n \cdot \mathbf{q}'(Var(\mathbf{q}))^{-1}\mathbf{q} \\ &= n \cdot (\boldsymbol{\beta}_{Fix} - \boldsymbol{\beta}_{Random})'(Var(\mathbf{q}))^{-1}(\boldsymbol{\beta}_{Fix} - \boldsymbol{\beta}_{Random}) \\ &\sim \chi^{(2)}_K \end{aligned}$$

La segunda, es una prueba para determinar si existen efetos aleatorios. La forma de determinarlo es si existe un componente de varianza asociado a ν, de forma que la hipótesis quedará planteada como una prueba basada en Multiplicadores de Lagrange (LM) de Breush y Pagan:

$$\begin{aligned} H_0 &: \sigma_\nu^2 = 0 \\ H_1 &: \sigma_\nu^2 \neq 0 \end{aligned}$$

6.3. Conclusión

- El modelo de efectos fijo y aleatorios supone que existe información individual que fue omitida.
- La distinción entre efectos fijos y efectos aleatorios es si el efecto no observable individual está correlacionado con los regresores o si no lo está. Así, esto no tiene que ver con si dicho efecto es estocástico o no.
- Cuando todo falla, la alternativa es estimar un modelo pool para los datos panel.

7

Métodos de estimación basados en verosimilitud

7.1. Introducción

En esta sección discutiremos el tratamiento general para la estimación por máxima verosimilitud. La diferencia significativa en estos casos es que asumiremos una distribución subyacente. Noten que los metodos utilizados anteriormente no requieren de que asumamos una distribución particular para realizar la estimación.

El planteamiento general del método de máxima verisimilitud considera un conjunto de observaciones independientes e idénticamente distribuidas descritas por la familia: $\{y_i, \mathbf{x}_i : y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^K\}$.

En estos casos también supondremos que buscamos etimar la regresión de y_i es $\mathbf{x}_i$ y no el caso contario. Por lo tanto, el método requiere de la función de densidad condicional de y_i en $\mathbf{x}_i$, es decir:

$$f(y_i|\mathbf{x}_i)$$

Así, al método en realidad se le conoce como método de máxima verosimilitud condicional. Su aplicación es amplia en casos como:

1. Modelos lineales: regresión de una sola ecuación, sistemas de ecuaciones, modelos de datos panel y modelos generalizados con heterocedásticidad y autocorrelación

2. Modelos no lineales: repuesta binaria, respuesta categórica, respuesta multinivel

Ejemplo. Sea y_i^* una variable aleatoria latente que se pueda expresar como:

$$y_i^* = \mathbf{x}_i\boldsymbol{\theta} + \varepsilon_i$$

Donde ε_i es independiente de $\mathbf{x}_i$, el cual es un vector de $1 \times K$ que contienen un término constante. $\boldsymbol{\theta}$ es un vector de $K \times 1$ parámetros a estimar. Finalmente, asumiremos que:

$$\varepsilon_i \sim Normal(0, 1)$$

Supongamos que en lugar de y_i^* nosotros sólo podemos observar la variable binaria indicadora:

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* \leq 0 \end{cases}$$

En este caso, la pregunta que pretendemos responder sería: ¿cuál es la probabilidad de que $y_i = 1$ condicional en que tenemos el vector $\mathbf{x}_i$?

$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0|\mathbf{x}_i) \\ &= P(\mathbf{x}_i\boldsymbol{\theta} + \varepsilon_i > 0|\mathbf{x}_i) \\ &= P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) \end{aligned}$$

Dada la distribución ε_i, entonces:

$$P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = \int_{-\mathbf{x}_i\boldsymbol{\theta}}^{\infty} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i$$

Sabemos que toda función de densidad cumple con:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i \\ &= \int_{-\infty}^{-\mathbf{x}_i\boldsymbol{\theta}} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i + \int_{-\mathbf{x}_i\boldsymbol{\theta}}^{\infty} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i \\ &= \int_{-\infty}^{-\mathbf{x}_i\boldsymbol{\theta}} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i + \int_{-\infty}^{\mathbf{x}_i\boldsymbol{\theta}} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i \\ &= \Phi(-\mathbf{x}_i\boldsymbol{\theta}) + \Phi(\mathbf{x}_i\boldsymbol{\theta}) \\ &= (1 - \Phi(\mathbf{x}_i\boldsymbol{\theta})) + \Phi(\mathbf{x}_i\boldsymbol{\theta}) \\ &= P(\varepsilon_i < -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) + P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) \end{aligned}$$

De esta forma tenemos dos ecuaciones:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}) \\ P(y_i = 0|\mathbf{x}_i) &= P(\varepsilon_i < -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}) \end{aligned}$$

Estas dos funciones de probabilidad son independientes. De esta forma, la función de densidad condicional de y_i será:

$$\begin{aligned} f(y_i|\mathbf{x}_i) &= P(y_i = 1|\mathbf{x}_i)^{y_i} \cdot P(y_i = 0|\mathbf{x}_i)^{1-y_i} \\ &= \Phi(\mathbf{x}_i\boldsymbol{\theta})^{y_i} \cdot (1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}))^{1-y_i} \end{aligned}$$

7.2. Marco general de estimación por máxima verosimilitud

Sea $f(y_i|\mathbf{x}_i)$ la función de densidad condicional de y_i dado $\mathbf{x}_i$. Sea $\boldsymbol{\theta}$ un conjunto de parámetros de la función. Entonces la función de densidad conjunta de variables aleatorias independientes $\{y_i : y_i \in \mathbb{R}\}$ dados los valores $\{\mathbf{x}_i : \mathbf{x}_i \in \mathbb{R}^K\}$ estará dada por:

$$\Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = f(y_1, y_2, \ldots, y_n|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n;\boldsymbol{\theta}) = L(\boldsymbol{\theta}) \tag{7.1}$$

A la ecuación (7.1) se le conoce como ecuación de verosimilitud. El problema de máxima verosimilitud entonces será:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}) \tag{7.2}$$

Dado que el logaritmo natural es una transformación monotona, podemos decir que el problema de la ecuación (7.2) es equivalente a:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} lnL(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ln\Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} lnf(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \tag{7.3}$$

Para solucionnar el problema se tiene que determinar las condicones de primer y segundo orden, las cuales serán:

$$\frac{\partial}{\partial\boldsymbol{\theta}} lnL(\boldsymbol{\theta}) = \nabla lnL(\boldsymbol{\theta}) \tag{7.4}$$

$$\frac{\partial^2}{\partial^2\boldsymbol{\theta}} lnL(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} lnL(\boldsymbol{\theta}) \cdot \frac{\partial}{\partial\boldsymbol{\theta}} lnL(\boldsymbol{\theta}') = H(\boldsymbol{\theta}) \tag{7.5}$$

La solución estará dada por aquel valor de $\hat{\boldsymbol{\theta}}$ que hace:

$$\frac{\partial}{\partial \boldsymbol{\theta}} lnL(\hat{\boldsymbol{\theta}}) = 0$$

A su vez, la varianza será aquella que resulta de:

$$Var[\hat{\boldsymbol{\theta}}|\mathbf{X}] = (-\mathbb{E}_{\hat{\boldsymbol{\theta}}}[H(\boldsymbol{\theta})])^{-1}$$

7.3. Pruebas de hipótesis

1. Para el caso de pruebas de hipótesis sobre cada uno de los coeficientes estimados $\hat{\theta}_i$, dada la construcción de la función de verosimilitud y que se trata de grandes muestras, la función de distribución de ese coeficiente sera una normal, porlo que las hipótesis planteadas serán asumiendo:

$$Z = \frac{\hat{\theta}_i - \theta_i}{\sqrt{Var(\hat{\theta}_i)}} \sim N(0, Var(\hat{\theta}_i)) \tag{7.6}$$

2. Para pruebas globales utilizaremos la prueba de Wald:

$$W = c(\hat{\boldsymbol{\theta}})'(\hat{C}Var(\hat{\boldsymbol{\theta}})\hat{C}')^{-1}c(\hat{\boldsymbol{\theta}}) \sim \chi^2_{(Q)} \tag{7.7}$$

   Donde $c(\hat{\theta})$ es una función de los coeficienntes, $\hat{C}$ es el jacobiano de $c(\hat{\theta})$, Q es el número de restricciones.

3. Alternativamente, para pruebas sobre los coeficientes podremos utilizar pruebas de razón de verosimilitud:

$$LR = 2[lnL(\hat{\boldsymbol{\theta}}_0) - lnL(\hat{\boldsymbol{\theta}}_R)] \sim \chi^2_{(Q)} \tag{7.8}$$

8

Estimación de modelos no lineales

8.1. Modelos de respuesta binaria

8.1.1. Planteamiento general

En el caso de modelos de respuesta binaria, asumiremos que y_i es una variable aleatoria que toma sólo valores de 0 y 1. Los ejemplos pueden ser amplios. Decisiones sobre consumir o no, pagar o no, ir a una escuela determianda, viajar por tierra o aire. En general diremos que 1 es éxito y 0 es fracaso.

Como en el caso de modelos lineales diremos que y_i es la variable dependiente y $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ un veector de variables independientes o explicativas. Así, en los modelos no lineales de respuesta binaria estaremos interesados en la probabilidades:

$$P(y_i = 1|\mathbf{x}_i) = P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{iK})$$

Dicho lo anterior y de forma similar al de la regresión lineal, podemos establecer el efecto marginal para el caso de una variable independiente x_j continua:

$$EMg_j = \frac{\partial}{\partial x_j} P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{iK})$$

Cuando la variable independiente es una variable dummy:

$$\begin{aligned} EMg_j &= P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ij} = 1, \ldots, x_{iK}) - \\ &\quad P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ij} = 0, \ldots, x_{iK}) \end{aligned}$$

Este tipo de modelos se basan en el modelo de Bernoulli, cuya función de densidad de probabilidad condicional y podemos platearlo como:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= p(x) \\ P(y_i = 0|\mathbf{x}_i) &= 1 - p(x) \\ Var(y_i|\mathbf{x}_i) = p(x) \cdot (1 - p(x)) & \end{aligned}$$

8.1.2. Algunos planteamientos adicionales

Cuando la variable dependiente es binaria (0 y 1) no se puede implementar una ecuación lineal de la forma tradicional, ya que no es posible determinar una ecuación:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

Por otro la, tambien enfrentaremos el mismo problema con modelos de probabilidad lineal:

$$P(y_i = 1|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

Y:

$$P(y_i = 0|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

En ambos casos es dificl lograr la estimación.

