

Econometría

Marcos Bujosa

03/11/2023

Puede encontrar la última versión de este material en

<https://github.com/mbujosab/Ectr>



Marcos Bujosa. Copyright © 2008–2023

Algunos derechos reservados. Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-sa/4.0/> o envie una carta a Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Índice

I Regresión Lineal (estadística descriptiva)	2
LECCIÓN 1: Geometría del ajuste MCO	3
1 Ajuste por Mínimos Cuadrados Ordinarios (MCO)	3
1.1 Introducción	3
1.2 Ajuste mínimo cuadrático	4
1.2.1 Una nota sobre la linealidad	5
1.2.2 Recordatorio sobre longitud, perpendicularidad y el Teorema de Pitágoras en \mathbb{R}^n	6
1.2.3 Criterio del ajuste mínimo cuadrático	7
1.2.4 Las ecuaciones normales	9
1.2.5 Ausencia de multicolinealidad exacta (regresores linealmente independientes)	11
1.2.6 ¡Ojo! El lenguaje habitual en econometría se toma ciertas licencias	12
LECCIÓN 2: Modelos con 1, 2 y 3 regresores	13
2 Espacios euclídeos	13
2.1 Productos escalares en general	13
2.1.1 Propiedades de los productos escalares	13
2.1.2 Los productos escalares sirven para medir longitudes y ángulos.	13
2.2 Producto escalar usual en el espacio euclídeo \mathbb{R}^N	14
2.2.1 Descomposición de los semi-productos escalares	14
2.3 El producto escalar de uso más común en estadística	15
3 Ajuste MCO con uno, dos, tres ó k regresores	16
3.1 Una constante como único regresor (media, desviación típica y varianza)	17
3.1.1 De la geometría a la interpretación estadística	21
3.2 Modelo Lineal Simple: ajuste MCO con una constante más un segundo regresor	22
3.3 Ajuste MCO con una constante y otros dos regresores adicionales	25
3.4 Ecuaciones normales del ajuste MCO con k regresores	25
Problemas de la Lección 2	26
LECCIÓN 3: Propiedades algebraicas del ajuste MCO. Medidas de ajuste	27

4 Propiedades algebraicas del ajuste MCO	27
4.1 Sumas de cuadrados y descomposición de la varianza	28
4.1.1 La norma usual y las “sumas de cuadrados”	28
4.1.2 La norma de la estadística y la descomposición de la varianza	30
4.2 Medidas de ajuste	31
4.2.1 Coeficiente de determinación o R^2	32
<i>Problemas de la Lección 3</i>	36
II Modelo Clásico de Regresión Lineal	37
LECCIÓN 4: Variables aleatorias y momentos condicionados	38
5 Las variables aleatorias como modelo.	38
5.1 Relación con las lecciones anteriores (Parte I).	39
6 Definición de Espacio Euclídeo de Probabilidad	40
6.1 Esperanza matemática	42
6.2 Subespacios probabilísticos	42
7 Esperanza condicional	43
7.1 Momentos de variables aleatorias.	44
7.2 Momentos condicionados.	46
7.3 Relación con las lecciones anteriores (Parte II).	49
7.3.1 (\mathbb{R}^N, s) es un Espacio Euclídeo Probabilístico	49
7.3.2 La regresión o ajuste MCO es un caso particular de esperanza condicional	49
7.3.3 Y si \mathbb{R}^N es un subespacio de probabilidad ¿dónde está el azar?	50
7.4 Diferencias con las lecciones previas.	51
7.5 La regresión como descomposición ortogonal	52
<i>Problemas de la Lección 4</i>	54
8 Apéndice: interpretación geométrica de la Probabilidad	54
8.1 Interpretación geométrica en un modelo para el lanzamiento de una moneda	56
8.2 “Clases de equivalencia” de variables aleatorias y la esperanza condicional	58
8.2.1 “Clases de equivalencia” de variables aleatorias	58
8.2.2 La esperanza condicional	59
LECCIÓN 5: Especificación y Estimación del Modelo Lineal General	61
9 Modelo Clásico de Regresión Lineal	61
9.1 Los cuatro primeros supuestos en el Modelo Clásico de Regresión Lineal	61
9.2 Primer supuesto	61
9.3 Segundo supuesto	62
9.4 Tercer supuesto	63
9.5 Cuarto supuesto	64
10 Regresión cuando \mathcal{E} es \mathbb{R}^N	65
10.1 Dos casos particulares de MLG	66
10.1.1 Modelo con una constante como único regresor	66
10.1.2 Modelo Lineal Simple	66
11 Estimación del Modelo Clásico de Regresión Lineal	67
11.1 Estimación de un modelo clásico de regresión con una muestra de datos	68
<i>Problemas de la Lección 5</i>	70
LECCIÓN 6: Propiedades estadísticas de los estimadores MCO	72
12 Espacio Euclídeo de Probabilidad para un Muestreo Aleatorio Simple	72
12.1 Momentos muestrales	74
12.1.1 Media muestral	74

12.1.2 La varianza muestral	74
12.1.3 La cuasi-varianza muestral	74
13 Propiedades estadísticas de los estimadores MCO	74
13.1 Estimador MCO de β	75
13.2 Esperanza del Estimador	76
13.3 Varianza del estimador	77
13.3.1 Consistencia del estimador MCO	79
13.4 Caso particular: la constante como único regresor	79
13.5 Caso particular: modelo lineal simple	79
13.6 Momentos de los valores ajustados y los errores	80
14 Distribución de los estimadores MCO bajo la hipótesis de Normalidad	81
14.1 Quinto supuesto del Modelo Clásico de Regresión Lineal	82
14.2 Estimación de la varianza residual y la matriz de covarianzas	82
14.3 Más sobre eficiencia de los estimadores	83
15 Más sobre medidas de ajuste	83
16 Apéndice de definiciones y resultados	84
<i>Problemas de la Lección 6</i>	86
III Inferencia en el Modelo Clásico de Regresión lineal	88
LECCIÓN 7: Inferencia. Contrastes de hipótesis lineales	89
17 Introducción a la contrastación de hipótesis	89
18 Estadístico t de Student	90
19 Contraste de hipótesis sobre coeficientes individuales de la regresión	91
19.1 Contrastes de dos colas	91
19.2 Contrastes de una sola cola	92
19.3 Reglas de decisión para contrastes de una cola y de dos colas usando el p -valor	93
<i>Problemas de la Lección 7</i>	94
LECCIÓN 8: Inferencia. Contrastes de hipótesis lineales (combinaciones lineales de parámetros). Intervalos y regiones de confianza	95
20 Contraste de hipótesis sobre combinaciones lineales de coeficientes de la regresión	95
20.1 El test F	96
20.2 t versus F	97
21 Regiones e intervalos de confianza	99
<i>Problemas de la Lección 8</i>	102
LECCIÓN 9: Mínimos cuadrados restringidos y contrastes de hipótesis lineales	103
22 Estimación bajo restricciones lineales generales	103
22.1 Propiedades estadísticas del estimador MCR	105
22.2 Contraste de la F mediante sumas residuales	105
22.2.1 Test de Chow	107
23 Contraste de normalidad Jarque-Bera	108
24 Apéndice: Demostraciones	109
24.1 Demostración de la Proposición 22.1 (<i>minimización de residuos sujeto a restricción</i>)	109
24.2 Demostración de la Proposición 22.3 (<i>varianza del estimador restringido</i>)	110
<i>Problemas de la Lección 9</i>	111

IV Interpretación	112
LECCIÓN 10: Interpretación de coeficientes en modelos con logaritmos	113
25 Interpretación de los parámetros en un modelo de regresión	113
25.1 Interpretación “ <i>Ceteris páribus</i> ”	113
25.2 Regresor y variable explicativa no son siempre lo mismo	113
26 Función exponencial, función logaritmo y elasticidad	114
27 Modelos con logaritmos	117
27.1 Relaciones lineales en las variables	118
<i>Cálculo de la elasticidad en un modelo Lin-Lin</i>	118
27.2 Relaciones Lin-Log	119
<i>Cálculo de la elasticidad en un modelo Lin-Log</i>	119
27.3 Relaciones semi-logarítmicas (Log-lineal)	120
<i>Interpretación de coeficientes en un modelo Log-Lin</i>	121
<i>Comparación del ajuste entre un modelo Lin-Lin y otro Log-Lin</i>	122
27.4 Modelos Log-Log	122
<i>Elasticidades en la demanda del transporte en autobús</i>	122
LECCIÓN 11: Interpretación de coeficientes en modelos con regresores cualitativos	124
28 Dummies	124
28.1 Interpretación de los coeficientes de las variables ficticias	124
<i>Diferencias salariales entre hombres y mujeres</i>	125
<i>Diferencias salariales entre hombres y mujeres (cont.)</i>	126
28.2 Contrastes de homogeneidad	127
28.2.1 Más contrastes de homogeneidad: uso dummies para contrastar cambios estructurales	128
28.3 Términos de interacción	131
<i>Cambio estructural en la participación de las mujeres en el mercado laboral</i>	132
<i>Modelo Log-lin con variables ficticias: diferencias salariales entre grupos.</i>	132
<i>Precio de viviendas unifamiliares.</i>	133
LECCIÓN 12: Multicolinealidad	134
29 Multicolinealidad	134
<i>Construcción de vivienda nueva</i>	136
<i>Construcción de vivienda nueva (transformación a datos per-cápita)</i>	136
<i>Gastos de mantenimiento de los automóviles</i>	137
<i>Ortogonalización de regresores mediante regresiones auxiliares</i>	137
<i>Pobreza y sus determinantes (omisión de variables)</i>	138
LECCIÓN 13: Errores de especificación	139
30 Incorrecta especificación por omisión de variables explicativas	139
30.1 Esperanza del estimador	139
30.2 Varianza del estimador	140
31 Incorrecta especificación por inclusión errónea de variables	142
<i>Problemas de la Lección 13</i>	144
Soluciones	146

En estas notas encontrará todas las transparencias de clase junto con más información.

El desarrollo de la mayoría de los resultados matemáticos está propuesto en forma de problemas al final de cada lección. Creo que aligerando los detalles se logra una exposición más directa, pero es importante que en una segunda lectura intente resolver los problemas. Estos problemas de contenido matemático están indicados en el margen del

texto (como por ejemplo a la izquierda de este párrafo). El número que sigue a la *P* es el número del problema (además, debajo y entre paréntesis se indica la página donde aparece el enunciado).

Al final de las notas encontrará las soluciones, pero no debe mirarlas hasta que haya dado con “su respuesta”. Consultar las respuestas de otro sin haber resuelto antes el ejercicio por uno mismo sirve de muy poco. Recuerde que el aprendizaje es una tarea activa, es decir, usted debe encontrar la solución activamente (individualmente o en “tándem” con algún compañero). Nunca debe limitarse a mirar la solución proporcionada por otro.

Naturaleza y objetivos de la econometría

Sin duda ésta es una sección importante e interesante, pero pude prepararla sin ayuda del profesor, así que ¡léase alguno de los capítulos introductorios de las referencias de la bibliografía recomendada! Por ejemplo Wooldridge (2006, Capítulo 1) o Gujarati (2003, Introducción y Capítulo 1).

Las ideas centrales son:

(Lección 0)

T-1 Introducción: ¿Por qué modelar?

Modelado consiste en intentar ajustar un modelo matemático (estadístico) a un conjunto de datos (“la muestra”). El modelo es útil cuando (pese a ser *simple*) *capta las características* de los datos que consideramos más interesantes. Los objetivos por los que se construyen modelos son variados:

F1

(Lección 0)

T-2 Algunos ejemplos

- **Estimación:**

sensibilidad de un valor financiero a movimientos de un índice de referencia (evaluación de exposición al riesgo y cobertura con derivados sobre el índice)

- **Previsiones:**

probabilidad de impago de préstamos (función de las características de la operación y del solicitante)

- **Simulación:**

rendimiento de una cartera de valores en diferentes escenarios

- **Control:**

bancos centrales: intervención de tipos para controlar la inflación

F2

Part I

Regresión Lineal (estadística descriptiva)

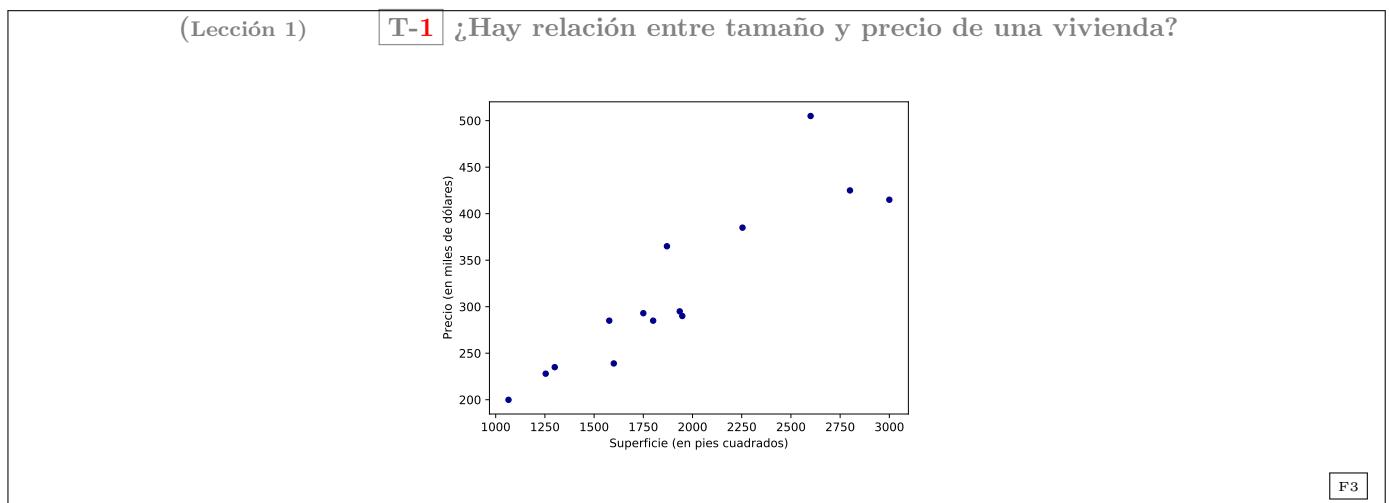
LECCIÓN 1: Geometría del ajuste MCO

1 Ajuste por Mínimos Cuadrados Ordinarios (MCO)

- Capítulos 2 y 3 de Wooldridge (2006)
- Apéndice E1 de Wooldridge (2006)

1.1 Introducción

En ocasiones se logra una mejor comprensión de una variable si se relaciona con otras.



Observando la representación de precios y tamaños de 14 casas unifamiliares del Campus de San Diego en California (de un ejemplo sacado del libro de Ramanathan (2002)) puede pensarse que los pares de datos (*precio, superficie*) están dispuestos de manera “*aproximada*” a lo largo de una recta

$$y = a + bx,$$

y que, consecuentemente, podríamos describir su relación como

$$\text{Precio} = a + b(\text{Superficie}) + \text{OtrasCosas};$$

donde *OtrasCosas* es lo que desplaza los puntos por encima y por debajo de esa hipotética recta.

La cuestión en esta primera lección es: ¿podemos encontrar “*la recta que mejor se ajusta*” a esta nube de puntos? es decir ¿podemos “*aproximar*” los datos de *precios* con una recta de la forma

$$\widetilde{\text{precio}} = \tilde{a} \mathbf{1} + \tilde{b} \mathbf{superficie}$$

donde *superficie* es el vector con los datos de superficie y $\mathbf{1}$ es un vector cuyas componentes son todas iguales a 1? ¿Y cómo calcular las magnitudes de los parámetros \tilde{a} y \tilde{b} ? es decir, ¿cuáles son los valores de la constante \tilde{a} y de la pendiente \tilde{b} de esta aproximación lineal? Antes de continuar, recordemos el concepto de combinación lineal:

Definición 1 (Combinación lineal de vectores de \mathbb{R}^n). *Sean los vectores $\mathbf{b}_1, \dots, \mathbf{b}_n$. Llamamos combinación lineal a cualquier suma de múltiplos de dichos vectores:*

$$a_1 \mathbf{b}_1 + a_2 \mathbf{b}_2 + \cdots + a_n \mathbf{b}_n$$

donde los números “ a_i ” son los *coeficientes* de la combinación lineal.

Por tanto, podemos reformular la pregunta de más arriba del siguiente modo: ¿podemos “*aproximar*” los *precios* de esas 14 casas mediante una combinación lineal de los vectores *superficie* y $\mathbf{1}$?

1.2 Ajuste mínimo cuadrático

Ejemplo: Función de consumo

Suponga que consumo (*con*) y renta disponible (*rd*) de las familias siguen la relación:

$$con = \beta_1 + \beta_2 rd + otrascosas$$

donde *otrascosas* son otros aspectos distintos de la renta (activos financieros, edad, lugar de residencia, etc.).

Disponiendo datos de *consumo* y *renta disp.* de N familias como vectores de \mathbb{R}^N , podemos construir una aproximación (*con*) del consumo con una combinación lineal de la renta disponible (*rd*) y de un término cte. (1) (ignorando las *otrascosas*):

$$\widetilde{con} = \widetilde{\beta}_1 \mathbf{1} + \widetilde{\beta}_2 \mathbf{rd} = [\mathbf{1}; \mathbf{rd}] \begin{pmatrix} \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \end{pmatrix}.$$

Nomenclatura

- *regresando*: vector de datos de *consumo* (*con*)
- *regresores*: vector de unos ($\mathbf{1}$) y de rentas disp. (*rd*): $\mathbf{X} = [\mathbf{1}; \mathbf{rd}]$ donde $\mathbf{X}_{|1} = \mathbf{1}$ y $\mathbf{X}_{|2} = \mathbf{rd}$.
- *vector de parámetros*: $\widetilde{\beta} = \begin{pmatrix} \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \end{pmatrix}$

Otro ejemplo: Un modelo para los salarios

salario = $\beta_1 + \beta_2 \text{educ} + \beta_3 \text{exper} + \beta_4 \text{IQ} + \text{otrascosas}$;

(disponiendo de datos de N trabajadores) el **ajuste** es

$$\widetilde{\text{salario}} = \widetilde{\beta}_1 \mathbf{1} + \widetilde{\beta}_2 \text{educ} + \widetilde{\beta}_3 \text{exper} + \widetilde{\beta}_4 \text{iq}$$

donde *educ* son los años de formación del trabajador, *exper* son sus años de experiencia laboral e *IQ* es una medida de la habilidad del trabajador (por ejemplo el coeficiente intelectual).

(Lección 1)

T-2 Ajuste MCO: función lineal en los parámetros

La aproximación (o ajuste) $\tilde{\mathbf{y}}$ es una combinación lineal de los *regresores* $\mathbf{X}_{|j}$:

$$\begin{pmatrix} \widetilde{y}_1 \\ \vdots \\ \widetilde{y}_N \end{pmatrix} = \widetilde{\beta}_1 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \widetilde{\beta}_2 \begin{pmatrix} x_{12} \\ \vdots \\ x_{N2} \end{pmatrix} + \cdots + \widetilde{\beta}_k \begin{pmatrix} x_{1k} \\ \vdots \\ x_{Nk} \end{pmatrix}$$

ó

$$\begin{aligned} \tilde{\mathbf{y}} &= \widetilde{\beta}_1 \mathbf{1} + \widetilde{\beta}_2 \mathbf{X}_{|2} + \widetilde{\beta}_3 \mathbf{X}_{|3} + \cdots + \widetilde{\beta}_k \mathbf{X}_{|k} \\ &= [\mathbf{1}; \mathbf{X}_{|2}; \dots; \mathbf{X}_{|k}] \begin{pmatrix} \widetilde{\beta}_1 \\ \vdots \\ \widetilde{\beta}_k \end{pmatrix} = \mathbf{X} \widetilde{\beta}; \end{aligned}$$

$$\text{donde } \widetilde{\beta} = \begin{pmatrix} \widetilde{\beta}_1 \\ \vdots \\ \widetilde{\beta}_k \end{pmatrix}; \quad \mathbf{X} = [\mathbf{1}; \mathbf{X}_{|2}; \dots; \mathbf{X}_{|k}].$$

Así los valores ajustados son $\tilde{\mathbf{y}} = \mathbf{X} \widetilde{\beta} \in \mathbb{R}^N$

F5

1.2.1 Una nota sobre la linealidad

La Econometría tiene una importante componente algebraica. Para consultar sus dudas sobre Álgebra Lineal dispone de una enorme cantidad de excelentes referencias, pero aquí aludiré de manera recurrente al *Curso de Álgebra Lineal con notación asociativa y un módulo para Python* (Bujosa, 2022a), pues aquí usaré la misma notación que en dicho curso¹.

En cualquier caso, y para facilitar la lectura de estos apuntes, iré intercalando algunas definiciones o resultados del Álgebra Lineal que son importantes en un curso de Econometría... Y dado que el título de la transparencia anterior alude al concepto de *función lineal*,² recordemos la definición de *función lineal* y veamos una nota sobre el producto “matriz por vector” que aclarará el uso del término “*función lineal*” en el título de la citada transparencia.

Definición 2 (Función lineal). *Sean \mathcal{D} y \mathcal{V} dos espacios vectoriales. Decimos que la función $f: \mathcal{D} \rightarrow \mathcal{V}$ es lineal si satisface las siguientes propiedades:*

1. *Para todo $\vec{x}, \vec{y} \in \mathcal{D}$, $f(\vec{x} + \vec{y}) = f(\vec{x}) + f(\vec{y})$.*
2. *Para todo $\vec{x} \in \mathcal{D}$ y para todo $\alpha \in \mathbb{R}$, $f(\alpha \vec{x}) = \alpha f(\vec{x})$.*

Nota 1 (sobre la expresión matriz por vector $\mathbf{A}\mathbf{x}$). *Por una parte, el producto de una matriz \mathbf{A} de orden N por k por un vector \mathbf{x} de \mathbb{R}^k es el vector de \mathbb{R}^N que se obtiene tomando la combinación lineal de las k columnas de \mathbf{A} cuyos parámetros son los k elementos de \mathbf{x} . Es decir, si \mathbf{A} tiene N filas y k columnas, entonces*

$$\mathbf{A}\mathbf{x} = x_1(\mathbf{A}_{|1}) + x_2(\mathbf{A}_{|2}) + \cdots + x_k(\mathbf{A}_{|k}) \in \mathbb{R}^N.$$

Por otra parte, como $\mathbf{A}\mathbf{x}$ es lineal por la derecha, es decir, como $\mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y}$ y $\mathbf{A}(\alpha\mathbf{x}) = \alpha\mathbf{A}\mathbf{x}$, el producto $\mathbf{A}\mathbf{x}$ transforma linealmente el vector \mathbf{x} de \mathbb{R}^k en el vector $\mathbf{A}\mathbf{x}$ de \mathbb{R}^N .