8.1.3. Modelos Logit y Probit

Este tipo de modelos suponene que tenemos que existe una variable latente que se puedes expresar como una ecuación lineal dada por:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

Donde ε_i es una variable aleatoria con función de densidad con media cero y distribución simetrica al rededor de cero. Dado lo anterio, para nosotros sólo es visible que:

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* < 0 \end{cases}$$

De esta forma tenemos una estrucutura de la probabilidad dada por:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i > 0) = P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta}) = G(\mathbf{x}_i\boldsymbol{\beta}) \\ P(y_i = 0|\mathbf{x}_i) &= P(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i < 0) = P(\varepsilon_i < -\mathbf{x}_i\boldsymbol{\beta}) = 1 - G(\mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

Donde $\mathbf{x}_i$ es un vector de dimensión $K \times 1$ que contiene al menos el término constante y $\boldsymbol{\beta}$ es un vector de parámetros a estimar, de forma que asumiremos:

$$\mathbf{x}_i\boldsymbol{\beta} = \beta_1 + x_{i2}\beta_2 + \ldots + x_{iK}\beta_K$$

Asumiremos que $G(\cdot)$ es uan función de densidad acumulada de forma que:

$$0 < G(\mathbf{x}_i\boldsymbol{\beta}) < 1 \ , \ \forall \mathbf{x}_i\boldsymbol{\beta} \in \mathbb{R}$$

En este caso utilizaremos dos modelos que dependen de la forma funcional de $G(\cdot)$ que está determinada por la distribución de ε_i. De esta forma tendremos dos modelos: Probit y Logit:

1. **Modelo Probit**, cuando

$$G(\mathbf{x}_i\boldsymbol{\beta}) = \Phi(\mathbf{x}_i\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}_i\boldsymbol{\beta}} \phi(\nu)d\nu$$

   Donde $\phi(\cdot)$ es la densidad de probabilidad normal estándar:

$$\phi(\mathbf{x}_i\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}} e^{-1/2\phi(\mathbf{x}_i\boldsymbol{\beta})}$$

2. **Modelo Logit**, cuando

$$G(\mathbf{x}_i\boldsymbol{\beta}) = \Lambda(\mathbf{x}_i\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}}$$

   Donde $\lambda(\cdot)$ es la función de densidad acumulada logística.

Sin importar el modelo que estemos ocupando, la forma de interpretar el modelo es mediante el efecto marginal, cuando x_j sea una variable continua:

$$EMg_j = \frac{\partial}{\partial x_j} P(y = 1|\mathbf{x}_i) = \frac{\partial}{\partial x_j} G(\mathbf{x}_i\boldsymbol{\beta}) = g(\mathbf{x}_i\boldsymbol{\beta})\beta_j$$

Por su parte, cuando x_j sea una variable dicotómica:

$$\begin{aligned} EMg_j &= P(y=1|\mathbf{x}_i, x_j=1) - P(y=1|\mathbf{x}_i, x_j=0) \\ &= G(\mathbf{x}_i\boldsymbol{\beta}|x_j=1) - G(\mathbf{x}_i\boldsymbol{\beta}|x_j=0) \\ &= G(\beta_1 + x_2\beta_2 + \ldots + \beta_j + \ldots + x_K\beta_K) \\ &\quad -G(\beta_1 + x_2\beta_2 + \ldots + 0 + \ldots + x_K\beta_K) \end{aligned}$$

8.1.4. Estimación

Supongamos n observaciones de una variables aleatoria independientes e idénticamente distribuidas. En estos casos la función de densidad para cada y_i, $i = 1, 2, \ldots, n$, estará dada por:

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i}, \text{ donde } y_i = 0, 1$$

En estos términnos la función de verosimilitud estará dada por:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \Pi_{i=1}^n f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) \\ &= \Pi_{i=1}^n [G(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i} \end{aligned}$$

En versión logaritmica:

$$\begin{aligned} lnL(\boldsymbol{\beta}) &= \sum_{i=1}^n ln(f(y_i|\mathbf{x}_i;\boldsymbol{\beta})) \\ &= \sum_{i=1}^n y_i ln[G(\mathbf{x}_i\boldsymbol{\beta})] + \sum_{i=1}^n (1-y_i) ln[1 - G(\mathbf{x}_i\boldsymbol{\beta})] \end{aligned}$$

Las condiciones de primer orden serán:

$$\frac{\partial}{\partial\boldsymbol{\beta}} lnL(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \frac{g(\mathbf{x}_i\boldsymbol{\beta})}{G(\mathbf{x}_i\boldsymbol{\beta})}\mathbf{x}_i' - \sum_{i=1}^n (1-y_i)\frac{g(\mathbf{x}_i\boldsymbol{\beta})}{1 - G(\mathbf{x}_i\boldsymbol{\beta})}\mathbf{x}_i'$$

Finalmente, las condiciones de segundo orden y varianza estará dada por:

$$\begin{aligned} \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} lnL(\boldsymbol{\beta}) &= -\sum_{i=1}^n \frac{y_i g(\mathbf{x}_i\boldsymbol{\beta})^2}{G(\mathbf{x}_i\boldsymbol{\beta})[1 - G(\mathbf{x}_i\boldsymbol{\beta})]}\mathbf{x}_i'\mathbf{x}_i \\ &= H(\boldsymbol{\beta}) \end{aligned}$$

Por lo tanto:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= [-\mathbb{E}[H(\boldsymbol{\beta})]]^{-1} \\ &= \left[\sum_{i=1}^{n} \frac{y_i g(\mathbf{x}_i\boldsymbol{\beta})^2}{G(\mathbf{x}_i\boldsymbol{\beta})[1-G(\mathbf{x}_i\boldsymbol{\beta})]}\mathbf{x}_i'\mathbf{x}_i\right]^{-1} \end{aligned}$$

8.2. Modelos de respuesta multinomial y ordenada

En esta sección analizaremos modelos de respuesta discreta con más de 2 resultados posibles, los cuales se dividen:

1. Modelos de respuesta no ordenada, son modelos conocidos como de respuesta nominal, donde los valores de los distintos resultados son arbitrarios y no tienen un efecto en la estimación. Ejemplos de estos modelos son casos de selección de seguro, lugar de hospedaje, etc.

2. Modelos de respuesta ordenada, en estos modelos se asigna a cada un valor no arbitrario a la respuesta. Por ejemplo, modelos de calificación de crédito, modelos de preferencia de bienes, etc.

8.2.1. Modelos de respuesta multinomial

Logit multinomial

Este primer modelo aplica en situaciones en las que la unidad de respuesta o elección depende de las características individuales de los elementos de la muestra, pero no de la atribuciones de la elección.

Dado lo anterior, definiremos el modelo en términos de variables aleatorias que representan a la población subyacente. Sea y_i una variable aleatoria que toma valores en un conjunto $\{0, 1, 2, \ldots, J\}$. Sea $\mathbf{x}_i$ un conjunto de valores o regresores para $i = 1, 2, \ldots, n$.

Como en el modelo de respuesta binaria, estimaremos el efecto de una variable en la probabilidad de ocurrencia de uno de los valores de respuesta. Dichas variables de respuesta estarán dados por:

$$P(y_i = j|\mathbf{x}_i), \text{ para } j = 1, 2, \ldots, J$$

Dado que la suma de probabilidades debe ser 1, se sule tomar como práctica estimar $P(y_i = 0|\mathbf{x}_i)$ como diferencia del resto de los casos $j = \{1, 2, \ldots, J\}$. Así, sea $\mathbf{x}_i$ un vector de dimensión $1 \times K$. De esta forma, el modelo Logit tendrá como probabilidad de cada una de las respuestas a:

$$P(y_i = j|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_j}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}}$$

Donde cada una de las $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$ es de dimensión $K \times 1$. Es facil observar que si la suma de probabilidades es 1, entonces, la probabilidad de $y = 0|\mathbf{x}_i$ estará dada por:

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i) &= \frac{e^0}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \\ &= \frac{1}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \end{aligned}$$

De esta forma tenemos:

$$\begin{aligned} 1 &= \sum_{j=0}^{J} P(y_i = j|\mathbf{x}_i) \\ &= \frac{1}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \\ &\quad + \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_1}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \\ &\quad + \ldots + \\ &\quad + \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_J}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \end{aligned}$$

Para el proceso de estimación consideraremos la función de verosimilitud dada por:

$$\begin{aligned} L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) &= \Pi_{i=1}^{n} \left[(P(j = 0|\mathbf{x}_i))^{I(0)} \cdot (P(j = 1|\mathbf{x}_i))^{I(1)} \right. \\ &\quad \left. \cdots (P(j = J|\mathbf{x}_i))^{I(J)} \right] \end{aligned}$$

Donde $I(j)$ es la función indicadora que toma el valor de 1 o 0 dependiendo de cual de las opciones posibles $j = 0, 1, 2, \ldots, J$ es cierta para cada individuo $i = 1, 2, \ldots, n$. De forma similar al caso de respuesta binaria en los modelos Logit Multinomiales se interpretan en sus efectos marginales y no de forma directa en los coeficientes $\boldsymbol{\beta}_j$.

Para una variable x_{ik} continua y para cada opción $j = 0, 1, 2, \ldots, J$ el efecto marginal será:

$$\begin{aligned} EMg_k &= \frac{\partial P(y_i = j|\mathbf{x}_i)}{\partial x_{ik}} \\ &= P(y_i = j|\mathbf{x}_i) \cdot \left[\beta_{jk} - \frac{\sum_{h=1}^{J} \beta_{hk} \cdot e^{\mathbf{x}_i \boldsymbol{\beta}_h}}{1 + e^{\mathbf{x}_i \boldsymbol{\beta}_1} + e^{\mathbf{x}_i \boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i \boldsymbol{\beta}_J}} \right] \end{aligned}$$

Donde β_{hk} es el k-èsimo elemento del vector $\boldsymbol{\beta}_h$. En el caso de que x_{ik} se una variable dicotómica (o categórica) y para cada opción $j = 0, 1, 2, \ldots, J$ el efecto marginal será:

$$EMg_k = P(y_i = j|\mathbf{x}_i, x_{ik} = 1) - P(y_i = j|\mathbf{x}_i, x_{ik} = 0)$$

8.2.2. Modelos de respuesta ordenada: Logit y Probit Ordinal

Sea y_i una variable que representa ordenada que toma los valores de $\{0, 1, 2, \ldots, J\}$. En estos casos en que importa el órden de la respuesta existen dos mecanismos de estimación: el Modelo Probit y el Modelo Logit.