Así pues, decir que el ajuste $\tilde{\mathbf{y}}$ es una transformación lineal en $\tilde{\beta}$ es lo mismo que decir que $\tilde{\mathbf{y}}$ es de la forma: “una matriz por” el vector $\tilde{\beta}$; y decir que es combinación lineal de los regresores $\mathbf{X}_{|j}$ es lo mismo que decir que es de la forma: la matriz \mathbf{X} “por un vector”. Consecuentemente, la expresión $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\beta}$ significa ambas cosas (ahora puede volver a releer la última transparencia y entenderá su título).³

Ejemplo 1. Precio de las viviendas: considere los datos del siguiente cuadro con Precios de venta y Superficie útil de 14 casas unifamiliares en *University City*. San Diego, California. Año 1990. (Ramanathan, 2002, pp. 78).

n	price (\mathbf{y})	sqft (\mathbf{x})	$\widehat{\text{price}} (\tilde{\mathbf{y}})$
1	199.9	1065	?
2	228.0	1254	?
3	235.0	1300	?
4	285.0	1577	?
5	239.0	1600	?
6	293.0	1750	?
7	285.0	1800	?
8	365.0	1870	?
9	295.0	1935	?
10	290.0	1948	?
11	385.0	2254	?
12	505.0	2600	?
13	425.0	2800	?
14	415.0	3000	?

Table 1: Precio (miles de dólares) y superficie (pies al cuadrado). Ramanathan (2002, pp. 78).

Si asumimos que el precio y se relaciona con la superficie x del siguiente modo:

$$y_n = a + b x_n + \text{otras cosas}_n,$$

¹Además el libro es de acceso libre desde [GitHub](#).

²Aunque su contenido solo dice explícitamente que el ajuste es una combinación lineal de los regresores.

³Nótese que la expresión $\mathbf{A}\mathbf{x}$ también es lineal por la izquierda, es decir, $(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}$ y $(\alpha\mathbf{A})\mathbf{x} = \alpha\mathbf{A}\mathbf{x}$. Esto quiere decir que $\mathbf{X}\tilde{\beta}$ también es una función *lineal en los regresores*, pero lo más relevante en el ajuste MCO es que es lineal en los parámetros $\tilde{\beta}$. La linealidad en los regresores cobrará interés más adelante, al interpretar los resultados de los modelos estimados por MCO.

donde *otrascosas* son otros factores que influyen en el precio: localización, mantenimiento, servicios, calidades, etc.; podemos “*aproximar*” el vector de precios, \mathbf{y} , con una combinación lineal de los regresores:

$$\tilde{\mathbf{y}} = \tilde{\beta}_1 \mathbf{1} + \tilde{\beta}_2 \mathbf{x} = [\mathbf{1}; \mathbf{x}] \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \mathbf{x} \tilde{\boldsymbol{\beta}}.$$

De esta manera,

$$\tilde{\mathbf{y}} = (\mathbf{X}_{|1})\tilde{\beta}_1 + (\mathbf{X}_{|2})\tilde{\beta}_2 = \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \tilde{y}_3 \\ \tilde{y}_4 \\ \tilde{y}_5 \\ \tilde{y}_6 \\ \tilde{y}_7 \\ \tilde{y}_8 \\ \tilde{y}_9 \\ \tilde{y}_{10} \\ \tilde{y}_{11} \\ \tilde{y}_{12} \\ \tilde{y}_{13} \\ \tilde{y}_{14} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tilde{\beta}_1 + \begin{pmatrix} 1065 \\ 1254 \\ 1300 \\ 1577 \\ 1600 \\ 1750 \\ 1800 \\ 1870 \\ 1935 \\ 1948 \\ 2254 \\ 2600 \\ 2800 \\ 3000 \end{pmatrix} \tilde{\beta}_2 = \begin{pmatrix} 1 & 1065 \\ 1 & 1254 \\ 1 & 1300 \\ 1 & 1577 \\ 1 & 1600 \\ 1 & 1750 \\ 1 & 1800 \\ 1 & 1870 \\ 1 & 1935 \\ 1 & 1948 \\ 1 & 2254 \\ 1 & 2600 \\ 1 & 2800 \\ 1 & 3000 \end{pmatrix} \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \mathbf{x} \tilde{\boldsymbol{\beta}};$$

así por ejemplo, el precio ajustado para el séptimo piso de la muestra (cuya superficie es de 1800 pies cuadrados) sería

$$\tilde{y}_7 = (1)\tilde{\beta}_1 + (1800)\tilde{\beta}_2 = (1, 1800,) \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = {}_7|\mathbf{x} \tilde{\boldsymbol{\beta}} = {}_7|\tilde{\mathbf{y}}.$$

La cuestión es:

¿qué criterio empleamos para elegir $\tilde{\beta}_1$ y $\tilde{\beta}_2$ en el ajuste $\tilde{\mathbf{y}} = \mathbf{x} \tilde{\boldsymbol{\beta}}$?

Antes de establecer el criterio de selección de los parámetros de ajuste, veamos un:

1.2.2 Recordatorio sobre longitud, perpendicularidad y el Teorema de Pitágoras en \mathbb{R}^n

Primero recordemos la definición de *producto punto*:⁴

Definición 3. El *producto punto* de dos vectores \mathbf{a} y \mathbf{b} de \mathbb{R}^n es⁵

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3 + \cdots + a_n b_n = \sum_{i=1}^n a_i b_i.$$

El producto punto nos dota de una métrica con la que definir la longitud o norma de un vector:⁶

Definición 4. La *longitud* (o *norma*) de un vector \mathbf{a} es la raíz cuadrada de $\mathbf{a} \cdot \mathbf{a}$:

$$\text{longitud de } \mathbf{a} = \|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}}.$$

Por ejemplo, la longitud de $\mathbf{x} = \begin{pmatrix} 6 \\ 0 \\ -2 \\ 3 \end{pmatrix}$ es $\sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\begin{pmatrix} 6 \\ 0 \\ -2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 6 \\ 0 \\ -2 \\ 3 \end{pmatrix}} = \sqrt{6^2 + 0^2 + (-2)^2 + 3^2} = \sqrt{49} = 7$.

El *producto punto* nos permite definir la ortogonalidad o perpendicularidad entre vectores de \mathbb{R}^n .

⁴que es el producto escalar usual en \mathbb{R}^N (aunque ya veremos más adelante que no es el producto escalar usado en estadística).

⁵En la mayoría de los libros lo denotan con $\mathbf{a}^\top \mathbf{b}$, pero aquí seguiré la notación del curso de álgebra citado más arriba.

⁶En la siguiente lección veremos que en estadística se emplea una forma de medir ligeramente distinta.

Definición 5. Decimos que \mathbf{a} y \mathbf{b} son ortogonales o perpendiculares ($\mathbf{a} \perp \mathbf{b}$) cuando $\mathbf{a} \cdot \mathbf{b} = 0$.

Y ello da pie a repasar un conocido teorema aplicable a la suma de dos vectores perpendiculares entre si:

Teorema 1.1 (Teorema de Pitágoras en \mathbb{R}^n). Sean $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$; entonces $\mathbf{x} \cdot \mathbf{y} = 0$ (son perpendiculares) si y solo si

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

Demostración. (Véase Bujosa, 2022a, Lección 11). □

Fíjese que \mathbf{x} e \mathbf{y} corresponden a los catetos y el vector suma ($\mathbf{x} + \mathbf{y}$) a la hipotenusa de un triángulo rectángulo.

1.2.3 Criterio del ajuste mínimo cuadrático

Ahora ya estamos en condiciones de establecer el criterio de ajuste, que consistirá en minimizar una distancia. Pero antes de enunciar el criterio necesitamos definir el *error de ajuste*:

(Lección 1) T-3 Error de ajuste

Dados \mathbf{X} e \mathbf{y} , el “*error de ajuste*” al emplear $\tilde{\beta}$ es

$$\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\beta} = \mathbf{y} - \tilde{\mathbf{y}};$$

Así, descomponemos los datos observados \mathbf{y} en: $\mathbf{y} = \tilde{\mathbf{y}} + \tilde{\mathbf{e}}$.

Llamamos “Suma de los Residuos al Cuadrado” del ajuste $\tilde{\mathbf{y}}$ a

$$SRC(\tilde{\beta}) \equiv \sum_{n=1}^N \tilde{e}_n^2 = \tilde{\mathbf{e}} \cdot \tilde{\mathbf{e}} = \|\tilde{\mathbf{e}}\|^2$$

es decir, al cuadrado de la longitud del vector $\tilde{\mathbf{e}} = (\mathbf{y} - \tilde{\mathbf{y}})$. F8

Consecuentemente, el error cometido por el ajuste para la observación n -enésima es

$$\tilde{e}_n = y_n - \tilde{y}_n = {}_n|\mathbf{y} - {}_n|\tilde{\mathbf{y}} = {}_n|(\mathbf{y} - \mathbf{X}\tilde{\beta}) = {}_n|\tilde{\mathbf{e}};$$

y la suma de residuos al cuadrado también se puede expresar como

$$SRC(\tilde{\beta}) = \tilde{\mathbf{e}} \cdot \tilde{\mathbf{e}} = \sum_{n=1}^N \tilde{e}_n^2 = \sum_{n=1}^N (y_n - \tilde{y}_n)^2 = (\mathbf{y} - \tilde{\mathbf{y}}) \cdot (\mathbf{y} - \tilde{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\tilde{\beta}) \cdot (\mathbf{y} - \mathbf{X}\tilde{\beta})$$

(Lección 1) T-4 Criterio de ajuste MCO

Suponga $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ y $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$.

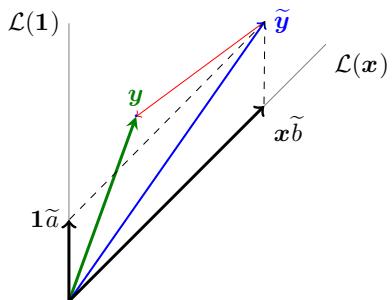
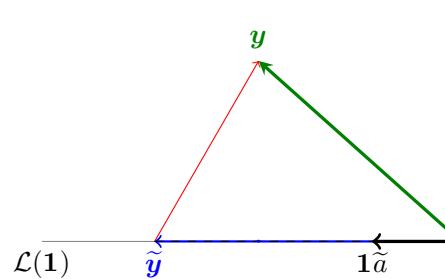
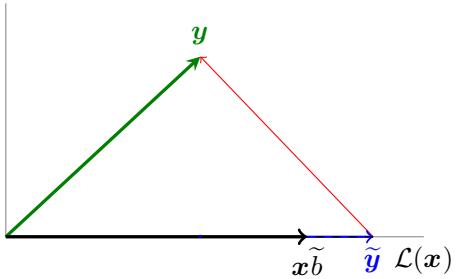
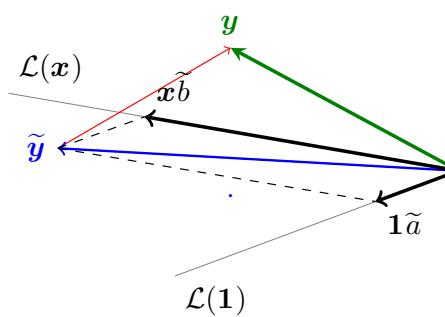
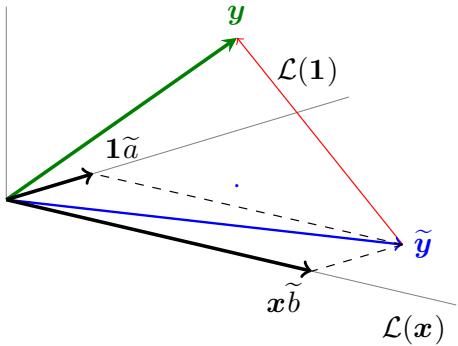
Como “criterio de ajuste” buscaremos un $\tilde{\beta}$ tal que $\mathbf{X}\tilde{\beta}$ esté lo más próximo posible a \mathbf{y} ; es decir, tal que

la componente $\tilde{\mathbf{e}}$ sea lo más pequeña posible en la descomposición:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\tilde{\beta} + \tilde{\mathbf{e}} \\ &= \tilde{\mathbf{y}} + \tilde{\mathbf{e}}. \end{aligned}$$

F9

Un \tilde{a} demasiado pequeño y un \tilde{b} demasiado grande.



$$\mathbf{X} = [1; \mathbf{x};];$$

$$\tilde{\beta} = \begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix};$$

$$\tilde{\mathbf{y}} = \mathbf{X}\tilde{\beta};$$

$$\mathbf{y} = \tilde{\mathbf{y}} + \tilde{\mathbf{e}};$$

$$\tilde{\mathbf{e}} = \mathbf{y} - \tilde{\mathbf{y}}$$

F10

Las cinco figuras muestran una representación del mismo ejemplo visto desde ángulos o perspectivas distintas (las dos primeras desde cierta altura, las dos siguientes desde el suelo y la última sería la “vista de pájaro” desde la vertical). En cada figura, el “suelo” es el plano (el subespacio) generado por los regresores $\mathbf{1}$ y \mathbf{x} ; es decir, el conjunto de todas las combinaciones lineales de los regresores, que en los libros de álgebra también se conoce como *espacio columna* de la matriz \mathbf{X} , y que con frecuencia se denota con $\mathcal{C}(\mathbf{X})$; es decir:

El suelo de las figuras es el conjunto de vectores $\mathbf{X}\mathbf{v}$ donde $\mathbf{v} \in \mathbb{R}^k$. Denotamos el conjunto con $\mathcal{C}(\mathbf{X})$.

Buscar la *combinación de regresores más próxima a \mathbf{y}* es buscar el punto del “suelo” más próximo a \mathbf{y} (más cercano al extremo superior de la flecha verde de la figura). Dicho de otro modo, buscar el vector de $\mathcal{C}(\mathbf{X})$ más próximo a \mathbf{y} es buscar la combinación lineal $\mathbf{X}\mathbf{v}$ que más se acerca a \mathbf{y} .

Si agudiza la vista verá que dicho punto aparece pintado en el suelo de las figuras. A partir de ahora a dicho punto (que es el ajuste MCO de \mathbf{y}) lo denominaremos con $\hat{\mathbf{y}}$.

En las figuras de más arriba se aprecia que la elección de los parámetros \tilde{a} y \tilde{b} es mejorable, pues usar \tilde{a} veces $\mathbf{1}$ y \tilde{b} veces \mathbf{x} nos conduce a un punto que no es el más próximo a \mathbf{y} , es decir, que

$$\hat{\mathbf{y}} \neq [1; \mathbf{x};] \begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix}.$$

El punto $\hat{\mathbf{y}}$ (el punto del “suelo” más próximo a \mathbf{y}) es la proyección ortogonal de \mathbf{y} sobre $\mathcal{C}(\mathbf{X})$. Para seguir con la exposición, recordemos la definición de la proyección ortogonal de un vector:

Definición 6. Llamamos proyección ortogonal de un vector \vec{y} sobre un subespacio \mathcal{V} a la función lineal $\text{Prj}_{\mathcal{V}}(\vec{y})$ tal que la diferencia $\vec{y} - \text{Prj}_{\mathcal{V}}(\vec{y})$ es ortogonal \mathcal{V} .

En el caso particular que nos ocupa, para cada vector de datos \mathbf{y} de \mathbb{R}^N , el vector $\hat{\mathbf{y}}$ del subespacio $\mathcal{C}(\mathbf{X})$ es la proyección ortogonal, $\text{Prj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$, de \mathbf{y} sobre el subespacio $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^m$ engendrado por las columnas de \mathbf{X} (i.e., por los regresores). Dicho vector existe y es único (véase la Lección 11 de Bujosa (2022a)). La diferencia $\mathbf{y} - \hat{\mathbf{y}}$ es el vector de errores $\hat{\mathbf{e}}$.

Lo más interesante de $\hat{\mathbf{y}}$ se afirma en la siguiente

Proposición 1.2. La proyección ortogonal de $\mathbf{y} \in \mathbb{R}^m$ sobre $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^m$ es el vector de $\mathcal{C}(\mathbf{X})$ más próximo a \mathbf{y} .

Demostración. Sea $\hat{\mathbf{y}}$ la proyección de \mathbf{y} sobre $\mathcal{C}(\mathbf{X})$ y tomemos un vector \mathbf{z} cualquiera de $\mathcal{C}(\mathbf{X})$. Veamos que \mathbf{y} está tan lejos o más de \mathbf{z} que de su proyección ortogonal $\hat{\mathbf{y}}$.

Como \mathbf{z} e $\hat{\mathbf{y}}$ están en $\mathcal{C}(\mathbf{X})$, su diferencia $(\hat{\mathbf{y}} - \mathbf{z})$ está en $\mathcal{C}(\mathbf{X})$ (pues $\mathcal{C}(\mathbf{X})$ es subespacio); consecuentemente $(\hat{\mathbf{y}} - \mathbf{z})$ es ortogonal a $(\mathbf{y} - \hat{\mathbf{y}}) = \hat{\mathbf{e}}$ (por ser $\hat{\mathbf{y}}$ la proyección ortogonal sobre $\mathcal{C}(\mathbf{X})$). Y como la suma de ambos vectores perpendiculares es $(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{z}) = (\mathbf{y} - \mathbf{z})$, por el Tma. de Pitágoras concluimos que

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|\hat{\mathbf{y}} - \mathbf{z}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \geq \|\mathbf{y} - \hat{\mathbf{y}}\|^2,$$

Por tanto, $\|\mathbf{y} - \mathbf{z}\| \geq \|\mathbf{y} - \hat{\mathbf{y}}\| = \|\hat{\mathbf{e}}\|$ para todo $\mathbf{z} \in \mathcal{C}(\mathbf{X})$. □

1.2.4 Las ecuaciones normales

El modo de encontrar los coeficientes $\hat{\beta}$ de la combinación de regresores $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ consiste en resolver un sistema de ecuaciones que se denomina *sistema de ecuaciones normales*.

Para verlo recordemos que dos vectores de \mathbb{R}^n son perpendiculares si y solo si su producto punto es cero

$$\mathbf{a} \perp \mathbf{b} \iff \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i = 0.$$

Consecuentemente, las filas de una matriz \mathbf{A} son perpendiculares a un vector \mathbf{b} si $\mathbf{A}\mathbf{b} = \mathbf{0}$, pues entonces $\mathbf{a}_i^\top \mathbf{A} \cdot \mathbf{b} = 0$. Por tanto, las columnas de una matriz \mathbf{C} son perpendiculares a \mathbf{b} si y solo si $(\mathbf{C}^\top)\mathbf{b} = \mathbf{0}$.

$$\mathbf{b} \perp \mathbf{C}_{|j} \iff \mathbf{C}^\top \mathbf{b} = \mathbf{0}$$

Así pues...

(Lección 1) T-6 Ecuaciones normales

El vector $\hat{\mathbf{e}}$ es mínimo cuando es perpendicular a cada regresor:

$$\hat{\mathbf{e}} \perp \mathbf{X}_{|j} \iff \mathbf{0} = \mathbf{X}^\top \hat{\mathbf{e}} = \mathbf{X}^\top (\mathbf{y} - \hat{\mathbf{y}}).$$

Consecuentemente

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \iff \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0} \iff \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{0}$$

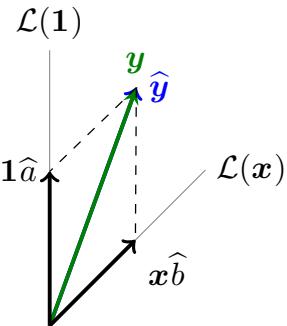
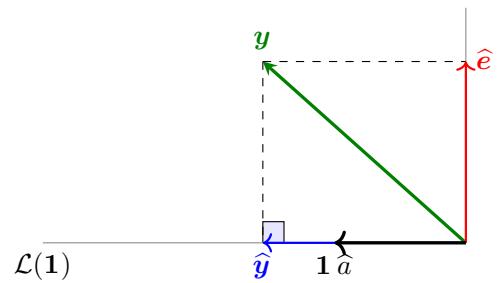
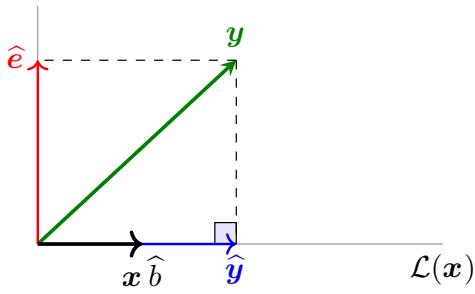
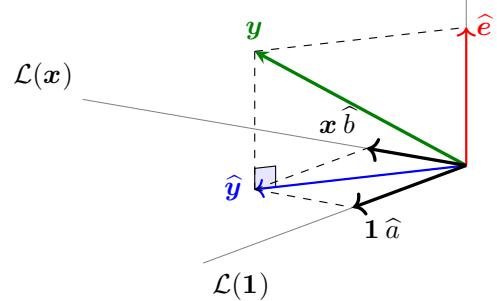
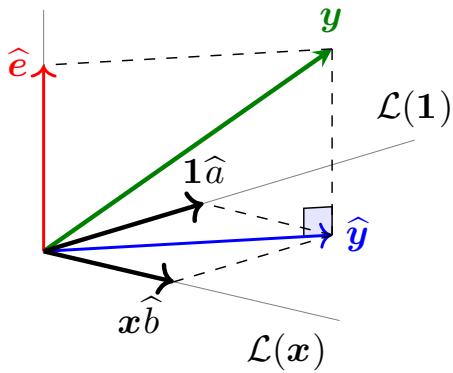
Es decir

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \quad \text{si y solo si} \quad (\mathbf{X}^\top \mathbf{X})\hat{\beta} = \mathbf{X}^\top \mathbf{y} \quad (1)$$

Las soluciones $\hat{\beta}$ son los parámetros del ajuste MCO $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$

(el ajuste que minimiza la longitud de $\hat{\mathbf{e}}$). F11

$$\mathbf{e} \perp \mathbf{X} \iff \hat{\beta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \text{ es solución de } \mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{y}.$$



$$\mathbf{X} = [\mathbf{1}; \mathbf{x}]; \quad \hat{\beta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}; \quad \boxed{\hat{y} = \mathbf{X}\hat{\beta}}; \quad \mathbf{y} = \hat{y} + \mathbf{e}; \quad \mathbf{e} = \mathbf{y} - \hat{y}$$

F12

El sistema de ecuaciones $(\mathbf{X}^\top \mathbf{X}) \hat{\beta} = \mathbf{X}^\top \mathbf{y}$ se denomina *sistema de ecuaciones normales*,⁷ y para obtener la proyección \hat{y} de \mathbf{y} sobre $\mathcal{C}(\mathbf{X})$ basta multiplicar \mathbf{X} por cualquier vector que sea solución de dicho sistema.⁸

Es llamativo que para resolver $\mathbf{X}\hat{\beta} = \hat{y}$ (para encontrar la combinación lineal de columnas de \mathbf{X} más próxima a \mathbf{y}) resolvamos el *sistema de ecuaciones normales*, donde no aparece \hat{y} . Este procedimiento indirecto funciona porque el *sistema de ecuaciones normales* y el sistema $\mathbf{X}\hat{\beta} = \hat{y}$ tienen el mismo conjunto de soluciones (fíjese en las implicaciones “si y solo si” de (1)).

⁷En este contexto geométrico, normalidad significa perpendicularidad, de ahí el nombre.

⁸Fíjese que aunque el punto \hat{y} es único, el sistema de ecuaciones normales puede tener infinitas soluciones, es decir, puede haber infinitas combinaciones de los regresores que sean iguales a \hat{y} . En la siguiente sección veremos la condición sobre \mathbf{X} para que la solución sea única.

1.2.5 Ausencia de multicolinealidad exacta (regresores linealmente independientes)

Para que la solución al sistema de ecuaciones normales (1) sea única (para que $\hat{\beta}$ sea único) es necesario que se verifique una condición sobre la matriz de regresores \mathbf{X} :

(Lección 1) **T-8** Condición para que las ecuaciones normales tengan solución única

Puesto que

$$\mathbf{X}\hat{\beta} = \hat{\mathbf{y}} \iff (\mathbf{X}^\top \mathbf{X})\hat{\beta} = \mathbf{X}^\top \mathbf{y}, \quad \text{donde } \mathbf{X}_{N \times k};$$

ambos sistemas tendrán *solución única si y sólo* si sus matrices de coeficientes son de *rango k*.

En tal caso, multiplicando ambos lados de las ecuaciones normales por $(\mathbf{X}^\top \mathbf{X})^{-1}$ tenemos que

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (2)$$

es la *única solución*.

F13

La condición de independencia lineal de las columnas de \mathbf{X} implica que $\mathbf{X}^\top \mathbf{X}$ es de rango completo⁹ y garantiza la unicidad de las soluciones. Es decir, para una matriz con k columnas $\mathbf{X}_{N \times k}$ (i.e., cuando hay k regresores):

$$\hat{\beta} \text{ es único} \iff \text{rg}(\mathbf{X}) = k \iff (\mathbf{X}^\top \mathbf{X})^{-1} \text{ es invertible.}$$

Cuando no se cumple la condición de rango existen infinitos $\hat{\beta}$'s (infinitos valores para los parámetros) tales que $\mathbf{X}\hat{\beta} = \hat{\mathbf{y}}$. En tal caso, se dice que hay *multicolinealidad perfecta* (i.e., *los regresores son linealmente dependientes*). Veamos un ejemplo.

Ejemplo 2. Ecuación de salarios: Supongamos el siguiente modelo (Ejemplo 3.2. Wooldridge, 2006)

$$Salar_n = e^{(\beta_1 + \beta_2(educ_n) + \beta_3(antig_n) + \beta_4(exper_n) + otrosas_n)};$$

donde $Salar_n$ es el salario del individuo n -ésimo, $educ_n$ son sus años de educación, $antig_n$ sus años de antigüedad en la empresa, y $exper_n$ sus años de experiencia en el sector y $otrosas_n$ son otros factores distintos de los anteriores.

Tomando logaritmos tenemos un modelo para la nueva variable $\ln(Salar_n)$

$$\ln(Salar_n) = \beta_1 + \beta_2(educ_n) + \beta_3(antig_n) + \beta_4(exper_n) + otrosas_n.$$

(donde ahora, el regresando $\ln(Salar_n)$ es una combinación lineal de los regresores más *otrosas*).

¿Qué pasa si jamás ninguno trabajador cambió de empresa? Entonces las columnas de la matriz de regresores \mathbf{X} correspondientes a *años de experiencia* y *años de antigüedad* son iguales; así que $\mathbf{X}^\top \mathbf{X}$ no es invertible.

Como *experiencia* y *antigüedad* coinciden, sólo podemos calcular su *efecto conjunto*:

$$\ln(Salar_n) = \beta_1 + \beta_2(educ_n) + (\beta_3 + \beta_4)exper_n + otrosas_n,$$

es decir, no es posible discriminar el “efecto” de la variables *antigüedad* y *experiencia* por separado.

Fíjese que si una de las columnas de \mathbf{X} es un múltiplo de otra; es decir, si la correlación entre dos regresores es *1 en valor absoluto*¹⁰ (por ejemplo, si la tercera columna es a veces la segunda) habrá multicolinealidad exacta. Pero no es necesario que haya correlación *uno en valor absoluto* entre algunos regresores para que haya multicolinealidad exacta; por ejemplo, si la tercera columna es a veces la primera más b veces la segunda, evidentemente las columnas de \mathbf{X} serán linealmente dependientes (aunque no haya correlación uno en valor absoluto entre dichas columnas, es decir, aunque ninguna columna sea un múltiplo de otra).

⁹lo que implica que para cada regresor hay al menos tantas observaciones como regresores tiene el modelo ($N \geq k$).

¹⁰Véase la correlación entre vectores alineados en la Página 20.

1.2.6 ¡Ojo! El lenguaje habitual en econometría se toma ciertas licencias

El modelo inicial para los salarios en el ejemplo anterior no es lineal en los parámetros, y una vez se ha transformado logarítmicamente dicho modelo *tampoco!*... pues lo podemos escribir como

$$\ln(Salary_n) = [1; educ; antig; exper;] \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \text{otras cosas}$$

o de manera más compacta

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \text{otras cosas}.$$

Dicho modelo no cumple las condiciones de la Definición 2 (ya que no es de la forma *matriz por vector*). No obstante, los libros de econometría se refieren a un modelo así como lineal en los parámetros porque se puede aproximar mediante el modelo lineal $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Por ello se dice que tomando logaritmos “se *linealiza* el modelo” inicial, pero la expresión no es estrictamente correcta. Como dicha expresión está completamente extendida en econometría, aquí también diremos que un modelo de la forma $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ es lineal en los parámetros, aunque realmente solo sería lineal si \mathbf{u} estuviera multiplicado por un parámetro adicional β_{k+1} . Así pues, en econometría, al decir que el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ es lineal en los parámetros lo que se quiere decir es que se puede aproximar por MCO mediante el modelo lineal $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

Enlace a algunas **prácticas de la Lección 1**

Fin de la lección

LECCIÓN 2: Modelos con 1, 2 y 3 regresores

Pensando en aquellos alumnos con inclinaciones matemáticas, comenzaré la lección con una introducción algo más formal de lo que suele ser habitual. Esta introducción enlazará con los espacios vectoriales de variables aleatorias que encontraremos más adelante. Como no todos los alumnos tienen dicha inclinación, en las transparencias esta introducción se pasa por alto.

Pues bien, como el entorno en que se desarrolla la Econometría son los los *espacios semi-euclídeos*,¹¹ empezaré dedicando unas páginas a los (semi-)productos escalares.

2 Espacios euclídeos

Un espacio semi-euclídeo es un espacio vectorial dotado de un semi producto escalar. Los espacios semi-euclídeos permiten medir distancias y ángulos. Más adelante veremos que los empleados en Estadística y Probabilidad tienen la particularidad de que son de tal manera que los vectores (o funciones) constantes “uno” (aquellos que solo toman el valor uno) siempre tienen longitud (o norma) igual a 1.

En este primer tema (donde tan solo trataremos con listas de datos) trabajaremos dentro del *espacio euclídeo* \mathbb{R}^N . Más adelante trabajaremos con espacios semi-euclídeos más generales, aunque formados por variables aleatorias.

2.1 Productos escalares en general

2.1.1 Propiedades de los productos escalares

Un producto escalar es una función bilineal $\langle \cdot | \cdot \rangle$ que satisface los siguientes axiomas para cualesquiera vectores \vec{x} , \vec{y} y \vec{z} de un espacio vectorial \mathcal{V} y para cualquier número real a .

- **Simetría:** $\langle \vec{x} | \vec{y} \rangle = \langle \vec{y} | \vec{x} \rangle$
- **Linealidad respecto al primer argumento:**
 1. $\langle a\vec{x} | \vec{y} \rangle = a\langle \vec{x} | \vec{y} \rangle$
 2. $\langle \vec{x} + \vec{y} | \vec{z} \rangle = \langle \vec{x} | \vec{z} \rangle + \langle \vec{y} | \vec{z} \rangle$
- **Positivo:** $\langle \vec{x} | \vec{x} \rangle \geq 0$
- **Definido:** $\langle \vec{x} | \vec{x} \rangle = 0 \Leftrightarrow \vec{x} = 0$.

La diferencia entre un *producto escalar* y un *semi producto escalar* es que éste último no es definido, es decir, no tiene por qué cumplir la última propiedad. Consecuentemente, con un *semi producto escalar* puede ocurrir que $\langle \vec{x} | \vec{x} \rangle = 0$ para algunos vectores \vec{x} no nulos ($\vec{x} \neq 0$).¹²

2.1.2 Los productos escalares sirven para medir longitudes y ángulos.

Fijado un producto escalar particular $\langle \cdot | \cdot \rangle$ en un espacio vectorial \mathcal{V} , la norma o longitud de un vector \vec{x} es

$$\| \vec{x} \| = \sqrt{\langle \vec{x} | \vec{x} \rangle}; \quad (3)$$

y el coseno del ángulo θ formado por dos vectores no nulos \vec{x} e \vec{y} es

$$\cos \theta = \frac{\langle \vec{x} | \vec{y} \rangle}{\| \vec{x} \| \cdot \| \vec{y} \|};$$

que toma valores entre -1 y 1 . Cuando el coseno es 1 en *valor absoluto* (1 ó -1) decimos que los vectores \vec{x} e \vec{y} están alineados (i.e., uno de los vectores es múltiplo del otro). Por otra parte, dos vectores \vec{x} e \vec{y} son **perpendiculares** si y solo si su (semi)producto escalar $\langle \vec{x} | \vec{y} \rangle$ es nulo:

$$\vec{x} \perp \vec{y} \quad \Leftrightarrow \quad \langle \vec{x} | \vec{y} \rangle = 0.$$

¹¹El término espacio semi-euclídeo está tomado de Golovina (1980) y corresponde a un espacio vectorial con un semi producto escalar.

¹²En \mathbb{R}^N tan solo emplearemos *productos escalares*. Pero en temas posteriores, cuando tratemos con variables aleatorias en general, necesitaremos emplear semi-productos escalares.

2.2 Producto escalar usual en el espacio euclídeo \mathbb{R}^N

El producto escalar “usual” entre vectores de \mathbb{R}^n es el *producto punto* (que ya hemos usado en la lección anterior):

$$\langle \mathbf{x} | \mathbf{y} \rangle_u = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^N x_i y_i.$$

(Fíjese en el subíndice “u” que indica que nos referimos al producto escalar “usual” del espacio euclídeo \mathbb{R}^N).

Nota: Aunque los textos de estadística usan mayoritariamente expresiones con sumatorios (como la de la derecha), yo usaré de manera preferente expresiones como la de la izquierda o la del centro, por ser mucho más breves.

Por (3) sabemos que el producto escalar de un vector por si mismo es su norma al cuadrado, es decir, es el cuadrado de su longitud. Así, empleando el producto escalar $\langle \cdot | \cdot \rangle_u$ usual en el espacio euclídeo \mathbb{R}^N tenemos que

$$\langle \mathbf{x} | \mathbf{x} \rangle_u = \mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|_u^2.$$

Consecuentemente llamamos *norma euclídea* de \mathbf{x} en \mathbb{R}^N a

$$\|\mathbf{x}\|_u = \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

(Fíjese en el subíndice “u”, que indica que estamos usando la norma del producto escalar “usual”).

2.2.1 Descomposición de los semi-productos escalares

Producto componente a componente en \mathbb{R}^N (o producto Hadamard). Definimos el producto *componente a componente*, $\mathbf{x} \odot \mathbf{y}$, entre dos vectores \mathbf{x} e \mathbf{y} de \mathbb{R}^N como el vector de \mathbb{R}^N cuya componente n -ésima es el producto de las componentes n -ésimas de \mathbf{x} e \mathbf{y} : $(x_1, \dots, x_n) \odot (y_1, \dots, y_n) = (x_1 y_1, \dots, x_n y_n)$.

$$\text{Es decir, } (\mathbf{x} \odot \mathbf{y})_{|n} = x_n y_n; \quad \text{por ejemplo,} \quad \mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1 \\ -3 \\ 5 \\ 0 \end{pmatrix} \Rightarrow \mathbf{x} \odot \mathbf{y} = \begin{pmatrix} 1 \\ -6 \\ 15 \\ 0 \end{pmatrix}.$$

Es fácil comprobar que si a es un número y \mathbf{x} e \mathbf{y} son vectores de \mathbb{R}^N :

- $\mathbf{x} \odot \mathbf{y} = \mathbf{y} \odot \mathbf{x}$
- $(\mathbf{x} + \mathbf{y}) \odot \mathbf{z} = \mathbf{x} \odot \mathbf{z} + \mathbf{y} \odot \mathbf{z}$
- $\mathbf{1} \odot \mathbf{x} = \mathbf{x}$,
- $(ax) \odot \mathbf{y} = a(\mathbf{y} \odot \mathbf{x})$
- $\mathbf{0} \odot \mathbf{x} = \mathbf{0}$

donde $\mathbf{0}$ es el vector cuyas componentes son nulas y donde $\mathbf{1}$ es el vector cuyas componentes son todas iguales a 1.

Con \mathbf{y}^2 nos referiremos al vector $\mathbf{y} \odot \mathbf{y}$, es decir, al vector cuyas componentes son el cuadrado de las de \mathbf{y} . De esta manera tenemos que el cuadrado de la longitud de \mathbf{y} es $\|\mathbf{y}\|_u^2 = \langle \mathbf{y} | \mathbf{y} \rangle_u = \mathbf{y} \cdot \mathbf{y} = \sum_{i=1}^N y_i^2 = \mathbf{1} \cdot \mathbf{y}^2$.

Fíjese que si $\mathbf{x}_1 \odot \mathbf{y}_1 = \mathbf{x}_2 \odot \mathbf{y}_2$ entonces $\langle \mathbf{x}_1 | \mathbf{y}_1 \rangle_u = \mathbf{x}_1 \cdot \mathbf{y}_1 = \mathbf{x}_2 \cdot \mathbf{y}_2 = \langle \mathbf{x}_2 | \mathbf{y}_2 \rangle_u$, es decir, que

el conjunto de pares $\left\{ (\mathbf{x} \odot \mathbf{y}, \langle \mathbf{x} | \mathbf{y} \rangle_u) \mid \mathbf{x}, \mathbf{y} \in \mathbb{R}^N \right\}$ es una función.

(Véase la definición de función en el apéndice de la Lección 6 de *Curso de Álgebra Lineal con notación asociativa...*). Todo semiproducto escalar que cumple lo anterior se puede descomponer en dos operaciones: un producto componente a componente (o punto a punto) y alguna forma de agregación (que normalmente se denomina suma o *integral*¹³). Veámoslo en el caso del producto escalar usual en \mathbb{R}^N .

El producto escalar usual se descompone en dos operaciones: un *producto* de funciones componente a componente (punto a punto) y la *suma* de las componentes del vector resultante (véase la Figura 1).

$$\langle \mathbf{x} | \mathbf{y} \rangle_u = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^N x_i y_i = \boxed{\text{la } \textcolor{blue}{\text{suma}} \text{ de las componentes del producto } \mathbf{x} \odot \mathbf{y}}.$$

¹³Integral de Lebesgue.

$$\begin{array}{ccc}
 \mathbb{R}^N \times \mathbb{R}^N & \xrightarrow{\mathbf{x} \odot \mathbf{y}} & \mathbb{R}^N \\
 & \searrow \langle \mathbf{x} | \mathbf{y} \rangle_u & \downarrow \text{Suma de componentes de } \mathbf{x} \odot \mathbf{y} = \sum_{i=1}^N x_i y_i \\
 & & \mathbb{R}
 \end{array}$$

Figure 1: Diagrama del producto escalar usual en \mathbb{R}^N , que es la suma de las componentes del producto Hadamard.

2.3 El producto escalar de uso más común en estadística

El *promedio de un conjunto de datos* es una idea central en Estadística. Dicho promedio es un valor que queremos que, de algún modo, “resuma” el conjunto de datos. Esta idea condiciona el tipo de producto empleado en estadística.

Imagine que recabamos un conjunto de datos y todos ellos resultan ser “unos”, parece natural pensar que el valor que mejor “resume” dicho conjunto sea precisamente el número 1. Para lograr que así sea, en probabilidad y estadística se emplean productos escalares tales que *todo vector constante $\mathbf{1}$ tenga longitud 1*. Fíjese que el producto escalar $\langle \cdot | \cdot \rangle_u$ usual en \mathbb{R}^N no cumple este requisito cuando $N > 1$, pues

$$\|\mathbf{1}\|_u^2 = \mathbf{1} \cdot \mathbf{1} = \sum_{n=1}^N 1^2 = N$$

y por tanto $\|\mathbf{1}\|_u = \sqrt{N}$. En estadística el producto escalar más habitual es:¹⁴

$$\langle \mathbf{x} | \mathbf{y} \rangle_s = \frac{1}{N} (\mathbf{x} \cdot \mathbf{y}). \quad (4)$$

(Fíjese que el subíndice “s” indica que nos referimos al producto escalar frecuentemente usado en estadística).

Con este nuevo producto escalar la norma al cuadrado de un vector \mathbf{x} es

$$\|\mathbf{x}\|_s^2 = \langle \mathbf{x} | \mathbf{x} \rangle_s = \frac{1}{N} (\mathbf{x} \cdot \mathbf{x}) = \frac{1}{N} (\mathbf{1} \cdot \mathbf{x}^2); \quad (5)$$

y arroja el resultado deseado independientemente del número N de componentes (de unos) del vector $\mathbf{1}$, pues

$$\|\mathbf{1}\|_s^2 = \frac{1}{N} (\mathbf{1} \cdot \mathbf{1}) = \frac{N}{N} = 1; \quad \Rightarrow \quad \|\mathbf{1}\|_s = \sqrt{1} = 1$$

(donde el subíndice “s”, que indica que es la norma del producto escalar frecuentemente usado en estadística).

Para abreviar me referiré a este producto escalar y a esta norma como “*los de la estadística*” (pero recuerde que el nombre es engañoso, pues no son los únicos que se usan en estadística)¹⁵. A partir de ahora, si no se indica lo contrario, al decir “la norma de un vector \mathbf{x} ” nos referiremos a la norma en estadística $\|\mathbf{x}\|_s = \sqrt{N^{-1} \mathbf{x} \cdot \mathbf{x}}$.

Media aritmética Empleando este nuevo producto escalar podemos definir la *media aritmética* (o media):

Definición 7. La media aritmética, $\mu_{\mathbf{y}}$, de un vector \mathbf{y} es el producto escalar (de la estadística) entre $\mathbf{1}$ e \mathbf{y} :

$$\mu_{\mathbf{y}} = \langle \mathbf{1} | \mathbf{y} \rangle_s = \frac{1}{N} (\mathbf{1} \cdot \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N y_i. \quad (6)$$

Fíjese que la media es una forma de “sumar” (o agregar) que cumple el objetivo de que el “promedio” de un vector constante $\mathbf{1}$ sea siempre 1: $\boxed{\mu_{\mathbf{1}} = 1}$.

Fíjese también en que, como los productos escalares son lineales tanto en el primer como en el segundo argumento, la media aritmética (que es un producto escalar con $\mathbf{1}$) es lineal y por tanto tenemos que $\boxed{\mu_{(ay)} = a(\mu_{\mathbf{y}})}$ y

$$\boxed{\mu_{(\mathbf{x} + \mathbf{y})} = \mu_{\mathbf{x}} + \mu_{\mathbf{y}}}; \text{ o de manera más compacta } \boxed{\mu_{(ax + by)} = a\mu_{\mathbf{x}} + b\mu_{\mathbf{y}}}.$$

¹⁴Consecuentemente el producto escalar usual, $\langle \cdot | \cdot \rangle_u$ arroja valores N veces mayores (en valor absoluto) que el de la estadística, $\langle \cdot | \cdot \rangle_s$.

¹⁵Al calcular con medias ponderadas (donde hay ponderaciones distintas) estaremos empleando productos escalares diferentes.

Por último, como $\mathbf{x}_1 \odot \mathbf{y}_1 = \mathbf{x}_2 \odot \mathbf{y}_2 \implies \langle \mathbf{x}_1 | \mathbf{y}_1 \rangle_s = \langle \mathbf{x}_2 | \mathbf{y}_2 \rangle_s$, es decir, como de nuevo el conjunto de pares $\{(\mathbf{x} \odot \mathbf{y}, \langle \mathbf{x} | \mathbf{y} \rangle_s) \mid \mathbf{x}, \mathbf{y} \in \mathbb{R}^N\}$ es una función, también podemos factorizar el producto escalar de la estadística:

$$\langle \mathbf{x} | \mathbf{y} \rangle_s = N^{-1}(\mathbf{x} \cdot \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N x_i y_i = \boxed{\text{media del producto } \mathbf{x} \odot \mathbf{y}} = \mu_{(\mathbf{x} \odot \mathbf{y})};$$

es decir, el producto escalar de la estadística es la composición del producto Hadamard y una operación de tipo “suma” (que en este caso se denomina la *media aritmética*¹⁶). *En adelante recuerde que el producto punto $\mathbf{x} \cdot \mathbf{y}$ es el producto escalar usual de \mathbb{R}^N , y que la media aritmética del producto $\mu_{(\mathbf{x} \odot \mathbf{y})}$ es el producto escalar de la estadística.*

$$\begin{array}{ccc} \mathbb{R}^N \times \mathbb{R}^N & \xrightarrow{\mathbf{x} \odot \mathbf{y}} & \mathbb{R}^N \\ & \searrow & \downarrow \text{media del producto } \mathbf{x} \odot \mathbf{y}: \quad \sum_{i=1}^N \frac{x_i y_i}{N} = \mu_{(\mathbf{x} \odot \mathbf{y})} \\ \langle \mathbf{x} | \mathbf{y} \rangle_s & \rightarrow & \mathbb{R} \end{array}$$

Figure 2: Diagrama del producto escalar en \mathbb{R}^N en estadística; que es la media del producto componente a componente.