Modelo Probit Ordinal. Este modelo se puede derivar de forma similar al modelo probit de dos respuestas 0 y 1, probit binario. Así, partimos de una variable latente y_i^* definida como:

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i; \ \boldsymbol{\varepsilon}_i \sim N(0, 1)$$

Donde $\boldsymbol{\beta}$ es un vector $(K - 1) \times 1$ ya que en este caso asumiremos que no se considera un término constante, por las razones que mostramos más adelante. Sean $\alpha_1, \alpha_2, \ldots, \alpha_J$ un conjunto de umbrales que particionan la probabilidad de la siguiente forma:

$$\begin{aligned} y_i = 0 \quad &\text{si} \quad y_i^* \leq \alpha_1 \\ y_i = 1 \quad &\text{si} \quad \alpha_1 < y_i^* \leq \alpha_2 \\ &\vdots \\ y_i = j \quad &\text{si} \quad \alpha_j < y_i^* \leq \alpha_{j+1} \\ &\vdots \\ y_i = J \quad &\text{si} \quad \alpha_J < y_i^* \end{aligned}$$

Dado que $\boldsymbol{\varepsilon}_i$ tiene una distribución normal estándar, para casa uno de los casos anteriores podemos establecer:

$$\begin{aligned}
P(y_i = 0|\mathbf{x}_i) &= P(y_i^* \leq \alpha_1|\mathbf{x}_i) = P(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_1|\mathbf{x}_i) \\
P(y_i = 1|\mathbf{x}_i) &= P(\alpha_1 < y_i^* \leq \alpha_2|\mathbf{x}_i) = P(\alpha_1 < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_2|\mathbf{x}_i) \\
&\vdots \\
P(y_i = j|\mathbf{x}_i) &= P(\alpha_j < y_i^* \leq \alpha_{j+1}|\mathbf{x}_i) = P(\alpha_j < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_{j+1}|\mathbf{x}_i) \\
&\vdots \\
P(y_i = J|\mathbf{x}_i) &= P(\alpha_J < y_i^*|\mathbf{x}_i) = P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i)
\end{aligned}$$

Desarrollando las ecuaciones anteriores tenemos para el caso de $y_i = 0$:

$$\begin{aligned}
P(y_i = 0|\mathbf{x}_i) &= P(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_1|\mathbf{x}_i) \\
&= P(\boldsymbol{\varepsilon}_i \leq \alpha_1 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= \Phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})
\end{aligned}$$

En el caso de $y_i = j$, $j = 1, 2, \ldots, J-1$:

$$\begin{aligned}
P(y_i = j|\mathbf{x}_i) &= P(\alpha_j < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_{j+1}|\mathbf{x}_i) \\
&= P(\alpha_j - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= P(\boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) - (\boldsymbol{\varepsilon}_i < \alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= \Phi(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})
\end{aligned}$$

Finalmente, para el caso $y_i = J$:

$$\begin{aligned}
P(y_i = J|\mathbf{x}_i) &= P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\
&= P(\alpha_J - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\
&= 1 - P(\boldsymbol{\varepsilon}_i < \alpha_J - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= 1 - \Phi(\alpha_J - \mathbf{x}_i\boldsymbol{\beta})
\end{aligned}$$

Es sencillo observar y analizar que:

$$1 = \sum_{j=0}^{J} P(y_i = j|\mathbf{x}_i)$$

Notemos que en este modelo los cambios en la probabilidad no serán determinados por $\mathbf{x}_i$, ya que lo que determina el cambio es el umbral α_j,

$j = 0, 1, \ldots, J$. También en este caso los coeficientes no se interpretan de forma directa, en cambio se utiliza el efecto marginal que tiene expresiones similares al del probit bivariado como más adelante mostraremos.

Modelo Logit Ordinal. Este modelo se puede derivar de forma similar al modelo Probit Ordinal con una variable latente y_i^*, también con $\boldsymbol{\beta}$ es un vector $(K-1) \times 1$ ya que en este caso asumiremos que no se considera un término constante, por las razones que hemos mostrado anteriormente. Igualmente, sean $\alpha_1, \alpha_2, \ldots, \alpha_J$ un conjunto de umbrales que particionan la probabilidad considerando una función logistica $\Lambda(\cdot)$ de la siguiente forma:

$$\boldsymbol{\varepsilon}_i \sim \Lambda(\cdot)$$

$$P(y_i = J|\mathbf{x}_i) \;=\; P(\alpha_J < y_i^*|\mathbf{x}_i) = P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i)$$

Desarrollando las ecuaciones de probabilidad de forma similar al Probit Multinomial tenemos para el caso de $y_i = 0$:

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i) &= P(y_i^* \leq \alpha_1|\mathbf{x}_i) \\ &= P(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_1|\mathbf{x}_i) \\ &= P(\boldsymbol{\varepsilon}_i \leq \alpha_1 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= \Lambda(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

En el caso de $y_i = j$, $j = 1, 2, \ldots, J-1$:

$$\begin{aligned} P(y_i = j|\mathbf{x}_i) &= P(\alpha_j < y_i^* \leq \alpha_{j+1}|\mathbf{x}_i) \\ &= P(\alpha_j < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_{j+1}|\mathbf{x}_i) \\ &= P(\alpha_j - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= P(\boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) - (\boldsymbol{\varepsilon}_i < \alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= \Lambda(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Lambda(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

Finalmente, para el caso $y_i = J$:

$$\begin{aligned} P(y_i = J|\mathbf{x}_i) &= P(\alpha_J < y_i^*|\mathbf{x}_i) \\ &= P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\ &= P(\alpha_J - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\ &= 1 - P(\boldsymbol{\varepsilon}_i < \alpha_J - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= 1 - \Lambda(\alpha_J - \mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

También, es sencillo observar y analizar que en los Logit Multinomiales:

$$1 = \sum_{j=0}^{J} P(y_i = j|\mathbf{x}_i)$$

Notemos que en este modelo los cambios en la probabilidad no serán determinados por $\mathbf{x}_i$, ya que lo que determina el cambio es el umbral α_j, $j = 0, 1, \ldots, J$. También en este caso los coeficientes no se interpretan de forma directa, en cambio se utiliza el efecto marginal que tiene expresiones similares al del probit bivariado como más adelante mostraremos.

Estimación de modelos Probit y Logit Ordinal. Para la estimación implementaremos un proceso de maximización de de la siguiente función de verosimilitud:

$$L(\alpha, \boldsymbol{\beta}) = \Pi_{i=1}^{n} P(y_i = 0|\mathbf{x}_i)^{I(0)} \cdot P(y_i = 1|\mathbf{x}_i)^{I(1)} \cdots P(y_i = J|\mathbf{x}_i)^{I(J)}$$

Donde $I(j)$ es la función indicadora que toma el valor de 0 y 1 si el individuo i-ésimo tiene como respuesta alguno de los posibles valores de $j = 0, 1, \ldots, J$. La función es válida sin importar la función $\Phi(\cdot)$ o $\Lambda(\cdot)$ que utilicemos para determinar la $P(y_i = j|\mathbf{x}_i)$, para $j = 1, 2, \ldots, J$. Por lo cual simbolizaremos indistintamente estas funciones como $G(\cdot)$.

Finalmente, sin importar el modelo que estemos ocupando, Probit o Logit ordinal, la forma de interpretar el modelo es mediante el efecto marginal, cuando x_k sea una variable continua:

$$EMg_k = \frac{\partial}{\partial x_k} P(y_i = j|\mathbf{x}_i) = \frac{\partial}{\partial x_k} G(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}) = -\beta_k \cdot g(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})$$

Por su parte, cuando x_j sea una variable dicotómica:

$$\begin{aligned} EMg_k &= P(y_i = j|\mathbf{x}_i, x_{ik} = 1) - P(y_i = j|\mathbf{x}_i, x_{ik} = 0) \\ &= G(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i, x_{ik} = 1) - G(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i, x_{ik} = 0) \\ &= G(\alpha_j - (\beta_1 + x_2\beta_2 + \ldots + \beta_k + \ldots + x_K\beta_K)) \\ &\quad -G(\alpha_j - (\beta_1 + x_2\beta_2 + \ldots + 0 + \ldots + x_K\beta_K)) \end{aligned}$$

También es posible hacer un efecto marginal intra respuestas, es decir, para cambios entre una respuesta j y una $j + h$, la cual será:

$$EMg_{j-h} = \beta_k \cdot [g(\alpha_h - \mathbf{x}_i\boldsymbol{\beta}) - g(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})], \text{ con } h < j$$

Donde $k = 1, 2, \ldots, K - 1$ y $j = 0, 1, \ldots, J$.

8.3. Modelos de conteo y otras respuestas

Una variabe de conteo es una variable que toma valores enteros no negativos y que cumple con dos características:

1. No tienen, de forma natural, una cuota superior
2. El valor cero debe ser posible para algunos de los elementos de la muestra

Muchas variables que conocemos cumplen con estasn condiciones. Por ejemplo, el número de veces que un individuo es detenido al año, el número de cigarros que una persona fuma al día, el número de patentes que una empresa registra en un año, etc. Por el contrario, el número de hijos de una familia que se gradua de la universidad no cumple con la condición 1, ya que ese número está topado por el número de hijos que tiene una familia.

8.3.1. Modelos de conteo del tipo Poisson

En esta sección analiazremos sólo uno de los modelos de conteo, el mode de conteo basado en una función de distribución del tipo Poisson. Así, sea y una variable de conteo y $\mathbf{x}$ un vector de variables explicativas. De esta forma estaremos interesados en la regresión poblacional dada por:

$$\mathbb{E}[y|\mathbf{x}] = \mu(\mathbf{x})$$

El modelo Poisson es el más utilizado por su simplicidad para implementar mediante un procedimiento de máxima verosimilitud. Los supuestos del modelo son los siguientes.

En primer lugar, asumiremos que existe una función de densidad para y, $f(y|\mathbf{x}) = f(y|x_1, x_2, \ldots, x_K)$, que tiene la forma de Poisson:

$$f(y|\mathbf{x}) = \frac{e^{-\mu(\mathbf{x})} \cdot \mu(\mathbf{x})^y}{y!} \tag{8.1}$$

Donde $\mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ y $y = 0, 1, \ldots$. En segundo lugar, asumiremos que:

$$\mu(\mathbf{x}) = e^{\mathbf{x}\boldsymbol{\beta}}$$

donde $\mathbf{x}$ es un vecto de $1 \times K$ y $\boldsymbol{\beta}$ es una vector de $K \times 1$. Asumiremos que dada la muestra $\{(\mathbf{x}_i, y_i); i = 1, 2, \ldots, N\}$ es factible estimar los parámetros $\boldsymbol{\beta}$ mediante un procedimiento de Máxima Verosimilitud.