Ortogonalidad en \mathbb{R}^N . Si dos vectores son perpendiculares con el producto escalar usual $\langle \cdot | \cdot \rangle_u$, también lo son con el producto escalar de la estadística, $\langle \cdot | \cdot \rangle_s$, pues

$$\mathbf{x} \perp \mathbf{y} \Leftrightarrow \mathbf{x} \cdot \mathbf{y} = 0 \Leftrightarrow N^{-1}(\mathbf{x} \cdot \mathbf{y}) = 0 \Leftrightarrow \mu_{(\mathbf{x} \odot \mathbf{y})} = 0; \quad (7)$$

Además, como $\mathbf{x} \odot \mathbf{1} = \mathbf{x}$, concluimos que un vector \mathbf{x} es perpendicular a $\mathbf{1}$ si y solo si su media es cero:

$$\mathbf{x} \perp \mathbf{1} \Leftrightarrow \mu_{(\mathbf{x} \odot \mathbf{1})} = \mu_{\mathbf{x}} = 0. \quad (8)$$

Dos normas distintas! Como disponemos de dos productos escalares en \mathbb{R}^N , disponemos de dos normas (o maneras de medir) diferentes; y por cuestiones de tradición trabajaremos con ambas en la lección siguiente.¹⁷ Afortunadamente el paso de una a otra es sencillo cuando manejamos *el cuadrado de la norma*. Por (4) sabemos que el cuadrado de la norma usual $\|\mathbf{x}\|_u^2 = \mathbf{x} \cdot \mathbf{x}$ es N veces mayor que el cuadrado de la norma en estadística $\|\mathbf{x}\|_s^2 = N^{-1}(\mathbf{x} \cdot \mathbf{x})$; es decir,

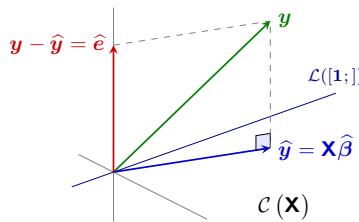
$$\|\mathbf{x}\|_u^2 = N \|\mathbf{x}\|_s^2. \quad (9)$$

3 Ajuste MCO con uno, dos, tres ó k regresores

(Lección 2) T-1 Geometría MCO

El ajuste de regresión MCO es una descomposición ortogonal:

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}; \quad \text{donde} \quad \hat{\mathbf{y}} = \mathbf{X} \hat{\beta} \perp \hat{\mathbf{e}}$$



donde los parámetros $\hat{\beta}$ se obtienen resolviendo $\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{y}$ y donde $\mathbf{1} \in \mathcal{C}(\mathbf{X})$. F16

*Solo hablamos de regresión (o ajuste) MCO cuando $\mathbf{1}$ pertenece al subespacio $\mathcal{C}(\mathbf{X})$ sobre el que se proyecta \mathbf{y} .*¹⁸

¹⁶O *esperanza matemática*, o en un contexto más general, integral de Lebesgue.

¹⁷En Econometría es tradicional referirse tanto a la *suma de cuadrados* (cuadrado de la norma que usa el producto escalar habitual) como a la *varianza* (cuadrado de la norma que usa el producto escalar de la estadística).

¹⁸Es decir, una proyección ortogonal es una *regresión* si y solo si los vectores constantes pertenecen al subespacio sobre el que se proyecta. Habitualmente esto se garantiza en Econometría imponiendo que la primera columna de \mathbf{X} sea el vector constante $\mathbf{1}$; aunque no siempre es así (véase la lección sobre variables ficticias).

3.1 Una constante como único regresor (media, desviación típica y varianza)

Estudiemos el caso más sencillo, donde la única columna de \mathbf{X} es el vector $\mathbf{1}$. Este caso corresponde al modelo

$$Y_n = a + \text{otrascosas}_n$$

(Lección 2) **T-2** Ajuste MCO con una constante (vector de unos) como único regresor

¿Qué es el ajuste MCO $\hat{\mathbf{y}}$ si $\mathbf{X} = [\mathbf{1}]$? ($Y_n = a + \text{otrascosas}_n$)

Las ecuaciones normales

$$\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$$

se reducen a una única ecuación

$$[\mathbf{1} \cdot \mathbf{1}](\hat{a},) = (\mathbf{1} \cdot \mathbf{y},) \implies (\hat{a},) = [N]^{-1}(\mathbf{1} \cdot \mathbf{y},)$$

Por tanto $\hat{a} = N^{-1}(\mathbf{1} \cdot \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N y_i = \mu_y$; así que

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = [\mathbf{1}](\hat{a},) = \mathbf{1} \mu_y \equiv \bar{\mathbf{y}}. \quad (10)$$

F17

Más arriba vimos que *la media μ_y es el producto escalar (en estadística) entre $\mathbf{1}$ e \mathbf{y}* ; pero ahora tenemos una segunda “forma de ver” la media aritmética; una forma que está íntimamente relacionada con la proyección ortogonal:

La media μ_y es el valor por el que hay que multiplicar $\mathbf{1}$ para obtener el vector constante más próximo a \mathbf{y} .

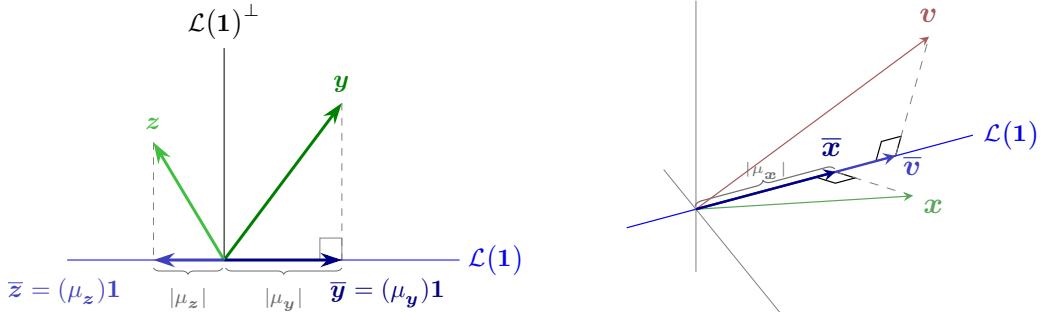
Demos nombre a dicho vector (que denotaremos con $\bar{\mathbf{y}}$):

Definición 8. Llamamos *vector de medias* de \mathbf{y} a la proyección $\bar{\mathbf{y}}$ de \mathbf{y} sobre los vectores constantes $\mathcal{L}(\mathbf{1})$.

Por tanto $\bar{\mathbf{y}} = \begin{pmatrix} \mu_y \\ \vdots \\ \mu_y \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu_y = \mathbf{1} \mu_y$, donde μ_y es la *media* de \mathbf{y} .

Fíjese que como las proyecciones ortogonales son funciones lineales (Definición 6 en la página 9), el vector de medias de una combinación lineal es una combinación lineal de vectores de medias: $\overline{(ax + by)} = a\bar{x} + b\bar{y}$.

(Lección 2) **T-3** El vector de medias $\bar{\mathbf{y}}$ y la media aritmética μ_y



$$\bar{\mathbf{y}} = (\mu_y) \mathbf{1} \quad \text{y} \quad \mu_y = \langle \mathbf{y} | \mathbf{1} \rangle_s.$$

F18

La media del *vector de medias* $\bar{\mathbf{x}}$ es igual a la media de \mathbf{x} , es decir $\mu_{\bar{\mathbf{x}}} = \mu_{\mathbf{x}}$.

Demostración. (Si piensa en la representación geométrica verá que el resultado es trivial. La demo también lo es.)

Como $\bar{\mathbf{x}} = (\mu_{\mathbf{x}})\mathbf{1}$ y la media aritmética es lineal, y como $\mu_{\mathbf{1}} = 1$; tenemos: $\mu_{\bar{\mathbf{x}}} = \mu_{((\mu_{\mathbf{x}})\mathbf{1})} = (\mu_{\mathbf{x}})\mu_{(\mathbf{1})} = \mu_{\mathbf{x}}$. \square

Por otra parte, como $\bar{\mathbf{x}}$ es la proyección ortogonal de \mathbf{x} sobre $\mathbf{1}$, se verifica que $\mathbf{1} \perp (\mathbf{x} - \bar{\mathbf{x}})$; de hecho,

la diferencia $(\mathbf{x} - \bar{\mathbf{x}})$ es la proyección ortogonal de \mathbf{x} sobre el subespacio $\mathcal{L}([\mathbf{1}])^\perp$ ortogonal a $\mathbf{1}$.

Por tanto, la media de $(\mathbf{x} - \bar{\mathbf{x}})$ es necesariamente cero (Ecuación 8 en la página 16).

Resumiendo, cuando proyectamos ortogonalmente un vector \mathbf{x} de \mathbb{R}^N sobre $\mathbf{1}$, lo descomponemos en una *componente constante* $\bar{\mathbf{x}}$ y otra componente $(\mathbf{x} - \bar{\mathbf{x}})$ que es perpendicular a la primera; y que muchos manuales de econometría denominan “*vector de desviaciones respecto a su media*”. Aquí, y por contraposición, la llamaré *componente variable*.¹⁹ La longitud de la *componente constante* $\bar{\mathbf{x}}$ es $\|\bar{\mathbf{x}}\|_s = |\mu_{\mathbf{x}}|$. A continuación nos fijaremos en la longitud de la *componente variable* $(\mathbf{x} - \bar{\mathbf{x}})$.

Desviación típica y varianza. De todos los múltiplos del vector $\mathbf{1}$, el vector de medias $\bar{\mathbf{x}}$ es el que está a la menor distancia de \mathbf{x} . Dicha distancia se denomina *desviación típica* (i.e., la longitud de la componente variable $\mathbf{x} - \bar{\mathbf{x}}$):

Definición 9. Llamamos *desviación típica* de \mathbf{x} a la norma (o longitud) $\|\mathbf{x} - \bar{\mathbf{x}}\|_s$; y la denotamos con $\sigma_{\mathbf{x}}$.

$$\sigma_{\mathbf{x}} = \|\mathbf{x} - \bar{\mathbf{x}}\|_s = \sqrt{\langle (\mathbf{x} - \bar{\mathbf{x}}) | (\mathbf{x} - \bar{\mathbf{x}}) \rangle_s} = \sqrt{\mu_{((\mathbf{x} - \bar{\mathbf{x}})^2)}} = \sqrt{\sum_{i=1}^N \frac{(x_i - \mu_{\mathbf{x}})^2}{N}}.$$

Como las proyecciones ortogonales son funciones lineales (página 6), la proyección de $\lambda\mathbf{x}$ sobre $\mathcal{L}(\mathbf{1})^\perp$ es $\lambda(\mathbf{x} - \bar{\mathbf{x}})$, cuya longitud es la desviación típica de $\lambda\mathbf{x}$, por tanto $\boxed{\sigma_{(\lambda\mathbf{x})} = |\lambda|\sigma_{\mathbf{x}}}$.

Definición 10. El cuadrado de la desviación típica de \mathbf{x} se denomina *varianza de \mathbf{x}* :

$$\sigma_{\mathbf{x}}^2 = \|\mathbf{x} - \bar{\mathbf{x}}\|_s^2 = \langle (\mathbf{x} - \bar{\mathbf{x}}) | (\mathbf{x} - \bar{\mathbf{x}}) \rangle_s = \mu_{((\mathbf{x} - \bar{\mathbf{x}})^2)} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{\mathbf{x}})^2.$$

Dado que al multiplicar \mathbf{x} por λ su desviación típica se multiplica por $|\lambda|$, tenemos que $\boxed{\sigma_{(\lambda\mathbf{x})}^2 = \lambda^2 \sigma_{\mathbf{x}}^2}$.

La ortogonalidad entre el vector de desviaciones (o *componente variable*) $(\mathbf{x} - \bar{\mathbf{x}})$ y el vector constante $\mathbf{1}$ tiene implicaciones geométricas que son permanentemente explotadas por la estadística ¡Resulta que la estadística es fundamentalmente una aplicación del Teorema de Pitágoras!

Tma. de Pitágoras y la estadística Recuerde que si \mathbf{a} y \mathbf{b} son perpendiculares, y $\mathbf{c} = \mathbf{a} + \mathbf{b}$, entonces:

P-1

$$\|\mathbf{c}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2,$$

donde \mathbf{a} y \mathbf{b} son los catetos y \mathbf{c} la hipotenusa de un triángulo rectángulo. Así, como $(\mathbf{y} - \bar{\mathbf{y}}) \perp \bar{\mathbf{y}}$, tenemos que

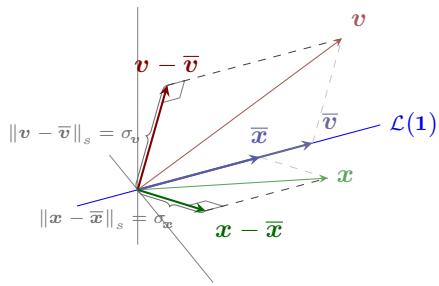
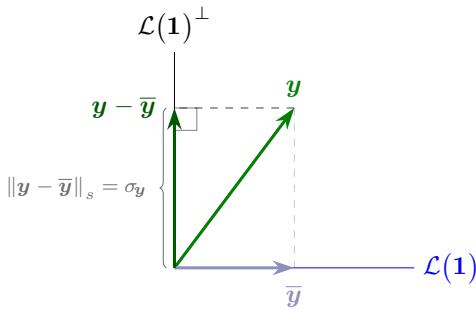
$$\|\mathbf{y}\|_s^2 = \|\mathbf{y} - \bar{\mathbf{y}}\|_s^2 + \|\bar{\mathbf{y}}\|_s^2 = \sigma_{\mathbf{y}}^2 + \|\bar{\mathbf{y}}\|_s^2,$$

es decir, que $\sigma_{\mathbf{y}}^2 = \|\mathbf{y}\|_s^2 - \|\bar{\mathbf{y}}\|_s^2$. Por (5) y (6) sabemos que $\boxed{\|\mathbf{y}\|_s^2 = \mu_{(\mathbf{y}^2)}}$; y como la longitud de $\mathbf{1}$ es 1 y $\bar{\mathbf{y}} = \mathbf{1}\mu_{\mathbf{y}}$, también sabemos que $\boxed{\|\bar{\mathbf{y}}\|_s^2 = |\mu_{\mathbf{y}}|}$. Consecuentemente llegamos a una segunda expresión para la varianza

$$\sigma_{\mathbf{y}}^2 = \mu_{(\mathbf{y}^2)} - (\mu_{\mathbf{y}})^2, \tag{11}$$

donde $\mu_{(\mathbf{y}^2)}$ es el cuadrado de la hipotenusa y los otros dos términos son el cuadrado de los catetos del triángulo.

¹⁹Pero fíjese que dicha *componente variable* pudiera ser en algún caso el vector constante $\mathbf{0}$.



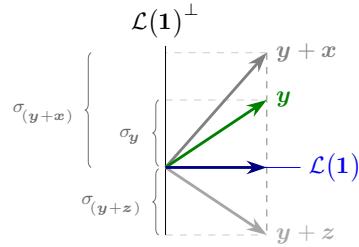
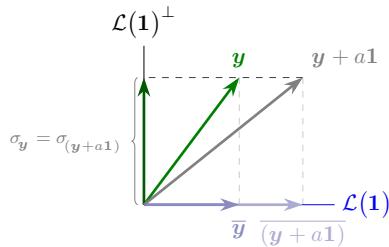
$$\text{Así, } \sigma_y^2 = \|\mathbf{y} - \bar{\mathbf{y}}\|_s^2 = N^{-1} \sum (y_i - \mu_y)^2 = \mu_{((\mathbf{y} - \bar{\mathbf{y}})^2)},$$

pero por el T. de Pitágoras, también

$$\sigma_y^2 = \|\mathbf{y}\|_s^2 - \|\bar{\mathbf{y}}\|_s^2 = \mu_{(\mathbf{y}^2)} - (\mu_{\mathbf{y}})^2.$$

F19

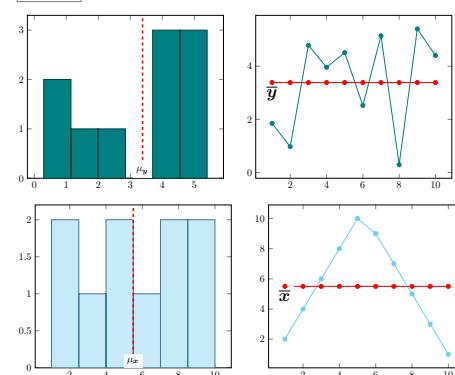
Fíjese que sumar un vector constante (paralelo a $\mathbf{1}$) no cambia la desviación típica, y que sumar un vector de media nula (perpendicular a $\mathbf{1}$) no cambia la media.



$$\sigma_{\mathbf{z}} = 0 \Leftrightarrow \mathbf{z} = a\mathbf{1}; \quad \mu_{\mathbf{z}} = 0 \Leftrightarrow \mathbf{z} \perp \mathbf{1} \quad (12)$$

F20

Recuerde que $\mu_{\mathbf{x}}$ es un valor numérico y que $\bar{\mathbf{x}}$ es un vector constante (un múltiplo de $\mathbf{1}$).



F21

Para finalizar esta sección vamos a añadir la definición de dos estadísticos más: la covarianza y la correlación.

Definición 11. La covarianza entre \mathbf{x} e \mathbf{y} es el producto escalar (de la estadística) de las proyecciones $(\mathbf{x} - \bar{\mathbf{x}})$ y $(\mathbf{y} - \bar{\mathbf{y}})$:

$$\sigma_{\mathbf{x}\mathbf{y}} = \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{y} - \bar{\mathbf{y}}))}.$$

Nótese que dado que $(\mathbf{y} - \bar{\mathbf{y}})$ es la proyección de \mathbf{y} sobre el subespacio de vectores ortogonales a los vectores constantes, $(\mathbf{y} - \bar{\mathbf{y}})$ es ortogonal a todo vector constante. En particular

$$\mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot \bar{\mathbf{y}})} = 0.$$

Y como $\mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{y} - \bar{\mathbf{y}}))} = \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot \mathbf{y})} - \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot \bar{\mathbf{y}})}$, también tenemos que

$$\sigma_{\mathbf{x}\mathbf{y}} = \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot \mathbf{y})} \quad \text{y también} \quad \sigma_x^2 = \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot \mathbf{x})},$$

pues $\sigma_{\mathbf{x}\mathbf{x}} = \sigma_x^2$. También podemos deducir un expresión alternativa para la covarianza:

P-2
(26)

$$\sigma_{\mathbf{x}\mathbf{y}} = \mu_{\mathbf{x} \odot \mathbf{y}} - \mu_{\mathbf{x}} \mu_{\mathbf{y}}. \quad (13)$$

Definición 12. Llamamos correlación entre \mathbf{x} e \mathbf{y} al coseno del ángulo formado por las proyecciones $(\mathbf{x} - \bar{\mathbf{x}})$ y $(\mathbf{y} - \bar{\mathbf{y}})$:

$$\rho_{\mathbf{x}\mathbf{y}} = \frac{\sigma_{\mathbf{x}\mathbf{y}}}{\sigma_x \sigma_y}. \quad (14)$$

Como es un coseno (véase la Sección 2.1.2) su valor siempre está comprendido entre -1 y 1 , es decir, $-1 \leq \rho_{\mathbf{x}\mathbf{y}} \leq 1$.

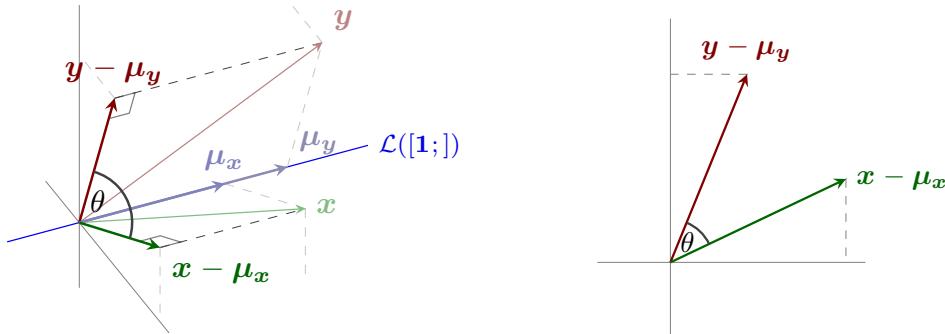


Figure 3: Perspectivas distintas representando el ángulo θ cuyo coseno es la correlación entre los vectores \mathbf{x} e \mathbf{y} .

Multiplicando y dividiendo por N obtenemos distintas expresiones para la correlación:

$$\rho_{\mathbf{x}\mathbf{y}} = \frac{N\sigma_{\mathbf{x}\mathbf{y}}}{\sqrt{(N\sigma_x^2) \times (N\sigma_y^2)}} = \frac{(\mathbf{x} - \bar{\mathbf{x}}) \cdot (\mathbf{y} - \bar{\mathbf{y}})}{\sqrt{((\mathbf{x} - \bar{\mathbf{x}}) \cdot (\mathbf{x} - \bar{\mathbf{x}})) \times ((\mathbf{y} - \bar{\mathbf{y}}) \cdot (\mathbf{y} - \bar{\mathbf{y}}))}} = \frac{\mathbf{x} \cdot (\mathbf{y} - \bar{\mathbf{y}})}{\sqrt{(\mathbf{x} \cdot (\mathbf{x} - \bar{\mathbf{x}})) \times (\mathbf{y} \cdot (\mathbf{y} - \bar{\mathbf{y}}))}}.$$

Vectores alineados: Cuando $\mathbf{y} = \lambda \mathbf{x}$ con $\lambda \neq 0$ (*cuando \mathbf{y} es un múltiplo no nulo de \mathbf{x}*) tenemos que

$$\sigma_{\mathbf{x}\mathbf{y}} = \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{y} - \bar{\mathbf{y}}))} = \mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot \lambda(\mathbf{x} - \bar{\mathbf{x}}))} = \lambda \mu_{((\mathbf{x} - \bar{\mathbf{x}})^2)} = \lambda \sigma_x^2.$$

Así, de (14), deducimos que si $\mathbf{y} = \lambda \mathbf{x}$

$$\rho_{\mathbf{x}\mathbf{y}} = \frac{\sigma_{\mathbf{x}\mathbf{y}}}{\sigma_x \sigma_y} = \frac{\lambda \sigma_x^2}{\sigma_x \sigma_{(\lambda \mathbf{x})}} = \frac{\lambda \sigma_x^2}{|\lambda| \sigma_x \sigma_x} = \frac{\lambda}{|\lambda|} = \pm 1,$$

es decir, 1 cuando λ es positivo y -1 cuando λ es negativo.

3.1.1 De la geometría a la interpretación estadística

Cuando proyectamos ortogonalmente un vector de datos \mathbf{x} sobre $\mathbf{1}$, lo descomponemos en una *componente constante* y otra *componente variable* que es perpendicular a la primera.

Denominamos *vector de medias*, $\bar{\mathbf{x}} = \mu_{\mathbf{x}} \mathbf{1}$, a la componente constante. Al ser $\bar{\mathbf{x}}$ el vector constante más próximo a \mathbf{x} , es habitual interpretar la media $\mu_{\mathbf{x}}$ como “*el valor central*” alrededor del cual se distribuyen los datos.

Por otra parte, denominamos *vector en desviaciones respecto a la media* ($\mathbf{x} - \bar{\mathbf{x}}$) a la componente variable de \mathbf{x} . La longitud de $(\mathbf{x} - \bar{\mathbf{x}})$ se denomina *desviación típica* y, de algún modo, indica la “*dispersión de los datos*” alrededor de su valor central.

Fíjese que al saltar a la interpretación estadística destacamos aspectos tales como el “valor central de los datos” o su grado de “concentración”, que son *interpretaciones subjetivas* de la información contenida en los datos, y que no son parte del formalismo matemático de las definiciones.²⁰ Exactamente lo mismo ocurre con la varianza, que es el cuadrado de una norma y cuya interpretación como medida de dispersión es, de nuevo, una interpretación o lectura subjetiva. Esto es extensivo a otros conceptos, como la probabilidad, que quedan fuera del contenido del curso.

La covarianza entre \mathbf{x} e \mathbf{y} es el producto escalar entre sus respectivas *componentes variables*, $\mu_{((\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{y} - \bar{\mathbf{y}}))} = N^{-1} \sum_i (x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})$. Es decir, es la suma de los elementos del vector $(\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{y} - \bar{\mathbf{y}})$ dividida por N . Consecuentemente, una covarianza positiva significa que en dicha suma “*dominan*” los productos positivos, es decir, “*dominan*” los casos en los que ambas desviaciones son positivas o ambas son negativas. Pero esto no quiere decir necesariamente que “*lo más frecuente*” sea que si x_i es mayor que $\mu_{\mathbf{x}}$ entonces y_i también es mayor que $\mu_{\mathbf{y}}$ (y viceversa). Bien podría ocurrir que lo más frecuente fuera lo opuesto, que en la mayoría de los casos donde $x_i > \mu_{\mathbf{x}}$ resultara que $y_i < \mu_{\mathbf{y}}$, pero que estas las diferencias respecto a los valores medios fueran tan pequeñas que la suma de los correspondientes productos fuera un número negativo pero no alejado de cero, y que sin embargo en unos pocos casos en los que $x_i \gg \mu_{\mathbf{x}}$ (en los que es mucho más grande) también ocurriera que $y_i \gg \mu_{\mathbf{y}}$ (y viceversa) dando lugar a algunos sumandos positivos muy grandes. En tal caso la suma (la covarianza) podrá resultar positiva aunque haya muchos más sumandos negativos, es decir, aunque sean mucho más frecuentes los casos en los que valores de \mathbf{x} por encima de la media de \mathbf{x} vengan acompañados de valores de \mathbf{y} por debajo de la media de \mathbf{y} .

La correlación entre \mathbf{x} e \mathbf{y} es el coseno del ángulo formado por sus *componentes variables* $(\mathbf{x} - \bar{\mathbf{x}})$ e $(\mathbf{y} - \bar{\mathbf{y}})$ (y tiene el mismo signo que la covarianza). Cuando los vectores $(\mathbf{x} - \bar{\mathbf{x}})$ e $(\mathbf{y} - \bar{\mathbf{y}})$ forman un ángulo pequeño la correlación está próxima a 1; y cuando $(\mathbf{x} - \bar{\mathbf{x}})$ e $(\mathbf{y} - \bar{\mathbf{y}})$ apuntan en direcciones casi opuestas, forman un ángulo muy abierto y la correlación está próxima a -1. Cuando la correlación es cero los vectores $(\mathbf{x} - \bar{\mathbf{x}})$ e $(\mathbf{y} - \bar{\mathbf{y}})$ son perpendiculares.

Tenga en cuenta que correlación nula no significa que \mathbf{x} e \mathbf{y} sean ortogonales; de hecho, \mathbf{x} e \mathbf{y} podrían estar casi alineados y, sin embargo, tener correlación nula (por ejemplo si, en la Figura 3, \mathbf{x} y \mathbf{y} estuvieran casi alineados con $\mathbf{1}$, pero contenidos en planos a escuadra cuya recta común son los vectores constantes, por ejemplo si \mathbf{x} está contenido en el plano horizontal e \mathbf{y} en el vertical). Por otra parte, dos vectores perpendiculares entre sí pueden tener correlación 1 en valor absoluto (por ejemplo si ambos vectores están contenidos en el mismo plano, pero forman un ángulo de 90 grados entre ellos). Por tanto, *no confunda vectores ortogonales con vectores con correlación nula*.

²⁰Recuerde también que disponemos otros modos alternativos para reflejar estos aspectos: la mediana, la moda, el coeficiente de variación, valores máximos y mínimos o rangos intercuartílicos; también representaciones gráficas como histogramas, diagramas de caja, diagramas de dispersión, etc.

3.2 Modelo Lineal Simple: ajuste MCO con una constante más un segundo regresor

(Lección 2)

T-7 Ajuste MCO con un regresor adicional a la constante

$$Y_n = a + bX_n + \text{otrascosas}_n$$

(Modelo Lineal Simple).

Las ecuaciones normales

$$\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y},$$

donde ahora

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = [\mathbf{1}; \mathbf{x}]; \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix};$$

se reducen a

$$\begin{bmatrix} (\mathbf{1} \cdot \mathbf{1}) & (\mathbf{1} \cdot \mathbf{x}) \\ (\mathbf{x} \cdot \mathbf{1}) & (\mathbf{x} \cdot \mathbf{x}) \end{bmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \mathbf{1} \cdot \mathbf{y} \\ \mathbf{x} \cdot \mathbf{y} \end{pmatrix}.$$

F22

A continuación puede ver la solución de este sistema de ecuaciones normales (correspondiente al modelo lineal simple). Su cálculo está propuesto como problema al final de la lección.

P-6
(26)

(Lección 2)

T-8 Solución para el modelo lineal simple

Para el Modelo Lineal Simple, la solución al sistema de ecuaciones normales es:

$$\hat{b} = \frac{\sigma_{\mathbf{x}} \mathbf{y}}{\sigma_{\mathbf{x}}^2} \quad (15)$$

y

$$\hat{a} = \mu_{\mathbf{y}} - \hat{b} \mu_{\mathbf{x}} \quad (16)$$

F23

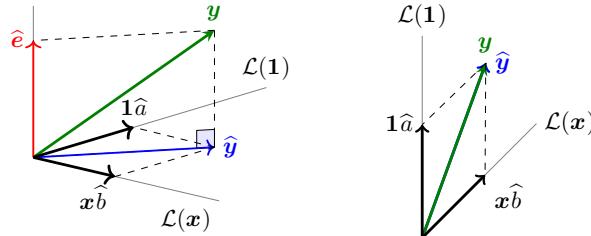
Así, cuando realizamos una regresión sobre un vector de unos más un regresor adicional, la pendiente de la recta de regresión ajustada, \hat{b} , es la covarianza entre \mathbf{x} e \mathbf{y} dividida por la varianza de \mathbf{x} , por tanto

$$\hat{b} = \frac{\sigma_{\mathbf{x}} \mathbf{y}}{\sigma_{\mathbf{x}}^2} = \frac{\mu((\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{y} - \bar{\mathbf{y}}))}{\mu((\mathbf{x} - \bar{\mathbf{x}}) \odot (\mathbf{x} - \bar{\mathbf{x}}))} = \frac{\mu(\mathbf{x} \odot (\mathbf{y} - \bar{\mathbf{y}}))}{\mu(\mathbf{x} \odot (\mathbf{x} - \bar{\mathbf{x}}))}$$

Multiplicando y dividiendo \hat{b} por $\sigma_{\mathbf{y}}$ tenemos $\hat{b} = \frac{\sigma_{\mathbf{x}} \mathbf{y}}{\sigma_{\mathbf{x}}^2} = \frac{\sigma_{\mathbf{x}} \mathbf{y}}{\sigma_{\mathbf{x}} \sigma_{\mathbf{x}}} \cdot \frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{y}}} = \frac{\sigma_{\mathbf{x}} \mathbf{y}}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}} \cdot \frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}} = \rho_{\mathbf{x} \mathbf{y}} \cdot \frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}}$; es decir

$$\hat{b} = \rho_{\mathbf{x} \mathbf{y}} \cdot \frac{\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}}.$$

Por tanto, la pendiente estimada en el Modelo Lineal Simple es el coeficiente de correlación entre el regresor \mathbf{x} y el regresando, multiplicado por la desviación típica del regresando y dividido por la desviación típica del regresor \mathbf{x} .



F24

Operando (o fijándose en la figura derecha de la transparencia anterior) se pueden deducir varios resultados. Primero:

$$\widehat{\mathbf{y}} = \widehat{a}\mathbf{1} + \widehat{b}\mathbf{x} = \bar{\mathbf{y}}. \quad (P-3) \quad (26)$$

que nos permite demostrar que la varianza de los valores ajustados es la varianza de \mathbf{x} por el cuadrado de la pendiente del ajuste \widehat{b} (que también se aprecia en la misma figura de la transparencia anterior). Es decir, que

$$\sigma_{\widehat{\mathbf{y}}}^2 = \widehat{b}^2(\sigma_{\mathbf{x}}^2) = \sigma_{b\mathbf{x}}^2. \quad (17) \quad (P-4) \quad (26)$$

Y por último, operando también llegamos a concluir una sencilla relación entre la covarianza entre \mathbf{y} y \mathbf{x} y la covarianza entre \mathbf{y} y los valores ajustados. La relación es

$$\sigma_{\mathbf{y}\widehat{\mathbf{y}}} = \widehat{b}(\sigma_{\mathbf{y}\mathbf{x}}) = \sigma_{b\mathbf{x}}. \quad (18) \quad (P-5) \quad (26)$$

Veamos un ejemplo de estimación de un modelo lineal simple²¹.

Ejemplo 3. Precio de las viviendas: precio de 14 viviendas en *University City*. San Diego, California. Año 1990. (Ramanathan, 2002, pp. 78).

■ CÓDIGO: EjPvivienda.inp

n	Precio (\mathbf{y})	Superficie (\mathbf{x})
1	199.9	1065
2	228.0	1254
3	235.0	1300
4	285.0	1577
5	239.0	1600
6	293.0	1750
7	285.0	1800
8	365.0	1870
9	295.0	1935
10	290.0	1948
11	385.0	2254
12	505.0	2600
13	425.0	2800
14	415.0	3000

Table 2: Superficie (pies al cuadrado) y precio de venta (miles de dólares)

Tratemos de ajustar por MCO los precios de las viviendas mediante la siguiente recta de regresión:

$$\widehat{\mathbf{y}} = \widehat{\beta}_1 \mathbf{1} + \widehat{\beta}_2 \mathbf{x} = [\mathbf{1}; \mathbf{x};] \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \mathbf{x}\widehat{\beta}.$$

²¹La primera práctica con ordenador del fichero *PracticasClase.pdf* reproduce este ejemplo.

Calculando

$$\begin{aligned} \mathbf{1} \cdot \mathbf{x} &= \sum_n X_n = 26753 & \mathbf{x} \cdot \mathbf{x} &= \sum_n X_n^2 = 55462515 \\ \mathbf{1} \cdot \mathbf{y} &= \sum_n Y_n = 4444.9 & \mathbf{x} \cdot \mathbf{y} &= \sum_n X_n Y_n = 9095985.5 \end{aligned}$$

y sustituyendo en el sistema de ecuaciones normales del Modelo Lineal Simple tenemos:

$$\begin{aligned} (14)\hat{a} + (26753)\hat{b} &= 4444.9 \\ (26753)\hat{a} + (55462515)\hat{b} &= 9095985.5 \end{aligned}$$

cuya solución $\hat{a} = 52.3509$ y $\hat{b} = 0.13875$ es el ajuste por mínimos cuadrados de a y b . Alternativamente, de (15), (16), y del valor de los estadísticos obtenemos idénticos valores:

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = 0.13875; \quad \hat{a} = \mu_y - \mu_x \hat{b} = 52.3509$$

Por tanto

$$\widehat{\text{precio}} = 52.35 + 0.139 (\text{sqft})$$

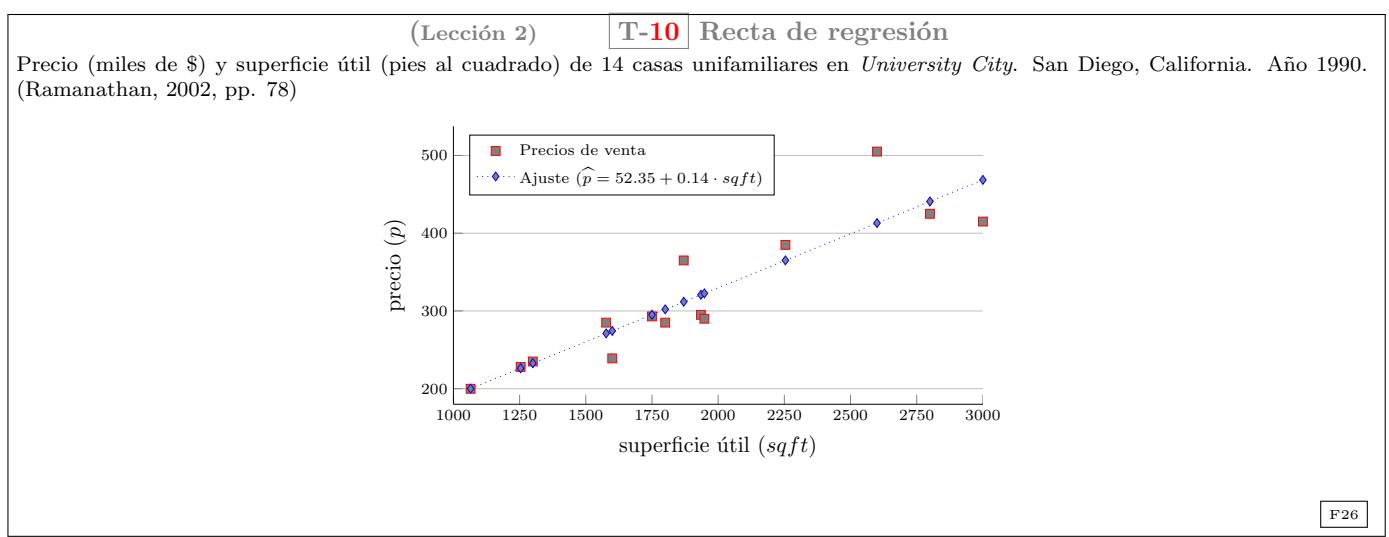
Así, por ejemplo, con esta muestra el precio de venta *ajustado* de una casa de 1800 pies cuadrados, será

$$\hat{y}_7 = 52.35 + 0.139(1800) = \mathbf{302.1} \text{ miles de dólares.}$$

No obstante, aunque la séptima casa de la muestra es de 1800 pies su precio solo fue $y_7 = \mathbf{285}$ miles de dólares. Esta discrepancia ($\hat{e}_7 = y_7 - \hat{y}_7$) puede sugerir que la casa tiene algo que ha disminuido su valor (mala ubicación, pocos servicios, mal estado, etc).²²

n	Precio	Superficie	Precio ajustado	Error \hat{e}
1	199.9	1065	200.1200	-0.22000
2	228.0	1254	226.3438	1.65619
3	235.0	1300	232.7263	2.27368
4	285.0	1577	271.1602	13.83984
5	239.0	1600	274.3514	-35.35142
6	293.0	1750	295.1640	-2.16397
7	285.0	1800	302.1015	-17.10148
8	365.0	1870	311.8140	53.18600
9	295.0	1935	320.8328	-25.83278
10	290.0	1948	322.6365	-32.63653
11	385.0	2254	365.0941	19.90587
12	505.0	2600	413.1017	91.89826
13	425.0	2800	440.8518	-15.85180
14	415.0	3000	468.6019	-53.60187

Table 3: Superficie (en pies al cuadrado), precio de venta (en miles de dólares), precios y errores estimados.



²²Esta interpretación del error de ajuste es especulativa y podría ser completamente errónea, entre otros motivos, porque el modelo ajustado fuera inadecuado. Usted siempre debe ser consciente de que la formalización matemática solo nos indica propiedades algebraicas, pero que nada dice sobre la adecuada interpretación de resultados empíricos (en eso las matemáticas nos ayudan poco).

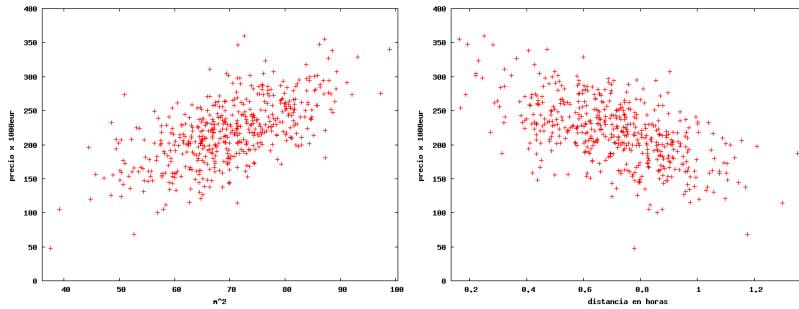
3.3 Ajuste MCO con una constante y otros dos regresores adicionales

En este caso $\mathbf{X} = \begin{bmatrix} 1 & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ 1 & x_{N2} & x_{N3} \end{bmatrix} = [\mathbf{X}_{|1}; \quad \mathbf{X}_{|2}; \quad \mathbf{X}_{|3}]$ donde $\mathbf{X}_{|1} = \mathbf{1}$; y $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \\ \widehat{\beta_3} \end{pmatrix}$.

Más abajo puede encontrar en forma de ejercicios el desarrollo algebraico de este modelo. Aquí solo vamos a ver, con un ejemplo simulado, que el ajuste lineal es un plano de regresión que “corta” una nube puntos. Para ello, ejecute el guión de Gretl de más abajo. Abra la ventana del modelo con el ajuste MCO. En el menú de “Gráficos” seleccione “gráfico de variable estimada y observada” --> “contra S y D”. Verá el plano de regresión que corta a la nube de puntos justo de la única manera en que se minimiza la suma de los errores al cuadrado del ajuste.

Ejemplo 4. Precio de las viviendas simulado (dos regresores):

Modelo simulado: $p = 100 + 3s - 130d + u$



☞ Código: [SimuladorEjPvivienda.inp](#)

En modelos con aún más regresores es imposible “visualizar” el ajuste del hiper-plano de $k - 1$ dimensiones que corta una nube de puntos en un espacio de k dimensiones. Afortunadamente no es necesaria tal visualización, basta analizar las propiedades algebraicas. Son las mismas, sea cual sea el número de regresores. Lo veremos en la Lección 3...

3.4 Ecuaciones normales del ajuste MCO con k regresores

Las matrices y vectores de las ecuaciones normales $(\mathbf{X}^\top)\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}$ en el caso general (k regresores) quedan del siguiente modo. La matriz de coeficientes $\mathbf{X}^\top\mathbf{X}$ es la matriz simétrica:

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} {}_{1|}(\mathbf{X}^\top\mathbf{X})_{|1} & {}_{1|}(\mathbf{X}^\top\mathbf{X})_{|2} & \cdots & {}_{1|}(\mathbf{X}^\top\mathbf{X})_{|k} \\ {}_{2|}(\mathbf{X}^\top\mathbf{X})_{|1} & {}_{2|}(\mathbf{X}^\top\mathbf{X})_{|2} & \cdots & {}_{2|}(\mathbf{X}^\top\mathbf{X})_{|k} \\ \vdots & \vdots & \ddots & \vdots \\ {}_{k|}(\mathbf{X}^\top\mathbf{X})_{|1} & {}_{k|}(\mathbf{X}^\top\mathbf{X})_{|2} & \cdots & {}_{k|}(\mathbf{X}^\top\mathbf{X})_{|k} \end{bmatrix}_{k \times k}$$

donde cada elemento de la matriz $\mathbf{X}^\top\mathbf{X}$ es de la forma ${}_{i|}(\mathbf{X}^\top\mathbf{X})_{|j} = (\mathbf{X})_{|i} \cdot (\mathbf{X})_{|j} = \sum_{n=1}^N x_{ni} x_{nj}$.

Si $\mathbf{X}_{|1} = \mathbf{1}$, entonces ${}_{1|}(\mathbf{X}^\top\mathbf{X})_{|1} = \mathbf{1} \cdot \mathbf{1} = N$; y ${}_{i|}(\mathbf{X}^\top\mathbf{X})_{|1} = {}_{1|}(\mathbf{X}^\top\mathbf{X})_{|i} = \mathbf{1} \cdot \mathbf{X}_{|i} = \sum_{n=1}^N x_{ni}$.

Por otra parte, el vector del lado derecho $\mathbf{X}^\top\mathbf{y}$ es

$$\mathbf{X}^\top\mathbf{y} = \begin{pmatrix} {}_{1|}\mathbf{X}^\top\mathbf{y} \\ {}_{2|}\mathbf{X}^\top\mathbf{y} \\ \vdots \\ {}_{k|}\mathbf{X}^\top\mathbf{y} \end{pmatrix} \in \mathbb{R}^k$$

donde ${}_{i|}\mathbf{X}^\top\mathbf{y} = \sum_{n=1}^N x_{ni} y_n$; y si $\mathbf{X}_{|1} = \mathbf{1}$, entonces su primer elemento es ${}_{1|}\mathbf{X}^\top\mathbf{y} = \mathbf{1} \cdot \mathbf{y} = \sum_{n=1}^N y_n$.

Problemas de la Lección 2

(L-2) PROBLEMA 1. Demuestre el Teorema de Pitágoras, es decir, que si \mathbf{a} y \mathbf{b} son perpendiculares (si $\langle \mathbf{a} | \mathbf{b} \rangle = 0$), y si $\mathbf{c} = \mathbf{a} + \mathbf{b}$, entonces $\|\mathbf{c}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$.

(L-2) PROBLEMA 2. Demuestre que $\sigma_{\mathbf{x}\mathbf{y}} = \mu_{\mathbf{x}\odot\mathbf{y}} - \mu_{\mathbf{x}}\mu_{\mathbf{y}}$

(L-2) PROBLEMA 3. Demuestre que $\bar{\mathbf{y}} = \hat{a}\mathbf{1} + \hat{b}\bar{\mathbf{x}}$.