Un tercer supuesto es que se cumple la propiedad de la distribución Poisson en la que:

$$Var(y|\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = \mu(\mathbf{x})$$

No obstante, es posible observar que planteamientos en los que asumimos que:

$$Var(y|\mathbf{x}) = \sigma^2\mathbb{E}[y|\mathbf{x}]$$

Donde $\sigma^2 > 0$ es una razón media de la varianza para cual observamos dos casos:

- Si $\sigma^2 > 1$ entonces la varianza es más grande que la media, por lo que estaremos en el caso en que existe una sobredispersión
- Si $\sigma^2 < 1$ entonces la varianza es menor que la media, por lo que estaremos en el caso en ue existe una subdispersión

Ahora platearemos el proceso de estimación. Para lo cual requermimos establecer la función de verosimilitud a partir de la ecuación (8.1) para quedar de la siguiente forma:

$$L(\boldsymbol{\beta}) = \Pi_{i=1}^{n} f(y_i|\mathbf{x}_i) = \frac{e^{-\mu(\mathbf{x}_i)} \cdot \mu(\mathbf{x}_i)^{y_i}}{y_i!} \tag{8.2}$$

Tomando el logaritmo tenemos:

$$lnL(\boldsymbol{\beta}) = \sum_{i=1}^{n} ln(f(y_i|\mathbf{x}_i)) = \sum_{i=1}^{n} -\mu(\mathbf{x}_i) + y_i ln(\mu(\mathbf{x}_i)) - ln(y_i!) \tag{8.3}$$

Donde $\mu(\mathbf{x}_i) = e^{\mathbf{x}_i\boldsymbol{\beta}}$. La estimación de los parámetros $\boldsymbol{\beta}$ serán resultado de igualar a cero y resolver la ecuación:

$$\frac{\partial lnL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} -\mu(\mathbf{x}_i) + y_i ln(\mu(\mathbf{x}_i)) - ln(y_i!) \tag{8.4}$$

La varianza de los estimadores estará dada por:

$$Var(\boldsymbol{\beta}) = [-\mathbb{E}[H(\boldsymbol{\beta})]]^{-1} = \left[-\mathbb{E}\left[\frac{\partial^2 lnL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right]\right]^{-1} \tag{8.5}$$

Finalmente, de forma similar a los modelos de elección binaria, multinomial y múltiple ordenada, los resultados se interpretan en sus efectos marginales. Esos efectos, en el caso de una variable continua, serán:

$$EMg_j = \frac{\partial \mathbb{E}[y_i|\mathbf{x}_i]}{\partial x_{ij}} = e^{\mathbf{x}_i\boldsymbol{\beta}} \cdot \beta_j$$

En el caso de variables dicótomicas:

$$EMg_j = \mathbb{E}[y_i|\mathbf{x}_i, x_{ij} = 1] - \mathbb{E}[y_i|\mathbf{x}_i, x_{ij} = 0]$$

9

Modelos con datos truncados, censurados, con endogeneidad por selección de muestra y otros casos

En esta sección discutiremos dos casos generales conocidos como de truncamiento y censura de datos. Ambos casos dan pie a a casos partículares como los modelos Tobit y de selección de muestra, en el cual el ejemplo más conocido es el ejemplo de Heckman.

Decimos que existe truncamiento cuando hacemos inferencia sobre una población a partir de una muestra que fue tomada de una sección de dicha población. Por otro lado, censura es una condición más común de los datos. Esta situación se presenta como un defecto de los datos, presumiblemente, si los datos no estubieran censurados estos deberían ser posiblemente observados mendiante muestreos de la población de interés. Analicemos primero una forma de truncamiento y luego una de censura.

9.1. Truncamiento

Truncamiento es una condición observada cuando se toma una muestra a partir de una población que ha sido restringida o truncada. Para resolver este tipo de problemas requerimos del uso de funciones de distribución truncadas. Las cuales son una parte por encima o por debajo de algún valor especifico

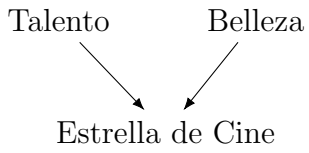


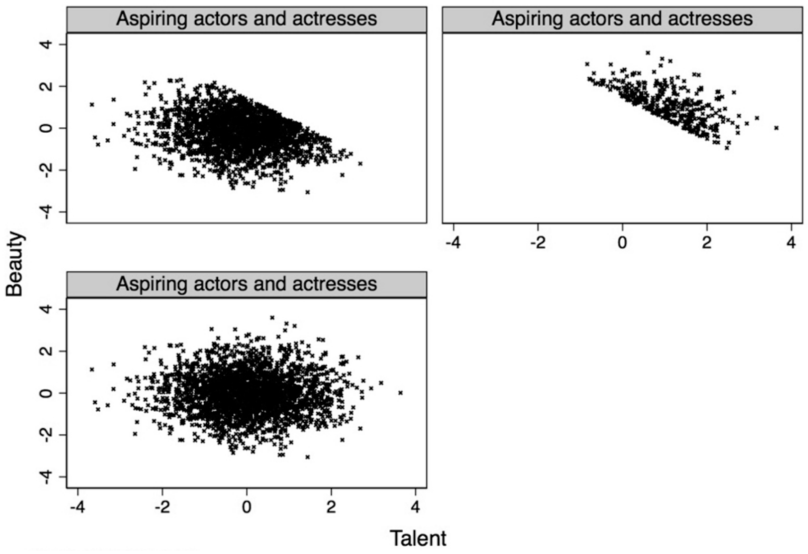
Figura 9.1: ¿Talento y la belleza están negativamente correlacionados?

en el soporte de la función original. Por ejemplo, cuando analizamos la distribución del ingreso de las personas para aquellos individuos que ganan más de una cierta cantidad dada.

Tomemos un ejemplo más retomado de Cunningham (2021) [Cun21]. Una publicación de blog de CNN de 2009 informó que Megan Fox, quien protagonizó la película Transformers, fue votada como la peor y más atractiva actriz de 2009 en una encuesta sobre estrellas de cine. La publicación infería que había una implicación de que el talento y la belleza están negativamente correlacionados. ¿Pero lo son? ¿Y por qué podrían serlo? ¿Qué pasa si en realidad son independientes entre sí pero están correlacionados negativamente en una muestra de estrellas de cine debido al sesgo del colisionador? Es ¿eso es posible? Un diagrama que ilustra la situación anterio es el mostrado en la figura 9.1.

En la figura 9.2 muestra una simulación de la situación. El panel inferior izquierdo muestra el diagrama de dispersión entre talento y belleza. Observe que las dos variables son extracciones aleatorias e independientes de la distribución normal estándar, lo que crea una nube de datos alargada. Pero debido a que .^estrella de cine"se encuentra en el percentil 85 superior de la distribución de una combinación lineal de talento y belleza, la muestra consta de personas cuya puntuación combinada se encuentra en la parte superior derecha de la distribución conjunta. Esta frontera tiene una pendiente negativa y se encuentra en la parte superior derecha de la nube de datos, lo que crea una correlación negativa entre las observaciones en la muestra de estrellas de cine.Sin embargo, sabemos que, de hecho, no hay relación entre las dos variables. Este tipo de selección de muestras crea correlaciones espurias. Una muestra aleatoria de la población completa sería suficiente para mostrar que no existe relación entre las dos variables, pero al dividir la muestra en estrellas de cine únicamente, introducimos correlaciones espurias entre las dos variables de interés.

Figura 9.2: Relación entre belleza y talento, retomado de [Cun21].



Demos un poco de formalidad el problema. Digamos que la función de densidad truncada sera aquella función continua de la variable x dada por $f(x)$ y que dada una constante a, la definiremos como:

$$f(x|x > a) = \frac{f(x)}{P(x > a)} \tag{9.1}$$

En estos casos, la condición más común es emplear una función de densidad normal truncada. Así, si x tiene una distribución normal con media μ y desviación estándar σ, entonces:

$$P(x > a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{x-\mu}{\sigma^2}} dx = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

Donde $\Phi(\cdot)$ es la función de probabilidad acumulada de la distribución normal estándar. De esta forma, podemos construir la función de densidad

truncada como:

$$\begin{aligned} f(x|x>a) &= \frac{f(x)}{1-\Phi(\alpha)} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2}}{1-\Phi(\alpha)} \\ &= \frac{\frac{1}{\sigma}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2}}{1-\Phi(\alpha)} \\ &= \frac{\frac{1}{\sigma}\cdot\phi(\alpha)}{1-\Phi(\alpha)} \end{aligned} \tag{9.2}$$

Donde $\phi(\cdot)$ es la función de densidad de probabilidad de una normal estándar, lo cual es cierto dado el Teorema del Límite Central.

Dado lo anterior podemos platear el concepto de regresión truncada. Para lo cual asumiremos que tenemos una ecuación de regresión del tipo:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

Donde, de forma convenxional, $\mathbf{x}_i$ es una vector de variables explicativas de $1 \times K$ y $\boldsymbol{\beta}$ es un vector de parámetros de $K \times 1$, y

$$\varepsilon|\mathbf{x}_i \sim N(0, \sigma^2)$$

Así:

$$y_i|\mathbf{x}_i \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$$

En este caso estamos interesados en la distribución de y_i dado que y_i es más gande que el punto de truncamiento a. Así, diremos que estamos interesados en la recta de regresión que estará dada por:

$$\mathbb{E}[y_i|y_i > a] = \mathbf{x}_i\boldsymbol{\beta} + \sigma\frac{\phi\left(\frac{a-\mathbf{x}_i}{\sigma}\right)}{1-\Phi\left(\frac{a-\mathbf{x}_i}{\sigma}\right)} \tag{9.3}$$

De esta forma la media condicional es una función no lineal de a, σ, $\mathbf{x}_i$ y $\boldsymbol{\beta}$. Digamos que tenemos:

$$\mathbb{E}[y_i|y_i > a] = \mathbf{x}_i\boldsymbol{\beta} + \sigma\lambda(\alpha_i) \tag{9.4}$$

Donde $\alpha_i = \left(\frac{a-\mathbf{x}_i}{\sigma}\right)$. Por simplicidad utilizaremos las siguiente notación: $\lambda_i = \lambda(\alpha_i)$ y $\delta_i = \delta(\alpha_i)$. De esta forma planteamos la ecuación a estimar será:

$$\begin{aligned} y_i|y_i > a &= \mathbb{E}[y_i|y_i > a] + \varepsilon_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + \sigma\lambda_i + \varepsilon_i \end{aligned}$$

Donde:

$$Var(\varepsilon_i) = \sigma^2(1 - \lambda_i^2 + \lambda_i\alpha_i) = \sigma^2(1 - \delta_i)$$

De forma similar:

$$Var(y_i|y_i > a) = \sigma^2(1 - \delta_i)$$

Es decir, una regresión truncada tiene errores heterocedásticos. La estimación de este tipo de modelos es por un procedimiento de máxima verosimilitud.

Ahora platearemos los efectos marginales como:

$$\begin{aligned} \frac{\partial\mathbb{E}[y_i|y_i > a]}{\partial\mathbf{x}_i} &= \boldsymbol{\beta} + \sigma^2\frac{\partial\lambda_i}{\partial\alpha_i}\frac{\partial\alpha_i}{\partial\mathbf{x}_i} \\ &= \boldsymbol{\beta} + \sigma^2(\lambda_i^2 - \alpha_i\lambda_i) - \frac{\boldsymbol{\beta}}{\sigma} \\ &= \boldsymbol{\beta}(1 - \lambda_i^2 + \alpha_i\lambda_i) \\ &= \boldsymbol{\beta}(1 - \delta_i) \end{aligned}$$

Ahora veamos un caso particular del truncamiento, partícularmente el caso de truncamiento incidental. Supongamos que tenemos dos variables aleatorias y y z, las cuales tienen una distribución conjunta bivariada con una coeficiente de correlación ρ. En estos casos estamos interesados en la distribución de y dado que z excede un valor particular.