(L-2) PROBLEMA 4. Demuestre que $\sigma_{\hat{\mathbf{y}}}^2 = \hat{b}^2(\sigma_{\mathbf{x}}^2)$.

(L-2) PROBLEMA 5. Demuestre que $\sigma_{\mathbf{y}\hat{\mathbf{y}}} = \hat{b}(\sigma_{\mathbf{y}\mathbf{x}})$

(L-2) PROBLEMA 6. Resuelva $\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$ para el caso de un regresor adicional a la constante, es decir, donde $\mathbf{X} = [\mathbf{1}; \mathbf{x}]$.

(L-2) PROBLEMA 7. ¿Cómo afectaría al ajuste MCO que el segundo regresor \mathbf{x} de un Modelo Lineal Simple fuera un vector de constantes \mathbf{c} ?

Ajuste MCO en modelos con tres regresores

(L-2) PROBLEMA 8. Obtenga el sistema de ecuaciones normales para el siguiente modelo con tres regresores: $\mathbf{y} = a\mathbf{1} + b\mathbf{x} + c\mathbf{z} + \text{otras cosas}$

(L-2) PROBLEMA 9. Obtenga la siguiente solución para el sistema de ecuaciones del ejercicio anterior.

$$\hat{a} = \mu_{\mathbf{y}} - \mu_{\mathbf{x}} \hat{b} - \mu_{\mathbf{z}} \hat{c} \quad (19)$$

$$\hat{b} = \frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{z}\mathbf{y}} - \sigma_{\mathbf{x}\mathbf{y}}\sigma_{\mathbf{z}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \quad (20)$$

$$\hat{c} = \frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{x}\mathbf{y}} - \sigma_{\mathbf{z}\mathbf{y}}\sigma_{\mathbf{x}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \quad (21)$$

(L-2) PROBLEMA 10. Si la covarianza entre \mathbf{x} y \mathbf{z} es cero en el modelo con tres regresores ¿con la estimación de qué modelo coincide la estimación de \hat{c} ?

Multicolinealidad perfecta:

(L-2) PROBLEMA 11. ¿Cómo afectaría al cálculo de los parámetros del ajuste MCO del modelo con tres regresores el hecho de que \mathbf{x} y \mathbf{z} tuvieran un coeficiente de correlación con valor absoluto igual a uno?

Fin de los Problemas de la Lección 2

Enlace a algunas **prácticas de la Lección 2**

Fin de la lección

LECCIÓN 3: Propiedades algebraicas del ajuste MCO. Medidas de ajuste

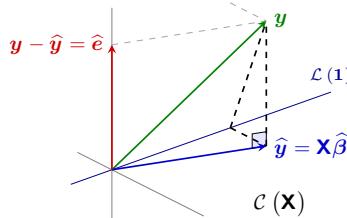
4 Propiedades algebraicas del ajuste MCO

- Capítulos 2, 3 y Apéndice de Wooldridge (2006)

(Lección 3) T-1 Geometría MCO

El Ajuste por regresión MCO es una descomposición ortogonal:

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}; \quad \text{donde} \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \perp \hat{\mathbf{e}}$$



donde los parámetros $\hat{\boldsymbol{\beta}}$ satisfacen $\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$ y $\mathbf{1} \in C(\mathbf{X})$. F29

Recuerde que dos vectores \mathbf{a} y \mathbf{b} de \mathbb{R}^N son ortogonales si, y solo si,²³ $\mathbf{a} \cdot \mathbf{b} = \sum a_i b_i = 0$. Por tanto, un vector \mathbf{a} es perpendicular a las columnas de \mathbf{X} si $\mathbf{a}^\top \mathbf{X} = \mathbf{0}$ o, lo que es lo mismo, si $\mathbf{X}^\top \mathbf{a} = \mathbf{0}$.

Dado que $\hat{\mathbf{y}}$ es la proyección de \mathbf{y} sobre las combinaciones lineales de los regresores $\mathbf{X}_{|j}$, la diferencia $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ es la proyección de \mathbf{y} sobre el subespacio de vectores perpendiculares a los regresores.²⁴ Por tanto $\hat{\mathbf{e}}$ es ortogonal a los regresores y, consecuentemente, a cualquier combinación lineal $\mathbf{X}\mathbf{v}$. En particular, los errores $\hat{\mathbf{e}}$ son ortogonales a los valores ajustados $\hat{\mathbf{y}}$ (por ser estos la combinación lineal $\mathbf{X}\hat{\boldsymbol{\beta}}$).

(Lección 3) T-2 Mínimos cuadrados ordinarios: Propiedades algebraicas

El cálculo MCO de $\hat{\boldsymbol{\beta}}$ en $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}}$ implica que (F11)

$$\hat{\mathbf{e}} \perp \mathbf{X}_{|j}, \quad \text{es decir} \quad \boxed{\hat{\mathbf{e}}^\top \mathbf{X} = 0}.$$

Y como $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, entonces $\hat{\mathbf{e}} \perp \hat{\mathbf{y}}$ (pues $\hat{\mathbf{e}} \cdot \hat{\mathbf{y}} = \hat{\mathbf{e}}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}^\top \hat{\boldsymbol{\beta}}$);

$$\text{es decir, } \hat{\mathbf{e}}^\top \mathbf{X} = 0 \Rightarrow \boxed{\hat{\mathbf{e}} \cdot \hat{\mathbf{y}} = 0} \quad (22)$$

Y como $\mathbf{1} \in C(\mathbf{X})$ tenemos que

$$\boxed{\mu_{\hat{\mathbf{e}}} = 0} \quad \text{y por tanto} \quad \boxed{\mu_{\mathbf{y}} = \mu_{\hat{\mathbf{y}}}}. \quad \text{Así que} \quad \boxed{\bar{\mathbf{y}} = \bar{\hat{\mathbf{y}}}}.$$

(Véase (F20) y la figura en (F29)) F30

Como $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$ y como $\hat{\mathbf{e}} \perp \mathbf{1}$ (pues \mathbf{X} siempre es tal que $\mathbf{1}$ es combinación de sus columnas)²⁵ tenemos que

$$\boxed{\mu_{\hat{\mathbf{e}}} = 0} \quad \text{y} \quad \boxed{\mu_{\mathbf{y}} = \mu_{\hat{\mathbf{y}}}}, \quad (23) \quad (P-1)$$

²³Si y solo si $\mu_{\mathbf{a} \odot \mathbf{b}} = 0$ si preferimos expresarlo con el producto escalar habitualmente usado en la estadística.

²⁴Es decir $C(\mathbf{X})^\perp$, la proyección sobre el complemento ortogonal del subespacio generado por los regresores $\mathbf{X}_{|j}$ (véase Bujosa, 2022b).

²⁵De hecho lo habitual es que la primera columna $\mathbf{X}_{|1}$ sea precisamente el vector constante $\mathbf{1}$.

ya que un vector tiene media nula si, y solo si, es perpendicular a las constantes (Ecuación 12 en la página 19).

4.1 Sumas de cuadrados y descomposición de la varianza

En la figura de la primera transparencia (F29) están presentes varios triángulos rectángulos²⁶. Aunque la figura destaca con colores el triángulo rectángulo cuya hipotenusa es \mathbf{y} y cuyos catetos son $\hat{\mathbf{y}}$ y $\hat{\mathbf{e}}$ (es decir, la descomposición ortogonal de \mathbf{y} en $\hat{\mathbf{y}}$ más $\hat{\mathbf{e}}$), también se destaca un segundo triángulo, *orientado perpendicularmente al vector de unos*, cuyos lados están marcados con líneas discontinuas.

Ese segundo triángulo está nuevamente representado en la siguiente transparencia (triángulo con lados rojo, azul y verde). El cateto vertical (rojo) corresponde al vector de errores de ajuste $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$. El cateto horizontal (en azul) es la diferencia $(\hat{\mathbf{y}} - \bar{\mathbf{y}})$ entre el vector $\hat{\mathbf{y}}$ ajustado por MCO y su proyección sobre los vectores constantes (es decir, su vector de medias, que como $\mathbf{1} \in \mathcal{C}(\mathbf{X})$, resulta ser $\bar{\mathbf{y}} = \bar{\mathbf{y}}$). Por último, la hipotenusa (en verde) es la diferencia $(\mathbf{y} - \bar{\mathbf{y}})$ entre el vector de datos y su proyección $\bar{\mathbf{y}}$ sobre los vectores constantes. Por tanto, en este segundo triángulo tanto la hipotenusa como los catetos están en un plano perpendicular a $\mathbf{1}$.

Esta orientación perpendicular a $\mathbf{1}$ es la que hace que la interpretación del Teorema de Pitágoras en este segundo triángulo sea interesante desde el punto de vista de la estadística y la econometría. Pero para verlo tenemos que dar significado a la longitud de los lados de este triángulo en particular.²⁷

Recuerde que disponemos de dos productos escalares en \mathbb{R}^N . El producto escalar usual ó *producto punto*: $\langle \mathbf{x} | \mathbf{y} \rangle_u = \mathbf{x} \cdot \mathbf{y}$; y el producto escalar asociado a la media aritmética: $\langle \mathbf{x} | \mathbf{y} \rangle_s = \mu_{\mathbf{x} \odot \mathbf{y}} = N^{-1}(\mathbf{x} \cdot \mathbf{y})$. Así que disponemos de “dos varas para medir” longitudes a la hora de interpretar el Teorema de Pitágoras: por una parte tenemos el cuadrado de la norma usual, $\|\mathbf{x}\|_u^2 = \mathbf{x} \cdot \mathbf{x}$, y por la otra el cuadrado de la norma en estadística, $\|\mathbf{x}\|_s^2 = \mu_{\mathbf{x}^2} = N^{-1}(\mathbf{x} \cdot \mathbf{x})$. Consecuentemente el *cuadrado de la norma usual* es N veces mayor que el cuadrado de la norma inducida por el producto escalar de la estadística

$$N\|\mathbf{x}\|_s^2 = \|\mathbf{x}\|_u^2$$

4.1.1 La norma usual y las “sumas de cuadrados”

(Lección 3)
T-3 MCO: T^a de Pitágoras y sumas de cuadrados

Como $(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \perp \hat{\mathbf{e}}$ y su suma es $(\mathbf{y} - \bar{\mathbf{y}}) = (\hat{\mathbf{y}} - \bar{\mathbf{y}}) + \hat{\mathbf{e}}$

$$\|(\mathbf{y} - \bar{\mathbf{y}})\|^2 = \|(\hat{\mathbf{y}} - \bar{\mathbf{y}})\|_u^2 + \|\hat{\mathbf{e}}\|_u^2 \quad (24)$$

Con la norma del producto escalar usual en \mathbb{R}^N

$$\underbrace{\|(\mathbf{y} - \bar{\mathbf{y}})\|_u^2}_{STC} = \underbrace{\|(\hat{\mathbf{y}} - \bar{\mathbf{y}})\|_u^2}_{SEC} + \underbrace{\|\hat{\mathbf{e}}\|_u^2}_{SRC}$$

F31

²⁶Cuatro en total, dos de ellos comparten $\hat{\mathbf{e}}$ como cateto vertical; y los otros dos comparten el vector de medias de $\bar{\mathbf{y}}$ como cateto sobre la recta de vectores constantes $\mathcal{L}(\mathbf{1})$.

²⁷Más adelante también daremos significado al coseno del ángulo entre la hipotenusa $(\mathbf{y} - \bar{\mathbf{y}})$ y el cateto horizontal $(\hat{\mathbf{y}} - \bar{\mathbf{y}})$ de este triángulo particular.

Sin mencionar explícitamente la interpretación geométrica (ni la norma utilizada), los libros de econometría dan nombre a cada uno de los lados del citado triángulo:

Suma Total de Cuadrados es el cuadrado de la norma usual del *vector de desviaciones* $\mathbf{y} - \bar{\mathbf{y}}$ de los datos \mathbf{y} respecto a su media (hipotenusa):

$$STC = \|(\mathbf{y} - \bar{\mathbf{y}})\|_u^2 = (\mathbf{y} - \bar{\mathbf{y}}) \cdot (\mathbf{y} - \bar{\mathbf{y}}) = \sum_{i=1}^N (y_i - \mu_y)^2.$$

Suma Explicada de Cuadrados es el cuadrado de la norma usual del *vector de desviaciones* $\hat{\mathbf{y}} - \bar{\mathbf{y}}$ del ajuste $\hat{\mathbf{y}}$ respecto a su media (cateto horizontal):

$$SEC = \|(\hat{\mathbf{y}} - \bar{\mathbf{y}})\|_u^2 = (\hat{\mathbf{y}} - \bar{\mathbf{y}}) \cdot (\hat{\mathbf{y}} - \bar{\mathbf{y}}) = \sum_{i=1}^N (\hat{y}_i - \mu_y)^2 \quad (\text{donde } \mu_y = \mu_{\hat{y}}).$$

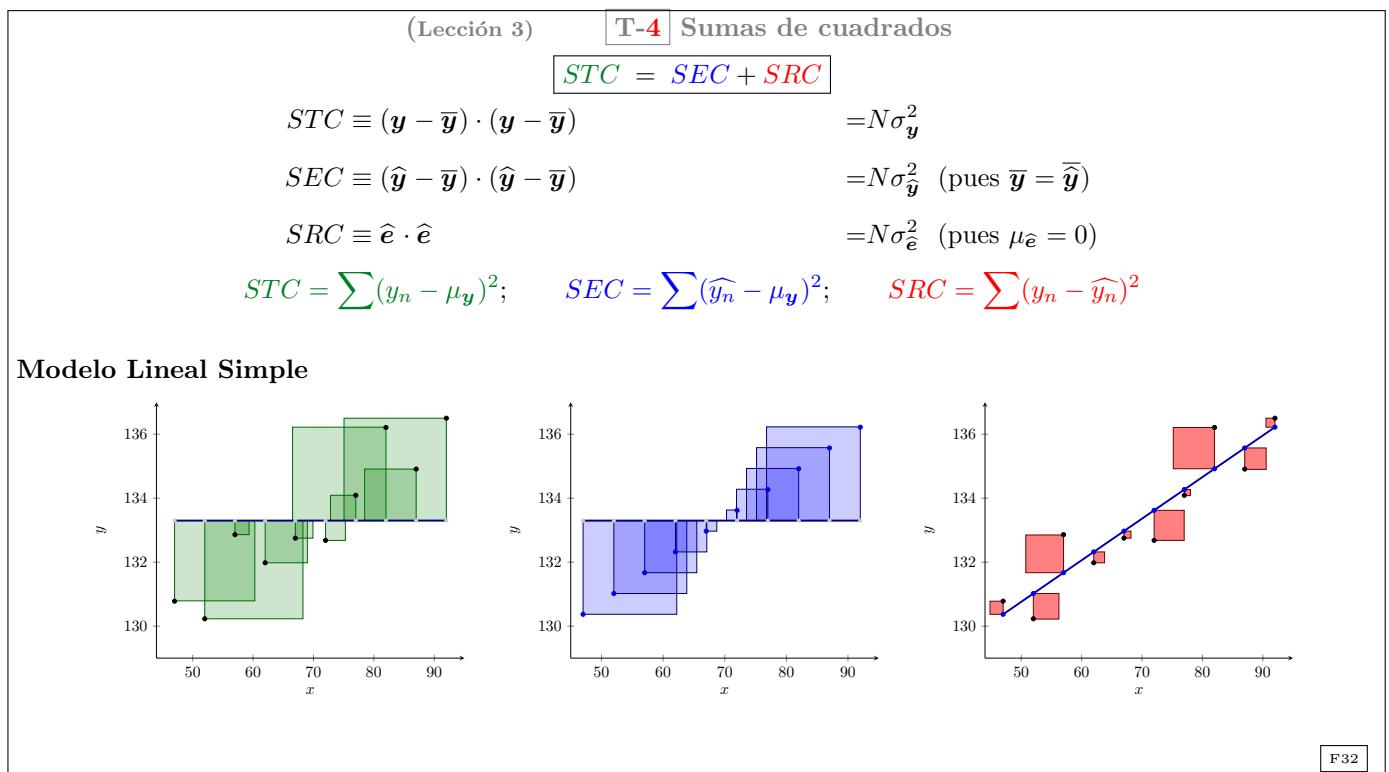
Suma de Residuos al Cuadrado es el cuadrado de la norma usual del *vector de desviaciones de errores de ajuste* $\hat{\mathbf{e}}$ respecto a su media, que es nula: $\hat{\mathbf{e}} - \bar{\hat{\mathbf{e}}} = \hat{\mathbf{e}}$ (cateto vertical).

$$SRC = \|\hat{\mathbf{e}}\|_u^2 = \hat{\mathbf{e}} \cdot \hat{\mathbf{e}} = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (e_i - \mu_e)^2 \quad (\text{donde } \mu_e = 0).$$

Dado que estamos midiendo el cuadrado de la longitud de los lados de un triángulo rectángulo, podemos acudir al Teorema de Pitágoras para concluir que

$$STC = SEC + SRC,$$

Es decir, la regresión MCO descompone la Suma Total de Cuadrados del regresando \mathbf{y} en dos partes, la Suma Explicada de Cuadrados del ajuste $\hat{\mathbf{y}}$ y la Suma de Residuos al Cuadrado.



4.1.2 La norma de la estadística y la descomposición de la varianza

En la Lección 2 vimos que la longitud (*medida con la norma de la estadística*) de la componente de \mathbf{x} ortogonal a $\mathbf{1}$ es su desviación típica (Sección 3.1). En consecuencia, el cuadrado de la longitud de cada lado del triángulo dispuesto perpendicularmente a $\mathbf{1}$ es una varianza. Así, cambiando la “vara de medir” tenemos que

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2. \quad (25)$$

La regresión descompone la varianza del regresando \mathbf{y} en la varianza del ajuste $\hat{\mathbf{y}}$ y la varianza del error de ajuste $\hat{\mathbf{e}}$.

El cambio en el modo de medir es inmediato: dado que el cuadrado de la norma usual es N veces mayor que el cuadrado de la norma de la estadística, $\|\mathbf{x}\|_u^2 = N\|\mathbf{x}\|_s^2$, tenemos que

$$STC = N\sigma_y^2; \quad SEC = N\sigma_{\hat{y}}^2; \quad SRC = N\sigma_e^2.$$

(Lección 3)
T-5 Dos varas para medir lo mismo: descomposición de la varianza

Veamos idéntica relación, pero medida con la norma de la estadística

$$\underbrace{\|(\mathbf{y} - \bar{\mathbf{y}})\|_s^2}_{\sigma_y^2} = \underbrace{\|(\hat{\mathbf{y}} - \bar{\mathbf{y}})\|_s^2}_{\sigma_{\hat{y}}^2} + \underbrace{\|\hat{\mathbf{e}}\|_s^2}_{\sigma_e^2}$$

$STC = SEC + SRC \xrightarrow{\text{dividiendo por } N} \sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2.$

F33

Fíjese que en las figuras anteriores podemos ver en total cuatro triángulos rectángulos (los ángulos rectos están marcados con pequeños cuadrados en cada uno de los correspondientes triángulos). A la relación que ya hemos visto:

$$\bullet \underbrace{(\mathbf{y} - \bar{\mathbf{y}}) \cdot (\mathbf{y} - \bar{\mathbf{y}})}_{STC} = \underbrace{(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \cdot (\hat{\mathbf{y}} - \bar{\mathbf{y}})}_{SEC} + \underbrace{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}_{SRC}, \quad \text{o bien} \quad \sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2 \quad (\text{Descomposición de la varianza}),$$

podemos añadir la de los dos triángulos que comparten el cateto: $\bar{\mathbf{y}}$ (y donde $\mu_{\mathbf{y}} = \mu_{\hat{\mathbf{y}}}$).

$$\bullet \mathbf{y} \cdot \mathbf{y} = \bar{\mathbf{y}} \cdot \bar{\mathbf{y}} + \underbrace{(\mathbf{y} - \bar{\mathbf{y}}) \cdot (\mathbf{y} - \bar{\mathbf{y}})}_{STC}, \quad \text{o bien} \quad \sigma_y^2 = \mu_{(\mathbf{y}^2)} - (\mu_{\mathbf{y}})^2, \quad (\text{Ecuación (11) en la página 18})$$

$$\bullet \hat{\mathbf{y}} \cdot \hat{\mathbf{y}} = \bar{\mathbf{y}} \cdot \bar{\mathbf{y}} + \underbrace{(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \cdot (\hat{\mathbf{y}} - \bar{\mathbf{y}})}_{SEC}, \quad \text{o bien} \quad \sigma_{\hat{y}}^2 = \mu_{(\hat{\mathbf{y}}^2)} - (\mu_{\hat{\mathbf{y}}})^2. \quad (\text{Ecuación (11) en la página 18})$$

Así como una relación que cumple la varianza de los errores debido a la descomposición ortogonal $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$:

$$\bullet \mathbf{y} \cdot \mathbf{y} = \hat{\mathbf{y}} \cdot \hat{\mathbf{y}} + \underbrace{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}_{SRC}; \quad \text{o bien} \quad \sigma_e^2 = \mu_{(\mathbf{y}^2)} - \mu_{(\hat{\mathbf{y}}^2)}.$$

En los cuatro casos, hemos aplicado el Teorema de Pitágoras, para obtener la descomposición del cuadrado de la respectiva hipotenusa (con la norma usual). Luego hemos dividido por N y hemos despejado para obtener expresiones para cada una de las tres varianzas teniendo en cuenta que $N^{-1}(\mathbf{y} \cdot \mathbf{y}) = \mu_{(\mathbf{y}^2)}$ y que $N^{-1}(\bar{\mathbf{y}} \cdot \bar{\mathbf{y}}) = (\mu_{\mathbf{y}})^2$.

Por otra parte, fíjese que los tres vectores: \mathbf{y} , $\bar{\mathbf{y}}$ e $\hat{\mathbf{y}}$ tienen idéntico vector de medias: $\bar{\mathbf{y}} = \bar{\mathbf{y}} = \hat{\mathbf{y}}$, pues el propio vector $\bar{\mathbf{y}}$ es la proyección ortogonal sobre $\mathcal{L}(\mathbf{1})$ de los tres. Dicho de otro modo: $\mu_{\mathbf{y}} = \mu_{\bar{\mathbf{y}}} = \mu_{\hat{\mathbf{y}}}$.

Comentario. El término “Suma Explicada de Cuadrados” proviene de la relación $STC = SEC + SRC$. Su intención es sugerir que el ajuste MCO descompone la variabilidad del vector \mathbf{y} en dos partes: SRC recoge la variabilidad de los residuos (aquellos que el ajuste no “explica”) y SEC recoge la variabilidad de los valores ajustados $\hat{\mathbf{y}}$ (aquella parte de la variabilidad de \mathbf{y} que se puede replicar (“explicar”) con la combinación lineal de los regresores $\hat{\mathbf{y}}$).