La intuición sugiere que si y y z están correlacionadas positivamente, entonces el truncamiento de z debería desplazar la distribución de y a la derecha. Así, estamos innteressados en:

1. La forma del truncamiento incidental

2. La media y varianza del truncamiento incidental

En estos casos la función de densidad conjunta de y y z estará dada por:

$$f(y, z|z > a) = \frac{f(y, z)}{P(z > a)}$$

Si y y z tienen una distribución normal conjunta bivarada con medias μ_y y μ_z; desviaciones estándar σ_y y σ_z, y un coeficiente de correlación ρ, entonces la media será:

$$\mathbb{E}[y|z > a] = \mu_y + \rho\sigma_y\lambda$$

Donde:

$$\lambda = \frac{\phi\left(\frac{a-\mu_z}{\sigma_z}\right)}{1 - \Phi\left(\frac{a-\mu_z}{\sigma_z}\right)}$$

Por su parte la varianza será:

$$Var(y|z > a) = \sigma_y^2[1 - \sigma^2\delta(\alpha_z)]$$

Donde:

$$\delta = \frac{\lambda}{\lambda - \left(\frac{a-\mu_z}{\sigma_z}\right)}$$

Como mencionamos anteriormente, la media condicional está desplazada en el sentido de la correlación entre las variables aleatorias y y z. En su caso el desplazamiento dependerá de si la correlación es positiva o negativa.

Adicionalmente, supongamos un modelo lineal subyacente:

$$z_i^* = \mathbf{w}_i\boldsymbol{\gamma} + u_i$$

y un proceso lineal dado por:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

Dado lo anterior, podemos establecer la regla de selección de muestra como y_i es observada simepte que z_i^* es más grande que cero. De esta forma planteamos la ecuación de regresión como:

$$\begin{aligned}
\mathbb{E}[y_i|y_i \text{ es observada}] &= \mathbb{E}[y_i|z_i^* > 0] \\
&= \mathbb{E}[y_i|\mathbf{w}_i\boldsymbol{\gamma} + u_i > 0] \\
&= \mathbb{E}[y_i|u_i > -\mathbf{w}_i\boldsymbol{\gamma}] \\
&= \mathbb{E}[\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i|u_i > -\mathbf{w}_i\boldsymbol{\gamma}] \\
&= \mathbf{x}_i\boldsymbol{\beta} + \mathbb{E}[\varepsilon_i|u_i > -\mathbf{w}_i\boldsymbol{\gamma}] \\
&= \mathbf{x}_i\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i \\
&= \mathbf{x}_i\boldsymbol{\beta} + \beta_\lambda\lambda_i
\end{aligned}$$

Donde:

$$\lambda_i = \frac{\phi\left(\frac{\mathbf{w}_i\boldsymbol{\gamma}}{\sigma_u}\right)}{\Phi\left(\frac{\mathbf{w}_i\boldsymbol{\gamma}}{\sigma_u}\right)}$$

Así, podemos plantear el modelo como:

$$\begin{aligned} y_i|z_i^* > 0 &= \mathbb{E}[y_i|\mathbf{w}_i\boldsymbol{\gamma} + u_i > 0] + v_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + \beta_\lambda\lambda_i + v_i \end{aligned}$$

Dicho lo anterior podemos representar el problema como, un modelo subyacente:

$$z_i^* = \mathbf{w}_i\boldsymbol{\gamma} + u_i$$

De forma que:

$$\begin{aligned} z_i^* = 1 \text{ si } z_i > 0 \\ z_i^* = 0 \text{ si } z_i \geq 0 \end{aligned}$$

Lo cual vamos a representar como:

$$\begin{aligned} P(z_i = 1|\mathbf{w}_i) = \Phi(\mathbf{w}_i\boldsymbol{\gamma}) \\ P(z_i = 0|\mathbf{w}_i) = 1 - \Phi(\mathbf{w}_i\boldsymbol{\gamma}) \end{aligned}$$

Junto con la ecuación lineal:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \text{ sólo si } z_i = 1$$

Donde:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim BivariateNormal(0, 0, 1, \sigma_\varepsilon, \rho)$$

Con estos ajustes podemos llegar a:

$$\mathbb{E}[y_i|z_i = 1, \mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i\boldsymbol{\beta} + \beta_\lambda\lambda_i$$

Donde:

$$\lambda_i = \frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}$$

Este modelo se le conoce como de Heckman (1979), el cual se estima en dos etapas:

1. Estimar el probit asociado para obtener $\hat{\boldsymbol{\gamma}}$ y

$$\hat{\lambda}_i = \frac{\phi\left(\mathbf{w}_i\hat{\boldsymbol{\gamma}}\right)}{\Phi\left(\mathbf{w}_i\hat{\boldsymbol{\gamma}}\right)}$$

$$\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \mathbf{w}_i\hat{\boldsymbol{\gamma}})$$

2. Estimamos $\hat{\boldsymbol{\beta}}$, $\hat{\beta}_\lambda = \rho\hat{\sigma}_\varepsilon$

9.2. Datos con censura

La censura es una condición más común que el truncamiento, ya que este último ocurre muchas veces por decisión del investigador. La censura es un problema observado en la distribución de la variable dependiente. Por ejemplo, es común encontrar variables que tienen un número considerable de valores cero: gasto en cietos bienes, varaibles que no son de la preferencia de todos los consumidores, etc. No obstate, existe censura en cualquier parte de la distribución.

En este tipo de modelos asumiremos un modelo subyacente dado por:

$$y_i = 0 \text{ si } y_i^* \leq b$$
$$y_i = y_i^* \text{ si } y_i^* > b$$

Donde b es una constate a partir de la cual se observa la censura y donde la ecuación del modelo subyacente es de la forma lineal:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

En este caso analizamos el problema con el modelo de regresión de Tobin (1958), denominado Tobit, en alución a Tobit y al uso de la función de densidad normal como función de distribución de la variable dependiente. En este modelo la forma del valor esperado de y_i estará dada por:

$$\mathbb{E}[y_i|\mathbf{x}_i] \neq \mathbf{x}_i\boldsymbol{\beta} \tag{9.5}$$

Lo cierto, es que a diferencia de la ecuación (9.5) el valor esperado de y_i^* sí es de una forma lineal como:

$$\mathbb{E}[y_i^*|\mathbf{x}_i] = \mathbf{x}_i\boldsymbol{\beta} \tag{9.6}$$

De esta manera, el valor esperado de y_i deberá ser ajustado para quedar como:

$$\mathbb{E}[y_i|\mathbf{x}_i] = \Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}_i\boldsymbol{\beta} + \sigma\lambda_i) \tag{9.7}$$

Donde σ es la desviación estándar de y_i y:

$$\begin{aligned}
\lambda_i &= \frac{\phi\left(\frac{b-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}{1-\Phi\left(\frac{b-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)} \\
&= \frac{\phi\left(\frac{0-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}{1-\Phi\left(\frac{0-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)} \\
&= \frac{\phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}
\end{aligned}$$

Donde $\phi(\cdot)$ y $\Phi(\cdot)$ son las funciones de densidad y acumulada de la normal estándar. Además de que hemos asumido que la constante b es absorvida por la constante contenida en el vector $\boldsymbol{\beta}$

Dicho lo anterior, los efectos marginales de la ecuación (9.7) ahora serán:

$$\frac{\partial\mathbb{E}[y_i|\mathbf{x}_i]}{\partial x_k} = \beta_k\Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \tag{9.8}$$

De forma particular, para la ecuación (9.6) el efecto marginal será dado por:

$$\frac{\partial\mathbb{E}[y_i^*|\mathbf{x}_i]}{\partial x_k} = \beta_k \tag{9.9}$$

Finalmente, en el caso de estimación, este es un caso más que se suele estimar por un proceso de máxima verosimilitud. Partamos de establecer la ecuación de verosimilitud como:

$$\begin{aligned}
L(\boldsymbol{\beta},\sigma) &= \Pi_{y_i>0}f(y_i^*|\mathbf{x}_i)\Pi_{y_i=0}P(y_i^*\leq 0) \\
&= \Pi_{y_i>0}\frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)^2}\Pi_{y_i=0}\Phi\left(-\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \\
&= \Pi_{y_i>0}\frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)^2}\Pi_{y_i=0}\left(1-\Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right)
\end{aligned}$$

En su versión logaritmica, la ecuación de verosimilitud quedará como:

$$\begin{aligned} lnL(\boldsymbol{\beta},\sigma) &= \Sigma_{y_i>0}\left[-\frac{1}{2}ln(2\Pi\sigma^2)-\frac{1}{2}\left(\frac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)^2\right] \\ &\quad +\Sigma_{y_i=0}\left(1-\Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right) \end{aligned}$$

De esta última expresión podemos obtener las condicones de primer orden para determinar un estimador para $\boldsymbol{\theta}' = (\boldsymbol{\beta},\sigma)$, así como la matriz Hesiana (H) para determinar la variaza de esto estimadores.

$$\frac{\partial lnL(\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\theta}} = 0 \tag{9.10}$$

$$\frac{\partial^2 lnL(\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = H(\boldsymbol{\theta}) \tag{9.11}$$

$$Var(\hat{\boldsymbol{\theta}}) = [-\mathbb{E}[H(\boldsymbol{\theta})]]^{-1} \tag{9.12}$$

Otro caso similar de

10

Introducción a inferencia causal

10.1. Estimación de modelos para determinar efectos de tratamiento

10.1.1. Motivación del procedimiento Difference - in - Differences (DiD)

La técnica de Differences-in-Differences (DiD) es una de las más populares en ciencias sociales para estimar efectos causales en conjuntos no experimentales. Los diseños de DiD son una estrategia cuasi-experimental para estimar, entre otras cosas, efectos causales. DiD es la técnica más popular en la investigación en ciencias sociales cuantitativas. En general, todos los buenos diseños DiD están basados en algún tipo de experimento natural.

La historia es famosa y atribuida a John Snow (1855)–a continuación retomamos la narración del caso en el libro de Cunningham (2021) [Cun21]. Snow analizó las causas de una serie de olas de la pandemia de cólera en Londres en el siglo XIX. El colera es una enfermedad que ataca a sus victimas de forma repentina con síntomas como vómito y diarrea. En el siglo XIX el cólera en Londres era una enfermedad mortal y que, inicialmente, se creía que se transmitía de forma aeróbica.

Snow era un médico en la epoca en que Londres enfrentó 3 olas de pandemía de cólera que enfermó y mató a decenas de miles de personas y que, en su momento, se consideró como una enfermedad misteriosa.

En aquel momento en el siglo XIX, los doctores no podían ayudar a las victimas de la enfermedad, ya que no estaban equivocados sobre el mecanismo

que causaba la cólera y que facilitaba el esparcimiento de la infermedad. Los médicos creían que la cólera se esparcía mediante el aiete por causa de partículas microscópicas. De esta forma, la forma en que los médicos abordaron el problema fue poner en cuarentena a las personas infectadas o petencialmente infectadas con el objeto de disminuir la plaga. Algo que hoy día sabemos era ineficiente.

Snow, inicialmente, creía en la teoría de que la enfermedad se esparcía o transmitía de forma aeróbica. No obstante, con el paso del tiempo, Snow desarrolló una nueva teoría sobre la forma en que se transmitía el cólera, desechando la teoría de la transmisión aeróbica. Snow concluyó que la causa de la enfermedad era un microrganismo que ingresaba al cuerpo humano mediante el consumo de agua y alimentos.

Snow encontró que el principal medio de transmisión del cólera era el agua consumida por la población y que se multiplicaba por el hecho de que los desechos sanitarios se depositaban en la principal fuente de agua de la época, el rio Támesis.

La forma en que Snow detectó la fuente principal y mediante la cual construyó su teoría fue una serie de anécdotas similares a las siguientes. Por ejemplo, notó que la transmisión del cólera solía seguir los patrones de comercio. En un caso de la llegada de un barco mercante a un puerto que tenía reportes de infecciones notó que la tripulación enfermaba sólo después de desembarcar, ya que la que no lo hacía no solía enfermarse. En otro caso notó que dos construcciones que tenían fuentes de agua distintas (origen del líquido distinto) tenían patrones de enfermedad distinta, en una había notablemente menos infecciones que en la otra.

Estas observaciones no eran consistentes con la teoría de la transmisión aeróbica. Así, Snow continuó recabando información. Un experimento natural ocurrió cuando notó que había el caso en dos proveedores de agua que tomaban agua de puntos distintos del Támesis, uno despúes del punto de deshague del drenaje y otro más desde un punto anterior. Con esta información, Snow recabó información de los lugares que ran atendidos por las empresas y encontro que el número de casos de cólera por cada 10,000 habitantes era como se muestra en el Cuadro 10.1.

10.1.2. Estrategia de estimación

En la literatura se ha disctutido algunos principios básicos que se deben cumplir en el caso de data sets en los que se puede implementar el

Cuadro 10.1: Tabla XII de Snow (1854)–Casos de cólera en Londres por cada 10,000 habitantes según proveedor de agua, retomado de Cunningham(2021, p. 472) [Cun21]

Nombre de la compañía	1849	1854
Southwark and Vauxhall	135	147
Lambert	85	19

DiD.[Rot+23] A estos principios se les puede denominar como el modelo canónico del DiD, donde solo existen dos periodos de tiempo disponibles, existe una población tratada por única vez a partir de que inicia el segundo periodo.

Otra forma de verlo es responder a estas preguntas:

1. ¿Todas las observaciones son tratadas al mismo tiemo?

2. ¿Está seguro de que se puede suponer que las tendencias entre todos los individuos son paralelas?

3. ¿Son suficientemente grandes los conjuntos de datos tratados y no tratados?

En esta sección describimos la forma en que estimaremos el modelo de DiD. La metodología DiD es una técnica utilizada en Econometría para medir el efecto sobre una variable dependiente o de respuesta cuantitativa y debido a un tratamiento o cambio, en un determinado período en el tiempo. Para tal efecto estimamos una ecuación como la siguiente:

$$y_{it} = \beta_1 + \beta_2 T_t + \beta_3 D_t + \theta T_t \times D_t + \boldsymbol{\gamma} \mathbf{X}_{it} + \varepsilon_{it} \tag{10.1}$$

Donde y_{it} es la variable sobre la cual queremos evaluar el efecto de un tratamiento, $\mathbf{X}_{it}$ es un conjunto de variables de control, que incluye a las variables de efectos fijos (temporales, de individuos o del tipo que se decida) y, en su caso, una tendencia, entre otras. Finalmente, las variables T_t y D_i son variables dicotómicas que identifican con el valor de 1 el momento a partir del cual se observa el tratamiento y con 0 en cualquier otro caso, y con 1 los individuos que fueron tratados y con 0 en cualquier otro caso, respectivamente.

De esta forma, el producto $T_t \times D_t$ sería una variable dicotómica que indicaría con un 1 a aquellos individuos que fueron tratados a partir del momento en que se implmenento el tratamiento y con 0 en cualquier otro caso. Bajo este escenario, θ es un coeficiente que captura el efecto del tratamiento, condicional a que se ha controlado por una serie de factores.

Otra forma de ver lo que se indica en la ecuación (10.1) es como la comparación de la diferencia en diferencia descrita por:

$$\begin{aligned} E &= [(\overline{y}_{exit}|treatment) - (\overline{y}_{baseline}|treatment)] \\ &\quad -[(\overline{y}_{exit}|placebo) - (\overline{y}_{baseline}|placebo)] \end{aligned} \tag{10.2}$$

La expresión (10.2) cuantifica el eecto que tiene el tratamiento como la diferencia de medias de dos grupos: un grupo tratado y otro no tratado. Dentro de los cuales se ha comparado las medias de respecto de una línea base. Podemos plantear esto desde esta otra perspectiva. Pensemos que solo tenemos dos periodos $t = 1, 2$, en el cual en 1 no ha ocurrido el tratamiento y en 2 ya ha ocurrido. Así, el cambio en la varaible de respuesta para los individuos tratados será (suponiendo que omitimos la matriz $\mathbf{X}_{it}$ en la ecuación (10.1)):

$$\begin{aligned} \mathbb{E}[y_{i2}|D_i = 1] - \mathbb{E}[y_{i1}|D_i = 1] &= (\beta_1 + \beta_2 + \beta_3 + \theta) - (\beta_1 + \beta_3 + \theta) \\ &= \beta_2 + \theta \end{aligned} \tag{10.3}$$

Ahora, sobre los no tratados:

$$\begin{aligned} \mathbb{E}[y_{i2}|D_i = 0] - \mathbb{E}[y_{i1}|D_i = 0] &= (\beta_1 + \beta_2) - (\beta_1) \\ &= \beta_2 \end{aligned} \tag{10.4}$$

Así, la diferencia en diferencia restulata de restar la ecuación (10.4) a la ecuación (10.3):

$$\begin{aligned} &(\mathbb{E}[y_{i2}|D_i = 1] - \mathbb{E}[y_{i1}|D_i = 1]) - (\mathbb{E}[y_{i2}|D_i = 0] - \mathbb{E}[y_{i1}|D_i = 0]) \\ &= \theta \end{aligned} \tag{10.5}$$

10.2. Control Sintético

10.2.1. Motivación

En esta sección utilizaremos como base el trabajo de Abadie (2021) [Aba21]. Por lo anterior pondremos énfasis en la factibilidad, los requerimientos de los datos, los requerimientos de contexto y algunos problemas

metodológicos asociados a la palicación empírica del control sintético. Particularmente, nos enfocaremos en la importancia que tiene las características que hacen posible la implementación del control sintético.

Susan Athey y Guido Imbens en 2017 describieron al proceso de control sintéricos como la innovación más importante en la evaluación de políticas públicas en los últimos 15 años (Athey y Imbebs (2017) [AI17]). En los últimos años, el control sintético ha sido aplicado a diversos dominios: migración, conexiones políticas y cumplimiento de leyes, impuestos, crimen organizado y terrorismo, salario mínimo, entre otros. El uso de controles sintéticos no se límita a las ciencias sociales. También ha sido aplicado en ingeniería y ciencias biomédicas, por ejemplo.

Los métodos de control sintético fueron propuestos originalmente por Abadie y Gardeazabal (2003) y Abadie, Diamond y Hainmueller (2010) con el próposito de estimar los efectos de intervenciones agregadas –intervenciones que son implementadas a un nivle agregado afectando a un pequeño número de unidades de in conjunto amplio de unidades como ciudades, regiones o países– en algún resultado agregado que sea de interés.

Consideremos una situación en la que una unidad agregada, tal como un estado o municipio, es expluesto a un evento o intervención de interés. En estos casos, los análisis tradicionales de regresión requieren muestras grandes y muchas observaciones dentro del los casos afectados por el evento. En un caso similar, el uso de series de tiempo para la evaluación de una política pública requiere de observaciones por muchos períodos para realiar un análisis adecuadamente.

El método de controles sintéticos está basado en la idea de que, cuando las unidades de observación son un número pequeño de entidades agregadas, una caombinación de unidades no afectadas podrían proveer una comparación más apropiada que una simple unidad observada de forma individual.

10.2.2. Planteamiento del método de estimación

Supongamos que tenemos datos para $J+1$ unidades, es decir, observamos $j = 1, 2, \ldots, J+1$. Sin pérdida de generalidad, asumiremos que la primera unidad ($j = 1$) es la unidad que ha sido tratatada, esto es, la unidad afectada por la intervención de política pública. El "donor pool.º conjunto de donadores que se constituye pot el conjunto de potenciales unidades de comparación, $j = 2, \ldots, J+1$, es una colección de unidades no tratadas o no afectadas por la intervención.

Asumiremos que nuestros datos son de una duración de T periodos y que los pimeros T_0 periodos son previos a la intervención. Para cada unidad, j, y tiempo, t, observamos el resultado de interés, Y_{jt}. Para cada unidad, j, observamos un conjunto de k predictores del resultado observado, $X_{1j}, \ldots, X_{kj}$ el cual podría incluir los valores de la variable analizada, Y_{jt}, pero en periodos previos y los cuales son en sí mismos no afectados por la intervención.

Los $K \times 1$ vectores $\mathbf{X}_1, \ldots, \mathbf{X}_{J+1}$ contienen los valores de los predictores de las unidades $j = 1, \ldots, J+1$, respectivamente. La matriz $k \times J$, $\mathbf{X}_0 = [\mathbf{X}_2, \ldots, \mathbf{X}_{J+1}]$ contiene los valores de los predictores de las J unidades no correlacionadas. Para cada unidad j y periodo de tiempo t definimos Y_{jt}^N como la respuesta potencial que se hubiera observado sin la intervención. Para la unidad afectada por la intervención $j = 1$ y para el periodo postintervención $t > T_0$ definimos Y_{1t}^I como la respuesta potencial bajo la intervención.

De esta forma, el efecto de la intervención de interés para el periodo t afectado ($t > T_0$) estará dado por:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \tag{10.6}$$

De esta forma, el reto de la evaluación de política pública es estimar Y_{1t}^N para $t > T_0$. Esto es, construir el contrafactual.