Pero debe tener presente que el término “explicación” es engañoso. En el ejemplo del precio de las viviendas y su superficie, es sensato suponer que los precios dependen de las características de las viviendas y , en particular, que parte de las variaciones de los precios entre distintos pisos se deben al diferente tamaño de éstos. En tal caso el nombre de “suma explicada de cuadrados” puede tener sentido (es sensato pensar que si se amplía el tamaño de una vivienda aumentará su precio).

Ahora bien, suponga que trata de ajustar una regresión lineal según la siguiente expresión:

$$\text{superficie} = \beta_1 + \beta_2 \text{precio} + (\text{otras cosas}). \quad (26)$$

Aquí la superficie aparece como función del precio de la vivienda. Si ajustamos este modelo por MCO tendremos que $STC = SEC + SRC$. No obstante, carece de sentido suponer que el tamaño de la vivienda es función del precio; de lo contrario deberíamos suponer que si la vivienda experimenta un alza o baja en su precio, su superficie aumentará o disminuirá en consecuencia.

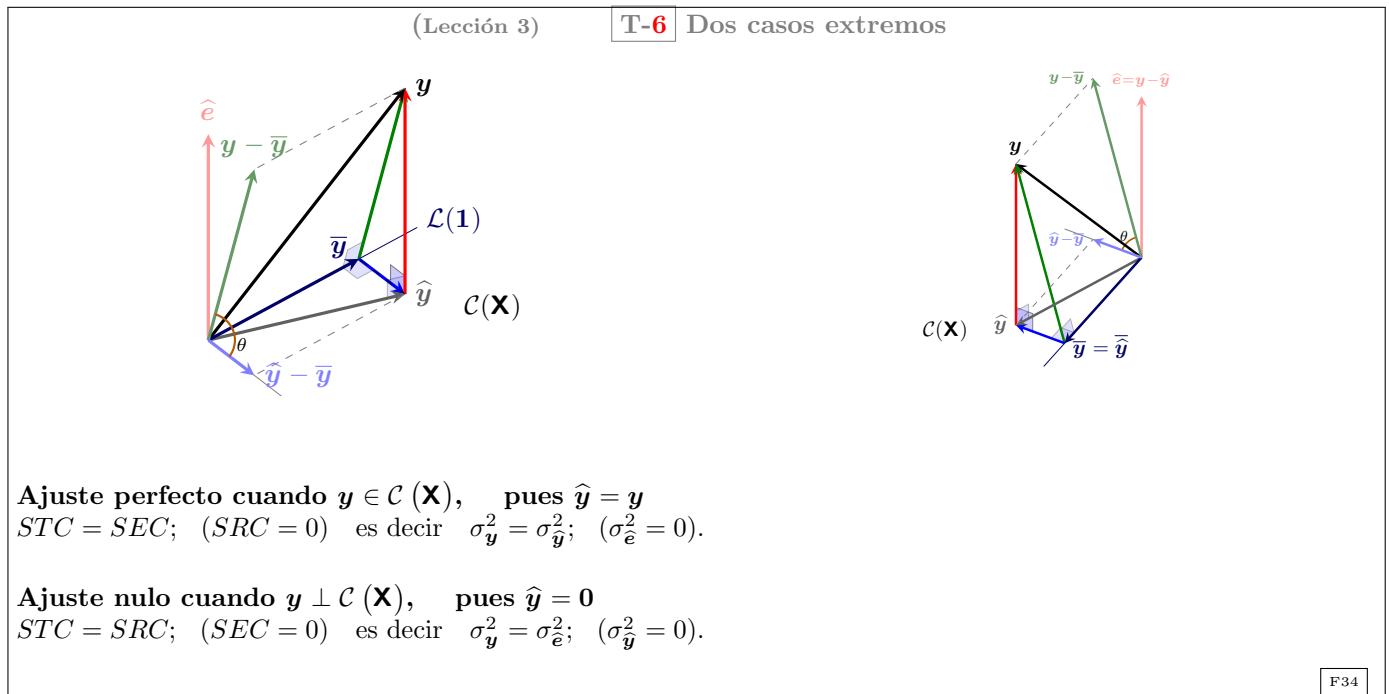
No olvide que la relación $STC = SEC + SRC$ es puramente algebraica (T^a de Pitágoras) y que su interpretación como suma de lo explicado más lo no explicado sólo puede ser aceptable si la función ajustada en la regresión “tiene sentido” desde un punto de vista de la Teoría Económica (o del “sentido común”).

La posible interpretación de los valores obtenidos mediante una regresión como la de la Ecuación 26 es de carácter exclusivamente estadístico (y no de Teoría Económica): si un piso tiene un precio muy elevado, cabe “esperar” que el piso sea grande (pero de ningún modo podemos deducir que el precio del piso “explica” su superficie).

4.2 Medidas de ajuste

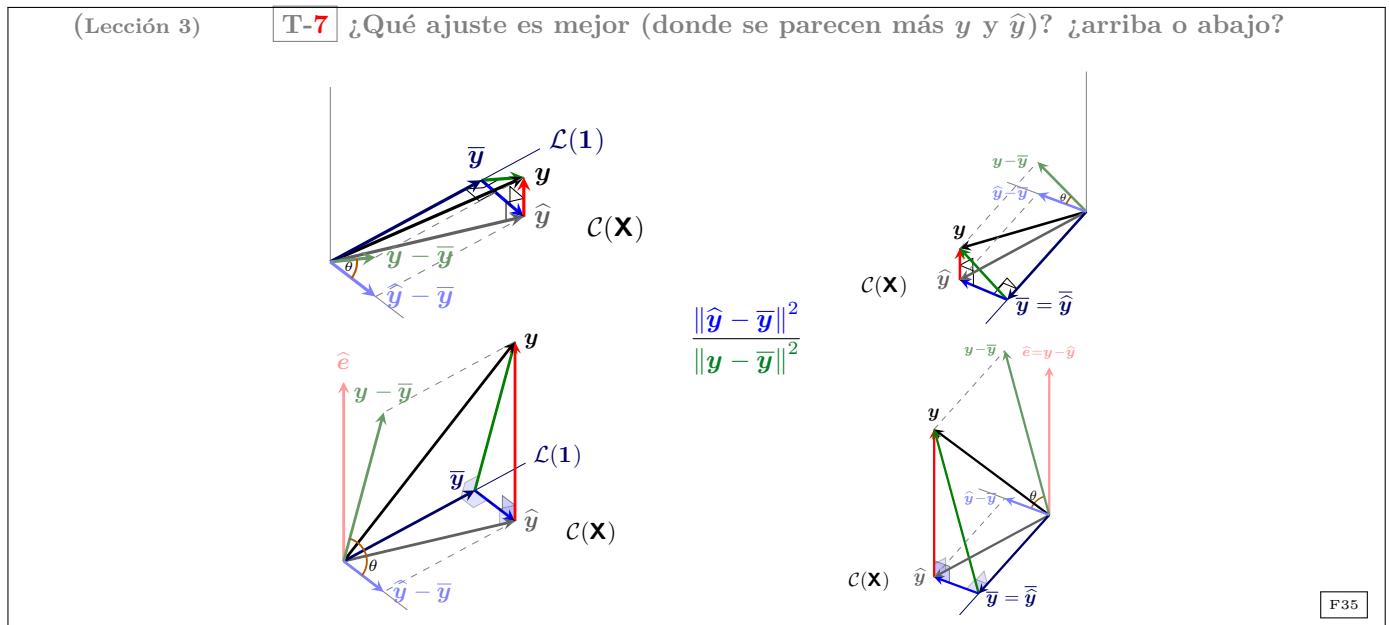
Al ajustar unos datos por MCO tenemos dos casos extremos, cuando \mathbf{y} es una combinación de los regresores resulta que el propio \mathbf{y} es la combinación de los regresores más próxima a sí mismo, por tanto $\hat{\mathbf{y}} = \mathbf{y}$. Como el el ajuste es perfecto el vector de errores es cero: $\hat{\mathbf{e}} = \mathbf{0}$.

El otro extremo sucede cuando \mathbf{y} es perpendicular a los regresores. En tal caso, la combinación de los regresores más próxima a \mathbf{y} es el vector nulo, $\hat{\mathbf{y}} = \mathbf{0}$, así que $\hat{\mathbf{e}} = \mathbf{y}$. Consecuentemente la Suma Explicada de Cuadrados es cero.



Las medidas de ajuste sirven para comparar la bondad del ajuste de modelos alternativos aplicados a un mismo

regresando \mathbf{y} . Los dos casos extremos que hemos visto (el ajuste nulo y el ajuste perfecto) nos dan una pista sobre cómo diseñar un posible criterio que permita comparar distintos ajustes.



4.2.1 Coeficiente de determinación o R^2

Dado que (en un modelo con más de un regresor, i.e., $k > 1$) el ajuste es perfecto cuando $SEC = STC$ y es nulo cuando $SRC = STC$; es fácil idear una función que tome valores entre cero (ajuste nulo) y uno (ajuste perfecto). Basta dividir el cuadrado de la norma de la componente variable del ajuste ($\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2$) por el cuadrado de la norma de la componente variable de los datos ($\|\mathbf{y} - \bar{\mathbf{y}}\|^2$)... siempre y cuando el regresando \mathbf{y} no sea constante.

Con este cociente estaremos empleando como criterio para evaluar la bondad del ajuste *la proporción de la variabilidad del regresando \mathbf{y} que es recogida por el vector ajustado $\hat{\mathbf{y}}$* .

$$R^2 = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} = \frac{SEC}{STC} = \frac{\sigma_{\hat{\mathbf{y}}}^2}{\sigma_{\mathbf{y}}^2} \quad (\text{con } \mathbf{y} \neq \mathbf{0}).$$

Esta función se denomina *coeficiente de determinación* y se denota con R^2 . Por el Teorema de Pitágoras sabemos que $\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = \|\mathbf{y} - \bar{\mathbf{y}}\|^2 - \|\hat{\mathbf{e}}\|^2$, así que alternativamente podemos expresar R^2 como

$$R^2 = \frac{\|\mathbf{y} - \bar{\mathbf{y}}\|^2 - \|\hat{\mathbf{e}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} = 1 - \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} = 1 - \frac{SRC}{STC} = 1 - \frac{\sigma_{\hat{\mathbf{e}}}^2}{\sigma_{\mathbf{y}}^2} \quad (\text{con } \mathbf{y} \neq \mathbf{0}). \quad (27)$$

Pese a que esta segunda expresión del R^2 es más larga, resulta ser la más frecuente en los manuales de Econometría (imagino que el motivo es que recuerda a la expresión de otra forma de medir el ajuste que veremos enseguida).

La notación “ R^2 ” del coeficiente de determinación proviene del hecho de que este cociente es el cuadrado de la correlación entre \mathbf{y} e $\hat{\mathbf{y}}$. Para comprobar que R^2 es $(\rho_{\hat{\mathbf{y}}\mathbf{y}})^2$ debemos recordar que $\mu_{\hat{\mathbf{y}}} = \mu_{\mathbf{y}}$ y tener en cuenta que $\hat{\mathbf{y}} \cdot \mathbf{y} = \hat{\mathbf{y}} \cdot \hat{\mathbf{y}}$ (y en consecuencia $\mu_{\hat{\mathbf{y}} \odot \mathbf{y}} = \mu_{\hat{\mathbf{y}} \odot \hat{\mathbf{y}}}$). Así, por la Ecuación 13 en la página 20 sabemos que

$$\sigma_{\hat{\mathbf{y}}\mathbf{y}} = \mu_{\hat{\mathbf{y}} \odot \mathbf{y}} - \mu_{\hat{\mathbf{y}}} \mu_{\mathbf{y}} = \mu_{\hat{\mathbf{y}} \odot \hat{\mathbf{y}}} - \mu_{\hat{\mathbf{y}}}^2 = \sigma_{\hat{\mathbf{y}}}^2.$$

Consecuentemente

$$R^2 = \frac{\sigma_{\hat{\mathbf{y}}}^2}{\sigma_{\mathbf{y}}^2} = \frac{\sigma_{\hat{\mathbf{y}}}^2}{\sigma_{\mathbf{y}}^2} \cdot \frac{\sigma_{\hat{\mathbf{y}}}^2}{\sigma_{\hat{\mathbf{y}}}^2} = \frac{(\sigma_{\hat{\mathbf{y}}}^2)^2}{\sigma_{\mathbf{y}}^2 \sigma_{\hat{\mathbf{y}}}^2} = \frac{(\sigma_{\mathbf{y}\hat{\mathbf{y}}})^2}{(\sigma_{\mathbf{y}} \sigma_{\hat{\mathbf{y}}})^2} = (\rho_{\hat{\mathbf{y}}\mathbf{y}})^2, \quad (28)$$

donde $\rho_{\hat{\mathbf{y}}\mathbf{y}} = \frac{\sigma_{\hat{\mathbf{y}}\mathbf{y}}}{\sigma_{\mathbf{y}} \sigma_{\hat{\mathbf{y}}}}$ es el coeficiente de correlación lineal simple entre $\hat{\mathbf{y}}$ e \mathbf{y} .

P-2
(36)

Y dado que la correlación de dos vectores es el coseno del ángulo formado por sus respectivas “componentes variables” (véase la Figura 3 en la página 20), también tenemos que el R^2 es el cuadrado del coseno del ángulo θ formado por el vector $(\mathbf{y} - \bar{\mathbf{y}})$ y el vector $(\hat{\mathbf{y}} - \bar{\mathbf{y}})$.²⁸

Caso especial:

El Modelo Lineal Simple. Ya sabemos (ecuaciones 17 y 18 en la página 23) que para el MLS se verifica tanto que, $\sigma_{\hat{\mathbf{y}}}^2 = \hat{b}^2 \cdot \sigma_{\mathbf{x}}^2$ como que $\sigma_{\mathbf{y}\hat{\mathbf{y}}} = \hat{b} \cdot \sigma_{\mathbf{y}\mathbf{x}}$; así pues, de (28) concluimos que en el MLS

$$R^2 = \left(\frac{\sigma_{\hat{\mathbf{y}}\mathbf{y}}}{\sqrt{\sigma_{\mathbf{y}}^2 \sigma_{\hat{\mathbf{y}}}^2}} \right)^2 = \left(\frac{\hat{b} \cdot \sigma_{\mathbf{y}\mathbf{x}}}{\sqrt{\sigma_{\mathbf{y}}^2 (\hat{b}^2 \cdot \sigma_{\mathbf{x}}^2)}} \right)^2 = \left(\frac{\hat{b} \cdot \sigma_{\mathbf{y}\mathbf{x}}}{\hat{b} \sqrt{\sigma_{\mathbf{y}}^2 \sigma_{\mathbf{x}}^2}} \right)^2 = (\rho_{\mathbf{y}\mathbf{x}})^2.$$

(Lección 3)

T-8 Medidas de ajuste

Coeficiente de determinación: R^2

$$\begin{aligned} R^2 &= \frac{SEC}{STC} &= 1 - \frac{SRC}{STC}; & 0 \leq R^2 \leq 1 \\ &= \frac{\sigma_{\hat{\mathbf{y}}}^2}{\sigma_{\mathbf{y}}^2} &= 1 - \frac{\sigma_{\hat{\mathbf{e}}}^2}{\sigma_{\mathbf{y}}^2} & = (\rho_{\hat{\mathbf{y}}\mathbf{y}})^2. \end{aligned}$$

En caso del MLS también tenemos que $R^2 = (\rho_{\mathbf{x}\mathbf{y}})^2$

Coeficiente de determinación corregido o ajustado: \bar{R}^2

$$\bar{R}^2 = 1 - \frac{\frac{\sigma_{\hat{\mathbf{e}}}^2}{\sigma_{\mathbf{y}}^2}}{1 - \frac{SRC}{STC}} = 1 - \frac{\frac{N-k}{N-1}}{\frac{STC}{N-1}} = 1 - \frac{N-1}{N-k}(1-R^2) \leq 1$$

donde $\frac{SRC}{N-k} \equiv \sigma_{\hat{\mathbf{e}}}^2$ es la *cuasi-varianza* de $\hat{\mathbf{e}}$; y $\frac{STC}{N-1} \equiv \sigma_{\mathbf{y}}^2$ es la *cuasi-varianza* de \mathbf{y}

F36

Coeficiente de determinación ajustado Llamamos modelos anidados a modelos que comparten el mismo regresando \mathbf{y} , pero en los que el conjunto de regresores de uno de los modelos es un subconjunto de los regresores del otro. Por ejemplo, son anidados el modelo que ajusta el precio de las viviendas en función de su superficie, y un segundo modelo que, además de la superficie, también incluye el número de dormitorios y cuartos de baño.

Pensemos qué ocurrirá si comparamos la bondad de ajuste de modelos anidados (y una misma muestra). El ajuste MCO de \mathbf{y} es la proyección ortogonal sobre el subespacio $\mathcal{C}(\mathbf{X})$ que contiene todos los vectores que son combinación lineal de los regresores (las columnas de \mathbf{X}). Es decir, de todos los vectores que hay en $\mathcal{C}(\mathbf{X})$, el vector $\hat{\mathbf{y}}$ es el que está más próximo a \mathbf{y} . Al añadir nuevos regresores a los ya existentes (nuevas columnas a \mathbf{X}), no desaparece ninguno de los vectores de $\mathcal{C}(\mathbf{X})$ que ya había; pero (si alguno de los nuevos regresores es linealmente independiente de los anteriores) habremos ampliado $\mathcal{C}(\mathbf{X})$ con una infinidad de nuevos vectores. En esta situación solo caben dos posibilidades: el vector más próximo a \mathbf{y} sigue siendo el mismo de antes, es decir, ninguno de los nuevos vectores de $\mathcal{C}(\mathbf{X})$ está más cerca de \mathbf{y} o bien puede que alguno de los nuevos esté más cerca de \mathbf{y} . Consecuentemente, al añadir nuevos regresores la longitud de los errores $\|\hat{\mathbf{e}}\|_s$ se mantiene (si los nuevos regresores son ortogonales a \mathbf{y}) o bien se reduce (que es lo más frecuente en la práctica).

Esto conduce a un problema práctico: siguiendo el criterio R^2 un modelo que incorpore nuevos regresores jamás será considerado peor, pues su vector de errores nunca será mayor que el correspondiente al modelo reducido.

Volvamos al ejemplo del precio de la vivienda e incorporemos algunos regresores sin sentido: el número de pares de zapatos que el arquitecto que firmó el proyecto de construcción tenía cuando se graduó. El segundo dígito del número de la licencia de conducir del conductor que llevó el primer porte de ladrillos a la obra y la presión en bares del

²⁸Fíjese que dicho ángulo θ está indicado en las figuras de la transparencia correspondiente al frame F35 en la página anterior. Mirando las figuras es inmediato darse cuenta que cuando el ángulo está próximo a cero el ajuste es bueno, y que cuando el ángulo está próximo a ser un ángulo recto el ajuste es muy malo.

neumático delantero derecho del vehículo en el que los transportó. Salvo en el remoto caso de que los nuevos regresores sean ortogonales a los precios (i.e., salvo que se dé la remotísima casualidad de que $\mathbf{y} \cdot \mathbf{X}_{\perp i} = 0$ para $i = 3, 4, 5$), el R^2 del modelo con 5 regresores (tres de ellos absurdos) será mayor que el del modelo con dos (constante y superficie).

Es evidente que no tiene ningún interés práctico preferir un modelo por el mero hecho de tener más regresores. Lo que buscamos es que los regresores incluidos en el modelo tengan capacidad predictiva (como ocurre con el tamaño cuando queremos predecir el precio de una vivienda, o la potencia del motor con el precio de un coche, o la altura con el peso de una persona, etc.). En consecuencia, para poder comparar sensatamente, es necesario penalizar la cantidad de regresores de manera se prefiera un modelo ampliado solo cuando los nuevos regresores mejoren el ajuste tanto como para compensar la penalización impuesta. La siguiente medida de ajuste penaliza el número de regresores (k).

Definición 13. El coeficiente de determinación corregido \bar{R}^2 de define como

$$\bar{R}^2 = 1 - \frac{s_{\hat{e}}^2}{s_y^2} = 1 - \frac{\frac{SRC}{N-k}}{\frac{STC}{N-1}} \quad \text{con } k > 1,$$

donde $s_{\hat{e}}^2 = \frac{SRC}{N-k}$ se denomina *cuasi-varianza de \hat{e}* y donde $s_y^2 = \frac{STC}{N-1}$ se denomina *cuasi-varianza de y* .

Como

$$\frac{s_{\hat{e}}^2}{s_y^2} = \frac{\frac{SRC}{N-k}}{\frac{STC}{N-1}} > \frac{\frac{SRC}{N}}{\frac{STC}{N}} = \frac{\sigma_{\hat{e}}^2}{\sigma_y^2} \quad (\text{cuando } k > 1),$$

resulta que $\bar{R}^2 < R^2$. Además, el coeficiente de determinación *corregido* es negativo cuando $\frac{SRC}{N-k} > \frac{STC}{N-1}$. Operando, también podemos deducir que $\bar{R}^2 = 1 - \frac{N-1}{N-k}(1 - R^2)$.

La ventaja del coeficiente de determinación *corregido* \bar{R}^2 es que *penaliza los modelos con un elevado numero de parámetros* (al corregir por el número de grados de libertad $N - k$), ello permite comparar el ajuste de modelos anidados. Muchos analistas consideran esta corrección como insuficiente. Consecuentemente se han propuesto otros criterios (*criterios de información*) que señalaremos en la Sección 15 y que penalizan aún más el número de parámetros.

Los programas econométricos suelen mostrar tanto el R^2 como el \bar{R}^2 . Por una parte el coeficiente de determinación R^2 nos indica la proporción (en tanto por uno) de la variabilidad del regresando que es captada por el ajuste; y por otra el coeficiente de determinación ajustado (o *corregido*) \bar{R}^2 nos permite comparar modelos cuando están anidados.²⁹

Fíjese en la salida de Gretl para el ejemplo del precio de 14 viviendas en *University City*. San Diego, California. Año 1990. (Ramanathan, 2002, pp. 78). En ella se pueden ver el R^2 y el \bar{R}^2 corregido.