Un control sintético es definido como el promedio ponderado de las unidades en el conjunto de donadores. Formalmente, un control sintérico puede ser representado por un vector $J \times 1$ de pesos $\mathbf{W} = (w_2, \times, w_{J+1})$. Dado el conjunto de pesos $\mathbf{W}$ los estimadores del control sintético de Y_{1t}^N y de τ_{1t} son:

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt} \tag{10.7}$$

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N \tag{10.8}$$

Para evitar un escenario en el que la estimación sea explosiva, restringimos los pesos a que no sean negativos y que su suma sea 1. Note que también se debe considerar que el control sintético requiere que las variables en el data set sean rescaladas correctamente para considerar las diferencias por las unidades (por ejemplo, ingreso per cápita) o si la corrección no es necesaria puesto que las variables en los datos no se requieren rescalar por tamaño (por ejemplo, precios).

De esta forma, el contro sintético se reduce a la pregunta de cómo deben ser los pesos $w_2, \ldots, w_{J+1}$. Dado un conjunto de constantes $v_1, \ldots, v_k$, podemos proponer la elección del control sintético $\mathbf{W}^* = (w_2^*, \ldots, w_{J+1}^*)'$ que minimiza:

$$||\mathbf{X}_1 - \mathbf{X}_0|| = \left(\sum_{h=1}^{k} v_h (X_{h1} - w_2 X_{h2} - \ldots - w_{J+1} X_{hJ+1})^2 \right)^{1/2} \tag{10.9}$$

Sujeto a que $w_j > 0$ y que $w_2 + \ldots + w_{J+1} = 1$.

De esta forma, la estimación del efecto derivado del tratamiento en la unidad tratada en el momento $t = T_0 + 1, \ldots, T$ es:

$$\hat{\tau} = Y_{1t} + \sum_{j=2}^{J+1} w_j^* Y_{jt} \tag{10.10}$$

Las constantes no negativas $\mathbf{V} = v_1, \ldots, v_k$ en la ecuación (10.9) reflejan la importancia relativa en el control sintético los valores de los k predictores para la unidad de tratamiento $X_{11}, X_{21}, \ldots, X_{k1}$. Así, la pregunta que queda por responder es cómo determinar $\mathbf{V}$. Una forma muy sencilla de determinar ese vecto es elegir cada v_h es la inversa de la varianza de $X_{h1}, \ldots, X_{hJ+1}$, $h = 1, \ldots, k$, lo que resulta en un efecto de rescalamiento de todos los renglones de $[\mathbf{X}_0 : \mathbf{X}_1]$ a una varianza unitaria.

10.2.3. ¿Por qué usar el control sintérico?

En principio consideremos que mediante el estimador de una refresión lineal es posible determinar el efecto del tratamiento facilmente mediante la construcción de un panel. Sea $\mathbf{Y}_0$ la matriz $(T - T_0) \times J$ de una resultado post-intervención para las unidades del conjunto de donadores. Sea $\overline{\mathbf{X}}_1$ y $\overline{\mathbf{X}}_0$ el resultado de agregar a las matrices $\mathbf{X}_1$ y $\mathbf{X}_0$, respectivamente, un vector de unos (1's). Para una matriz no singular $\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0'$, un estimador basado en una regresión del contrafactual Y_{1t}^N para $t > T_0$ es $\hat{\mathbf{B}}' \overline{\mathbf{X}}_1$, donde

$$\hat{\mathbf{B}} = (\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0')^{-1} \overline{\mathbf{X}}_0 \mathbf{Y}_0'$$

Lo cual es un estimador de regresión de $\overline{\mathbf{X}}_1$ en $\overline{\mathbf{X}}_0$.

También el control sintético evita los resultados explotativos, ya que los pesos en el control sintético son no negativos y suman 1.

Otra ventaja del control sintético es que existe transparecnia en el proceso de estimación y en la construcción del contrafactual.

En cuanto a los requerimientos, el control sintético requiere la disponibilidad de datos para las unidades tratadas y para las no tratadas.

Finalmente, respecto de las extenciones, múltiples estudios han discutido estimaciones e inferencia con controles sintéticos para casos donde existen múltiples unidades tratadas. En este caso se pueden crear problemas de estimación.

11

Estimación de modelos de duración

12

Introducción al Aprendizaje Estadístico

12.1. Motivación e introducción

El aprendizaje estadístico juega un rol essencial en muchas áreas de la ciencia, finanzas y la industtria. Algunos ejemplos son:

1. Predecir si un paciente–que se encuetra hospitalizado debido a un ataque al corazón–tenndrá un segundo ataque. La predicción estará basada en métricas demográficas, de la dieta y de regitros clínicos.

2. Predecir el precio de una acción en los siguienntes 6 meses; considerando la base de las medidas de desempeño de la compañía y de otros datos económicos.

3. Identificar los números en la digitalización de formas escritas a mano.

4. Identificar los factores de riesgo para el cáncer de próstata, basados en datos clínicos y de otras variables demográficas.

El aprendizaje estadístico comprende a un conjunto de herramientas para modelar y entender conjuntos de datos complejos. También se le conoce como Machine Learning (ML), el cual conjuga el desarrollo reciente en el área de la estadística junto con el crecimiento en paralelo de la computación.

El aprendizaje estadístico considera muchos métodos convencionales y de uso amplio como análisis de regresión, clasificación, árboles de decisión,

etc. También se refiere a una amplia gama de herramientas para entender o interpretar datos clasificadas como basadas en prueba y error, supervisadas y no supervisadas.

- Aprendizaje por refuerzo: se produce cuando una máquina aprende por medio de prueba y error hasta alcanzar la mejor manera de completar una tarea asignada.

- Aprendizaje supervisado: se produce cuando los algoritmos aprenden de datos previamente etiquetados y clasificados, es decir, cuenta con un conjunto muestra, con lo cual el algoritmo analiza varios elementos dados y los compara con su grupo muestra para etiquetarlos según esos criterios. Por ejemplo, puedes clasificar fotos con descripciones de los elementos que aparecen en ellas. El aprendizaje estadístico supervisado involucra la construcción de un modelo estadístico para predecir o estimar un resultado (o variable que se pede supervisar) basado en uno o más variables explicativas.

- Aprendizaje no supervisado: los algoritmos que trabajan bajo estos lineamientos son similares a los supervisados, solo que estos últimos adoptan un modelo predictivo, es decir, no tienen clasificaciones y etiquetas previas, sino que buscan ejemplos que se parezcan y puedan agrupar. El aprendizaje no supervisado considera variables explicativas, pero los resultados observados no son una variable explicada que sea suceptible de supervisión.

Ahora introduzcamos un poco de notación. Denotaremos a una variable independiente o explicativa con x_{ij}, si dicha variable es un conjunto de variables acomodadas en un vector utilizaremos $\mathbf{X}_i$; en estos casos denotaremos a un elemento o variable del vector $\mathbf{X}_i$ como x_{ik}, donde $i = 1, 2, \ldots, n$ denota a los individuos en la muestra y $k = 1, 2, \ldots, K$ denota al n úmero de variables. Por convención diremos que $x_{i1} = 1$ para todo $i = 1, 2, \ldots, n$, ya que en dicha variable consideraremos al término constante en la regresión.

Por su parte, los resultados, variables dependientes o variables de respuesta se denotarán como:

- y_i denotará una respuesta que es una cantidad continua

- g_i denotará una respuesta cualitativa, discreta o de grupo

Así, con $\mathbf{X}$, y $\mathbf{Y}$ y $\mathbf{G}$ denotaremos al a la matriz y vector columna que contiene a todos los valores de las variables dependientes y de respuesta apiladas para cada uno de los elementos en la muestra indexados con i.

Usaremos mayúsculas como X, Y o G para representar a los aspectos genéricos de las variables. De esta forma, debe ser claro que las letras minúsculas seran empleadas para representar a valores observados de las variables, así el valor observado de la variable k-ésima para el elemento de la muestra i-ésimo en $\mathbf{X}$ será representado como x_{ij}. Finalmente, con la notación $\hat{Y}$ o $\hat{G}$ representaremos a los valores estimados o predecidos.

El aprendizaje estadístico parte del establecimiento de que una variable dependiente Y es una función de un conjunto de variables explicativas $\mathbf{X} = [X_1, X_2, \ldots, X_p]$. De esta forma platearemos:

$$Y = f(\mathbf{X}) + \varepsilon \tag{12.1}$$

Así, el aprendizaje estadístico se trata de un conjunto de aproximaciones para f. ¿Por qué estimar f? La respuesta es por predicción y por inferencia. La más común de ambas razones es la predicción, con una predicción de Y podríamos establecer:

$$\hat{Y} = \hat{f}(\mathbf{X}) \tag{12.2}$$

12.2. Modelos linealles y el procedimiento de mínimos cuadrados

Los modelos lineales han sido utilizados de forma recurrente en estadística en por lo menos los últimos 30 años y sin duda se mantendrá como una de las herramientas más utilizadas.

Sean un conjunto de observaciones indexadas por $i = 1, 2, \ldots, n$ y un vector de variables independientes o explicativas X descrita de la forma:

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$$

Así, podemos describir a la variable Y (o y_i para una observación especifica) con una ecuación:

$$\hat{y}_i = \mathbf{X}_i \boldsymbol{\beta}$$

Donde $i = 1, 2, \ldots, n$ y el vector $\boldsymbol{\beta}$ incluye a un término constante.

Visto de esta forma, a partir de la función lineal $f(X) = \mathbf{X}'\boldsymbol{\beta}$ podemos determinar el vector gradiente $f'(X) = \boldsymbol{\beta}$, que es la ruta por la que la función $f(X)$ crece más rápidamente.

¿cómo estimamos a $\boldsymbol{\beta}$?, con un procedimiento de mínimos cuadrados visto en la sección previa:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

De igual forma, son validos los análisis de pruebas de hipótesis t y F descritos anteriormente.

12.3. Método de regresiones restringidas

Estte tipo de métodos permite restringir los posibles valores de las estimacionnes del parámetro $\boldsymbol{\beta}$ a un subconjunto seleccionado. Veámos dos casos particulaes i) una regresión de cresta destringida (ridge) y ii) la regresión Lasso.

12.3.1. Regresión Ridge

La regresión de ridge restringe los coeficientes de la regreesión mediante la imposición de una penalización en su magnitud. Los coeficientes esttimados por este método resultan de resolvel el problema:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda \sum_{k=2}^{K} \beta_k^2 \right] \tag{12.3}$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión.

La ideea de la penalización de los paramétros se deriva de la forma equivalente del problema en la ecuación (12.3) es:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right]$$

Sujeto a:

$$\sum_{k=2}^{K} \beta_k^2 \leq t$$

Note que la penalización no aplica al término constante, ¿por qué?, para garentizar que la estimación del hiperplano asociado pasa por la media de Y y no por el cero (0). ¿Qué implicaciones tiene para la estimación?

El término constante se estima considerando que este toma el valor de la media de Y dado por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Para los restantes coeficientes los determinaremos mediante un procedimiento dado por:

$$\min_{\boldsymbol{\beta}_R, \lambda} [(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R) + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R] \tag{12.4}$$

Donde $\boldsymbol{\beta}_R$ contiene sólo las pendientes. Pare determinar un valor estimado debemos resolver el problema descrito en la ecuación (12.4):

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_R} S(\boldsymbol{\beta}_R) &= \frac{\partial}{\partial \boldsymbol{\beta}_R}((\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R) + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}_R}(\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}_R + \boldsymbol{\beta}_R'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_R + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R) \\
&= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_R + 2\lambda \boldsymbol{\beta}_R
\end{aligned}$$

Determinando el mínimo:

$$\begin{aligned}
-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_R^{Ridge} + 2\lambda \hat{\boldsymbol{\beta}}_R^{Ridge} &= 0 \\
-\mathbf{X}'\mathbf{Y} + (\mathbf{X}'\mathbf{X} + \lambda \mathbb{I}_{K-1})\hat{\boldsymbol{\beta}}_R^{Ridge} &= 0
\end{aligned}$$

Por lo tanto, el estimador estará dado por:

$$\hat{\boldsymbol{\beta}}_R^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbb{I}_{K-1})^{-1}\mathbf{X}'\mathbf{Y} \tag{12.5}$$

Así, el problema de la estimación es que tiene 1 grado de libertad, λ, que es un valor de penalización y que resultta arbitrario.

12.3.2. Regresión Lasso (Least Absolute Shrinkage and Selection Operator)

La regresión Lasso se define por la solución al problema:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 \right] \tag{12.6}$$

Sujeto a:

$$\sum_{k=2}^{K} |\beta_k| \leq t$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión. El término constante se estima considerando que este toma el valor de la media de Y dado por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

La ideea de la penalización de los paramétros se deriva de la forma equivalente del problema en la ecuación (12.6) es:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 + \lambda \sum_{k=2}^{K} |\beta_k| \right] \tag{12.7}$$

12.4. Modelos lineales de clasificación

A continuación, analizamos el problema de clasificación y nos enfocamos en métodos lineales para casificación. Supongamos que nuestro predictor está dado por una función $G(X)$ que toma valores en un conjunto de clases o tipos indexado por $j = \{0, 1, 2, \ldots, J\}$. Este tipo de modelos se les conoce como discriminante por probabilidades:

$$\mathbb{P}(G = j | X = x) \tag{12.8}$$

De esta forma, supongamos un problema de determinar la probabilidad de dos clases $j = \{1, 2\}$, para cada un tendríamos:

$$\begin{aligned} \mathbb{P}(G = 1 | X = x) &= \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \\ \mathbb{P}(G = 2 | X = x) &= \frac{1}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \end{aligned}$$

De esta forma garantizamos que $1 = \mathbb{P}(G = 1|X = x) + \mathbb{P}(G = 2|X = x)$. Así, podemos tomar una transformación *Logit* de $ln(\theta/(1-\theta))$, ya que estamos ante el caso de un modelo binomial. En nuestro caso concreto tendríamos el caso:

$$ln \left\{ \frac{\mathbb{P}(G = 1|X = x)}{\mathbb{P}(G = 2|X = x)} \right\} = \mathbf{X}\boldsymbol{\beta} \tag{12.9}$$

A este modelo se le conoce como modelo *Logit* o regresión logística.

Regresión Logística

La regresión logística aparece por la necesidad de un modelo para determinar las probabilidades de ocurrencia de $J+1$ clases en función de una relación lineal de X y al mismo tiempo garantizar que la estimación respetará que la estimación se ubique en el intervalo $[0,1]$.

Supongamos un matriz $\mathbf{X}$ que contiene información de las variables explicativas y en su primera conlumna incluye al término constante. Consideremos un caso generalizado en el que la clases o valores que puede tomar $G(X)$ son $j = 0, 1, 2, \ldots, J$. Considerando esto, el modelo a continuación plateada será conocido como el modelo multinomial *logit*, para lo cual las probabilidades para cada clase estarán dadas por:

$$\mathbb{P}(G = j|\mathbf{X} = \mathbf{x}) = \frac{e^{\mathbf{X}\boldsymbol{\beta}_j}}{1 + \sum_{h=1}^{J} e^{\mathbf{X}\boldsymbol{\beta}_h}} \tag{12.10}$$

Donde $j = 1, 2, \ldots, J$ y $\boldsymbol{\beta}_j$ es un vector de dimensión $K \times 1$. Por su parte la probabilidad de el caso en que la clase es 0, de forma complementaria tendríamos:

$$\mathbb{P}(G = 0|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + \sum_{h=1}^{J} e^{\mathbf{X}\boldsymbol{\beta}_h}} \tag{12.11}$$

Estimación

Los modelos de regresión logítstica soos estimados usualmente por métodos de Máxima Verosimilitud mediante el uso de $\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)$, para $j = 0, 1, \ldots, J$ e $i = 1, 2, \ldots, n$. Así, plateamos la función de verosimilitud como:

$$L(\theta) = \Pi_{i=1}^{n} \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \tag{12.12}$$

Utilizando una transformación logarítmica:

$$lnL(\theta) = l(\theta) = \sum_{i=1}^{n} ln\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \tag{12.13}$$

Por simplicidad continuaremos con el caso dee una estimación para un modelo de respuesta binomial en el cual $G = \{0, 1\}$. De esta forma establez-

camos la función de verosimilitud como:

$$\begin{aligned} lnL(\theta) = l(\theta) &= \sum_{i=1}^{n} G_i ln\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \\ &\quad +(1 - G_i)ln(1 - \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)) \\ &= \sum_{i=1}^{n} G_i \mathbf{X}_i \boldsymbol{\beta} - ln(1 + e^{\mathbf{X}_i \boldsymbol{\beta}}) \end{aligned}$$

Las condiciones de primer orden son:

$$\frac{\partial l(\theta)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(G_i \mathbf{X}_i - \mathbf{X}_i \frac{e^{\mathbf{X}_i \boldsymbol{\beta}}}{1 + \mathbf{X}_i \boldsymbol{\beta}} \right) = (G_i - \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i))\mathbf{X}_i \quad (12.14)$$

Así, la solucción estará dada por aquellas $\boldsymbol{\beta}$ que satisfacen:

$$\sum_{i=1}^{n} \left(G_i - \frac{e^{\mathbf{X}_i \boldsymbol{\beta}}}{1 + \mathbf{X}_i \boldsymbol{\beta}} \right) \mathbf{X}_i = 0 \quad (12.15)$$

Ajuste y separación del conjunto de datos

Supomgamos la variable objetivo Y, un vector de variables explicativas o variables 'input' X y un modelo predictivo $\hat{f}(X)$ que es estimado a partir de un conjunto de entrenamiento τ.

Definiremos la función de pérdida derivada de la estimación y capturada por los errores entre Y y $\hat{f}(X)$ estará dada por:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{error cuadrático} \\ |Y - \hat{f}(X)| & \text{error absoluto} \end{cases} \quad (12.16)$$

De esta forma podemos establecer un error cuadrático como:

$$\begin{aligned} \mathbb{E}[Y - \hat{Y}]^2 &= \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \mathbb{E}[f(X) - \hat{f}(X)]^2 + Var(\varepsilon) \end{aligned}$$

Donde $\mathbb{E}[f(X) - \hat{f}(X)]^2$ es el único componente reducible. Para hacer predicciones requerimos de un connjunto de datos de entrenamiento y otro más de prueba – en el primero estimamos $f(\cdot)$ y enn el segundo hacemos predicciones –.

Finalmente, la condición de inferencia nos permite construir pruebas de hipótesis, estimadores que cumple ciertas propiedades, intervalos de confianza, etc. Así, dividiremos al conjunto de datos conforme se describe en la Figura 12.1.

Figura 12.1: División del conjunto de datos, retomado de Hastie, Tibshirani, y Friedman (2017, p. 222) [HTF17]

12.4.1. Evaluando los modelos de aprendizaje supervisado

Matriz de confusión

Una vez que se entrena un problema de aprendizaje automatizado supervisado en un conjunto de datos históricos, se prueba el modelo obtenido mediante el uso de datos del conjunto de entrenamiento. De esta forma, es posible comparar las predicciones del modelo entrenado con los valores reales de la variable sujeta de análisis. La matriz de confusión proporciona un medio para evaluar el éxito de un problema de clasificación y dónde se cometen errores (es decir, dónde se vuelve 'confuso').

En el Cuadro 12.1 se muestra un ejemplo de la forma en que se suele mostrar una matriz de confusión.

		Predicciones	
		Positivas (1)	Negativas (0)
Real	Positivas (1)	True Positive (TP)	False Negative (FN)
	Negativas (0)	False Positive (FP)	True Negative (TN)

Cuadro 12.1: Matriz de Confusión

Con base en la matriz de confusión en el Cuadro 12.1, se pueden construir métricas que se calculan de la siguiente manera:

$$Precision = \frac{TP}{TP + FP} \tag{12.17}$$

$$Recuperacion = \frac{TP}{TP + FN} \tag{12.18}$$

$$F1 = 2 \times \frac{Precision \times Recuperacion}{Precision + Recuperacion} \tag{12.19}$$

La medida $F1$ o $F1 - score$ combina precisión y recuperación mediante la media armónica de los dos valores. De esta forma, siempre se ubicará entre ambos indicadores.

12.5. Aprendizaje no supervisado

Bibliografía

[Aba21] Alberto Abadie. «Using synthetic controls: Feasibility, data requirements, and metodological acpects». En: *Journal of Econometric Literature* (2021).

[Ada20] Christopher P. Adams. *Learning Microeconometrics with R*. Estados Unidos: CRC Press, 2020.

[AI17] Susan Athey y Guido W. Imbens. «The State of Applied Econometrics: Causality and Policy Evaluation». En: *Journal of Economic Perspectives* 2 (2017), págs. 3-32.

[CT05] Colin Cameron y Pravin K. Trivedi. *Microeconometrics: Methods and Applications*. Estados Unidos: Cambridge University Press, 2005.

[Cun21] Scott Cunningham. *Causal inference: The mixtape*. Estados Unidos: Yale University Press, 2021.

[Gre12] William Greene. *Econometric Analysis*. Estados Unidos: Prentice Hall, 2012.

[HTF17] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. LLC: Springer, 2017.

[Jam+13] Gareth James y col. *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2013.

[Rot+23] Jonathan Roth y col. «What's trending in difference-in-differences? A synthesis of the recent econometrics literature». En: *Journal of Econometrics* (2023).

[Woo10] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Estados Unidos: The MIT Pres, 2010.

Apéndice A

Algunos resultados importantes