(Lección 3)		T-9 Ajuste en el ejemplo de las casas							
Código: EjPvivienda.inp								
Estimaciones MCO utilizando las 14 observaciones 1–14									
Variable dependiente: price									
		Coeficiente	Desv. Típica	Estadístico t	Valor p				
const	52,35	37,29		1,40	0,19				
sqft	0,14	0,02		7,41	0,00				
Media de la vble. dep.	317,4929	D.T. de la vble. dep.		88,49816					
Suma de cuad. residuos	18273,57	D.T. de la regresión		39,02304					
R^2	0,820522	R^2 corregido		0,805565					
$F(1, 12)$	54,86051	Valor p (de F)		8,20e-06					
Log-verosimilitud	-70,08421	Criterio de Akaike		144,1684					
Criterio de Schwarz	145,4465	Hannan–Quinn		144,0501					

Nota práctica importante: Los coeficientes de determinación nos dan información sobre la bondad del ajuste, pero jojo! No es recomendable darles demasiada importancia. Hay aspectos más relevantes a la hora de valorar la calidad de los modelos... por ejemplo, que las predicciones del modelo sean sensatas.

²⁹También muestran otros criterios basados en modelos probabilísticos (criterios de información de Akaike y de Schwartz que nombraremos en la Sección 15, cuando hablaremos de variables aleatorias y hayamos añadido el supuesto de normalidad sobre las perturbaciones).

Práctica 1. Reproduzca en Gretl el siguiente ejemplo sobre el cálculo del coeficiente de determinación ajustado para comparar modelos con distinto número de regresores.

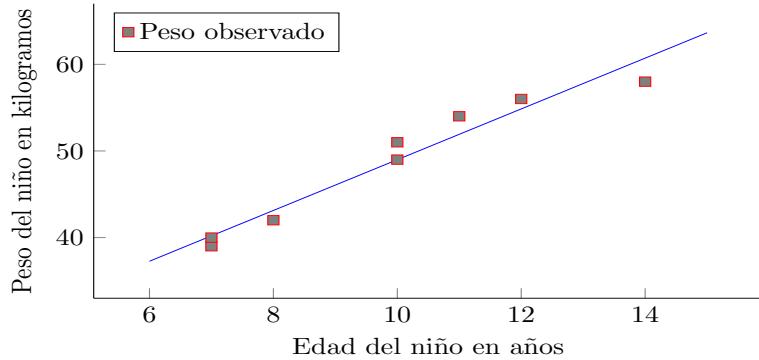
Ejemplo 5. Peso de niños según su edad:

• CÓDIGO: [PesoEdad.inp](#)

n	Peso Kg	Edad
1	39	7
2	40	7
3	42	8
4	49	10
5	51	10
6	54	11
7	56	12
8	58	14

Table 4: Peso (en kilogramos) y edad (en años)

Mod 1: $\text{peso} = \beta_1 1 + \beta_2 \text{edad} + \text{otras cosas}$



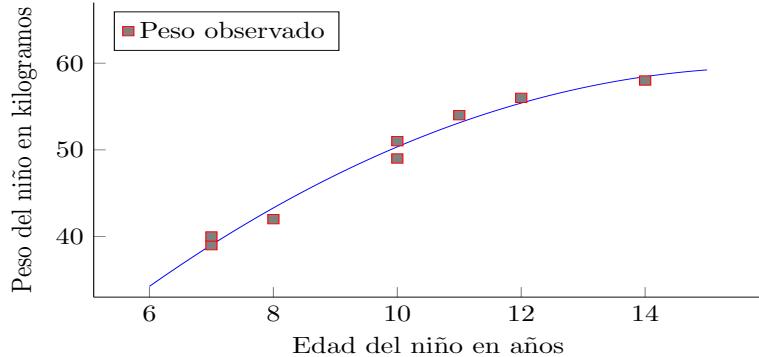
$$\widehat{\text{Peso.Kg}} = 19,6910 + 2,93003 \text{ Edad}$$

(6,999) (10,564)

$$T = 8 \quad \bar{R}^2 = 0,9405 \quad F(1,6) = 111,6 \quad \hat{\sigma} = 1,8161$$

(entre paréntesis, los estadísticos t)

Mod 2: $\text{peso} = \beta_1 1 + \beta_2 \text{edad} + \beta_3 \text{edad}^2 + \text{otras cosas}$



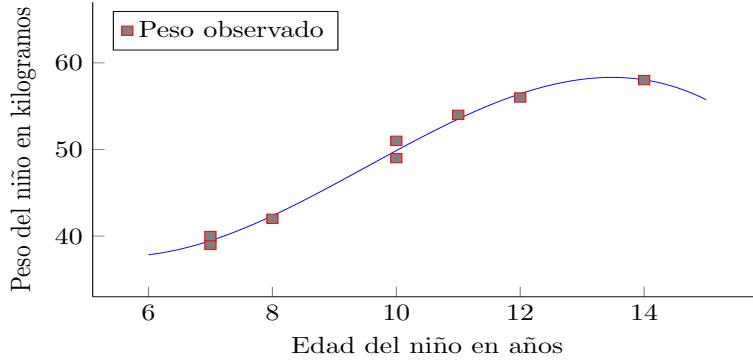
$$\widehat{\text{Peso_Kg}} = -5,11497 + 8,06835 \text{ Edad} - 0,252102 \text{ Edad}^2$$

$$(-0,664) \quad (5,159) \quad (-3,305)$$

$$T = 8 \quad \bar{R}^2 = 0,9776 \quad F(2, 5) = 153,57 \quad \hat{\sigma} = 1,1148$$

(entre paréntesis, los estadísticos t)

Mod 3: $\text{peso} = \beta_1 \text{edad} + \beta_2 \text{edad}^2 + \beta_3 \text{edad}^3 + \text{otrascosas}$



$$\widehat{\text{Peso_Kg}} = 81,7714 - 18,5964 \text{ Edad} + 2,37778 \text{ Edad}^2 - 0,0836541 \text{ Edad}^3$$

$$(1,904) \quad (-1,419) \quad (1,845) \quad (-2,043)$$

$$T = 8 \quad \bar{R}^2 = 0,9863 \quad F(3, 4) = 168,75 \quad \hat{\sigma} = 0,87188$$

(entre paréntesis, los estadísticos t)

Problemas de la Lección 3

Propiedades algebraicas del ajuste MCO

(L-3) PROBLEMA 1. Demuestre que en la descomposición $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$, si $\mu_{\hat{\mathbf{e}}} = 0$ entonces $\mu_{\mathbf{y}} = \mu_{\hat{\mathbf{y}}}$.

(L-3) PROBLEMA 2. Demuestre que $\hat{\mathbf{y}} \cdot \mathbf{y} = \hat{\mathbf{y}} \cdot \hat{\mathbf{y}}$.

Pista. Recuerde que $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$ y que $\hat{\mathbf{y}} \cdot \hat{\mathbf{e}} = 0$.

Medidas de ajuste MCO

(L-3) PROBLEMA 3. Calcule el coeficiente de determinación para un modelo en el que el único regresor es el vector de constantes. (*Pista:* piense cuanto vale SEC en este caso.)

Fin de los Problemas de la Lección 3

Enlace a algunas **prácticas de la Lección 3**

Fin de la lección

Part II

Modelo Clasico de Regresión Lineal

LECCIÓN 4: Variables aleatorias y momentos condicionados

(Léase el Capítulo 1 de Wooldridge (2006))

5 Las variables aleatorias como modelo.

Una vez vista la regresión por MCO en \mathbb{R}^N , es buen momento para avanzar otro paso y considerar un modelo para los datos que emplearemos en econometría. Dicho modelo asume que

los datos son realizaciones de variables aleatorias.

Antes de seguir debemos recordar qué significa esto. Si usted ha cursado asignaturas de estadística quizás ya lo sabe: una *variable aleatoria* es una *función*,³⁰ que asigna un número real a cada elemento (suceso elemental) de un conjunto Ω (denominado *conjunto de sucesos elementales*). Además, ciertos subconjuntos de Ω (llamados *sucesos*³¹) tienen asignada una medida (entre cero y uno) que es denominada la *probabilidad del suceso*. Así, cuando se usan variables aleatorias como modelización de los datos observados, se asume que cada observación (cada dato disponible) es una “realización” de una *variable aleatoria* (es decir, la imagen de un suceso elemental)... ¡menudo lío de nombres! ¿no?³²

Así pues, por “realización” se entiende cada uno de los posibles valores que toma la función (la variable aleatoria). Es decir, se asume que cada dato es la imagen de algún suceso elemental (algún elemento del dominio Ω). Así, de la misma forma que para la función $f(x) = x^2$ los valores 0, 1 y 4 son respectivamente la imagen de los elementos de los siguientes subconjuntos de su dominio: $\{0\}$, $\{-1, 1\}$ y $\{-2, 2\}$; en el caso de una *variable aleatoria* X , los valores (las *realizaciones*) tomados por dicha función $X(\omega)$ son las imágenes de los *sucesos elementales* $\omega \in \Omega$ que componen su dominio Ω . Y del mismo modo que si alguien nos dice que se está fijando en un punto particular de la gráfica de $f(x) = x^2$ en el que la función vale 9, esta información no permite saber si dicho punto es la imagen de $x = -3$ o de $x = 3$, conocer el valor tomado por una variable aleatoria generalmente tampoco nos permite saber de qué suceso es la imagen dicho valor.

La estadística pone nombres especiales³³ a conceptos que son de uso común en otras áreas de las matemáticas: llama *realizaciones* a los valores tomados por una función (a la que llama *variable aleatoria*), y al dominio de la función lo llama *conjunto de sucesos elementales*. Toda esta nomenclatura específica de la estadística puede hacer pensar que las variables aleatorias, sus realizaciones, los sucesos y la probabilidad de dichos sucesos tienen que ver algo con la “realidad”, pero no es así. Tan solo son construcciones intelectuales (meros objetos matemáticos). Su utilidad radica en que dichas construcciones mentales se emplean como modelos abstractos para preguntar cuestiones como ¿qué podemos concluir si interpretamos el lanzamiento de un dado “como si fuera una variable aleatoria” con determinada distribución de probabilidad?... y comparar lo que se deduce del modelo matemático con lo que se observa en la naturaleza. Si ese “modelo abstracto” describe las observaciones, entonces lo consideraremos un modelo adecuado.



Figure 4: (*Restos de una edificación visigoda*). En la naturaleza no hay ángulos rectos, rectas paralelas, planos, etc..

De manera similar, objetos matemáticos tales como rectas, ángulos rectos o planos no existen en la naturaleza. Así, en una habitación “con forma rectangular”, ni el suelo de la habitación es perfectamente plano, ni los ángulos de las esquinas son exactamente rectos, ni las paredes enfrentadas son perfectamente y paralelas entre sí (por no decir que las mediciones que tomemos tampoco serán exactas). Consecuentemente, cuando medimos el área de una habitación multiplicando la longitud de dos paredes adyacentes, estamos empleando datos tomados de un objeto real y actuando como si el objeto verificará las propiedades abstractas (e irreales) del modelo matemático. Con los modelos probabilísticos ocurre algo similar. No se deje persuadir por la sugerente nomenclatura de la estadística: la realidad y los objetos matemáticos están en esferas completamente distintas.

³⁰obsérvese lo poco atinado del nombre... se denomina “variable” aleatoria a lo que realmente es una función

³¹Fíjese que un *suceso elemental* es un elemento de Ω , pero que un *suceso* es un subconjunto de Ω .

³²La presentación tradicional de la probabilidad ni emplea una nomenclatura clara ni ayuda a interpretar la regresión.

³³y en mi opinión... confusos

Puesto que los datos numéricos de naturaleza económica son resultado de alguna serie de acontecimientos acaecidos en la economía (acontecimientos que de manera “difusa” asociamos mentalmente con un supuesto conjunto de posibles eventos o sucesos... no todos igualmente verosímiles), parece natural emplear variables aleatorias en los modelos económico, pues las variables aleatorias asignan “probabilidades” a “sucesos”.

Así, podemos emplear una variable aleatoria como “modelo” para el volumen diario de ventas en un establecimiento comercial. Por experiencia sabemos que algunos eventos raramente suceden, y que otros son más verosímiles. Si consideramos un establecimiento muy grande (como El Corte Inglés del Paseo de la Castellana) es muy muy raro que el volumen de ventas en una jornada laboral ascienda a tan solo 15 céntimos de euro (pues los acontecimientos que pudieran dar lugar a ese resultado son muy remotos). Que a lo largo de una jornada ocurra algo que suponga un volumen de ventas de solo 15 mil euros también es muy raro aunque más verosímil que un volumen de 15 céntimos. Así, al modelar con una variable aleatoria las ventas diarias de El Corte Inglés, se asume que los sucesos de Ω “asociados” a circunstancias muy inusuales deben tener asignada una probabilidad (una medida) muy pequeña.

Es decir, mediante una abstracción matemática, se intenta modelizar la verosimilitud de cada volumen de ventas con la probabilidad de ciertos subconjuntos de un conjunto abstracto Ω . Pero... ¿quién es el conjunto Ω para “el volumen de ventas de El Corte Inglés del Paseo de la Castellana”? ¿El conjunto de potenciales acontecimientos es finito? Si lo es, ¿se pueden enumerar todos sus elementos? Si es infinito ¿se puede describir formalmente?... No. ¡En el fondo la asociación entre potenciales acontecimientos de una jornada y Ω resulta ser completamente vaga! En matemáticas se trata con conjuntos de números, conjuntos de funciones, etc. ¿Es el conjunto de potenciales acontecimientos de la jornada un “conjunto matemático”?... ¿Son rectas las paredes de la construcción visigótica de la fotografía?

Entender que una variable aleatoria asocia cada posible acontecimiento en torno al funcionamiento del establecimiento con un determinado volumen de ventas es erróneo. La variable aleatoria no puede realizar dicha asociación porque no hay definida ninguna asociación entre un conjunto abstracto Ω y la lista de posibles acontecimientos de una jornada comercial. La citada asociación tan solo es una vaga y ambigua asociación mental que no podemos asimilar como una variable aleatoria. Sin embargo, denominar a Ω como el *conjunto de sucesos elementales* induce a pensar que si enumeráramos los elementos de Ω estaríamos enumerando los posibles acontecimientos que podrían afectar a las ventas del establecimiento comercial. Pensar así es similar a asumir que, por el mero hecho de calcular el área de una estancia usando las longitudes de las paredes, estamos dando por hecho que la construcción visigoda de la figura de más arriba está formada por estancias perfectamente rectangulares.

Esta absoluta vaguedad en la asociación entre los “hechos” del mundo real y el conjunto abstracto de *sucesos elementales* Ω queda habitualmente oculta en los libros de estadística. En ellos siempre se muestran ejemplos triviales: lanzamientos de monedas, de dados, etc. En tales ejemplos, una asociación entre un conjunto reducido de posibles resultados físicos y sus correspondientes valores numéricos es inmediata. Por ello estos ejemplos son tremadamente engañosos; inducen a pensar que los objetos matemáticos empleados (probabilidad del suceso cara o del suceso cruz, variable aleatoria, valor esperado del lanzamiento, etc.) son parte de la realidad en lugar de abstracciones matemáticas.

En cualquier caso (y una vez lanzadas estas advertencias) a partir de ahora usaremos como modelo teórico que

los datos son realizaciones de variables aleatorias.

5.1 Relación con las lecciones anteriores (Parte I).

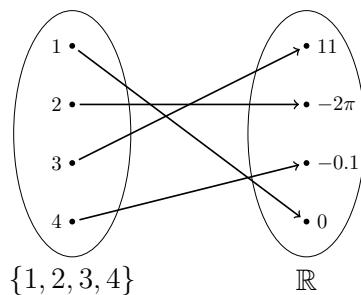
En las lecciones anteriores hemos empleado un tipo muy particular de funciones. Me refiero a los vectores de \mathbb{R}^N . Los vectores son un sistema (una lista ordenada) de números reales. Por ejemplo, el vector $\mathbf{x} = (\pi, -5, 0)$, es la lista *ordenada* \mathbf{x} cuyo primer elemento es π , el segundo es -5 y el tercero es 0 :

$$x_1 = \pi, \quad x_2 = -5 \quad \text{y} \quad x_3 = 0.$$

Por tanto, \mathbf{x} es una función definida sobre el conjunto de índices $\{1, 2, 3\}$ que asocia el índice 1 con el número π , el índice 2 con el -5 y el 3 con el 0. Visto así, \mathbb{R}^3 es un subespacio formado por funciones definidas sobre el conjunto de índices $\{1, 2, 3\}$ (es decir, si llamáramos *variable aleatoria* al vector \mathbf{x} entonces el conjunto Ω sería $\{1, 2, 3\}$).

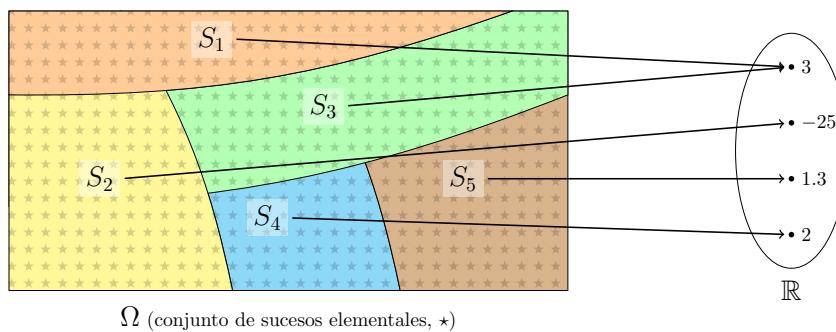
Lo relevante para esta lección es que, al igual que el conjunto de listas de tres números es el espacio vectorial \mathbb{R}^3 , el conjunto de variables aleatorias definidas sobre un conjunto de sucesos Ω también es un espacio vectorial (quien sea Ω resulta irrelevante en la práctica.³⁴). Pero disponer de un espacio vectorial no es suficiente, también necesitaremos semi-productos escalares con los que definir las proyecciones ortogonales con las que lograr construcciones similares a las descritas en las primeras lecciones.

³⁴De hecho, dos variables aleatorias pueden tener idéntica distribución, aunque el conjunto de sucesos elementales asociado a cada una de ellas sean completamente distintos. Siendo así, ya no debería sorprendernos que en general, nunca se describa explícitamente el conjunto Ω .

Vector de \mathbb{R}^4 : $(0, -2\pi, 11, -0.1)$ 

- El conjunto de vectores de \mathbb{R}^4 es un espacio vectorial
- Cada vector es una función que va de $\{1, 2, 3, 4\}$ a \mathbb{R}

F43



- El conjunto de VAs con varianza es un espacio vectorial
- Cada vector (VA) es una función que va del conjunto de sucesos elementales Ω a \mathbb{R}
- Sobre ciertos subconjuntos (S_i) se define una medida de probabilidad.

F44

6 Definición de Espacio Euclídeo de Probabilidad³⁵

La visión tradicional de que las variables aleatorias son funciones que asignan valores a distintos “sucesos elementales” y que los “sucesos” tienen asignada una probabilidad no es lo central en el modelo de regresión. Lo fundamental es tener presente que las variables aleatorias³⁶ son vectores de un espacio vectorial que tiene definido un semi producto escalar que nos permite tratar con proyecciones ortogonales (tal como hicimos en las primeras lecciones con los vectores de \mathbb{R}^N).

En esta sección se define el contexto en el que vamos a trabajar: los *espacios semi-euclídeos de probabilidad*. Se basa en una axiomatización del espacio de Hilbert de las variables aleatorias de varianza finita. Esta axiomatización alternativa de la probabilidad no es la que aparece en los manuales de probabilidad, aunque es equivalente.³⁷ La ventaja de este enfoque es que nos centra en lo verdaderamente relevante para un curso de regresión.

³⁵En esta sección (y las siguientes) realizaré una exposición que se aleja de lo tradicional pues emplea constantemente referencias al álgebra lineal. La justificación de por qué la probabilidad se puede acometer como parte del álgebra lineal requiere de una larga exposición que lamentablemente no cabe en el curso, por ello, aquí solo expondré las ideas, pero me saltaré los detalles (que son muchos y son extensos).

³⁶cuando tienen varianza definida.

³⁷Los detalles quedan fuera de los objetivos del curso.

Definición 6. Diremos que un espacio semi-euclídeo³⁸ (\mathcal{E}, η) es de probabilidad si cumple los siguientes axiomas.³⁹

A1 Existe un conjunto Ω tal que \mathcal{E} es un subespacio vectorial de⁴⁰ \mathbb{R}^Ω (es decir, \mathcal{E} es un espacio funcional).

A2 Para todo $\mathbf{X}, \mathbf{Y} \in \mathcal{E}$ se cumple que si $\mathbf{X}^2 - \mathbf{Y}^2 \geq 0 \Rightarrow \sqrt{\mathbf{X}^2 - \mathbf{Y}^2} \in \mathcal{E}$.

A3 Para todo $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_1, \mathbf{Y}_2 \in \mathcal{E}$ se cumple que $\mathbf{X}_1 \mathbf{Y}_1 = \mathbf{X}_2 \mathbf{Y}_2 \Rightarrow \langle \mathbf{X}_1 | \mathbf{Y}_1 \rangle_\eta = \langle \mathbf{X}_2 | \mathbf{Y}_2 \rangle_\eta$.⁴¹

A4 Para toda sucesión (\mathbf{X}_n) formada por funciones positivas de \mathcal{E} que sea creciente y de Cauchy, se cumple que la función \mathbf{X} definida por

$$\mathbf{X}(\omega) = \begin{cases} \sup_{n \in \mathbb{N}} \mathbf{X}_n(\omega) & \text{si } \{\mathbf{X}_n(\omega) \mid n \in \mathbb{N}\} \text{ está acotado} \\ 0 & \text{en caso contrario} \end{cases} \quad (\text{con } \omega \in \Omega)$$

pertenece a \mathcal{E} y $\lim_{n \rightarrow \infty} \|\mathbf{X}_n - \mathbf{X}\|_\eta = 0$.⁴²

A5 $\mathbf{1} \in \mathcal{E}$ ⁴³ y $\|\mathbf{1}\|_\eta = 1$.

Llamaremos *variables aleatorias* (con segundos momentos definidos) a los vectores de \mathcal{E} .

El hecho de que \mathcal{E} sea un subconjunto de \mathbb{R}^Ω nos dota de la operación producto punto a punto de \mathbb{R}^Ω , que es empleada tanto por A2 como por A3. El Axioma A3 equivale a decir que el conjunto de pares ordenados

$$\left\{ (\mathbf{XY}, \langle \mathbf{X} | \mathbf{Y} \rangle_\eta) \mid \mathbf{X}, \mathbf{Y} \in \mathcal{E} \right\}$$

es una función. Consecuentemente podemos factorizar el semi Producto Escalar como composición de dos funciones: Un producto punto a punto entre funciones, y una suma (o integral de Lebesgue) de la función que resulta tras el producto. Si denotamos el conjunto de pares ordenados como la función E_η y denotamos con $L_\mathcal{E}$ el dominio de dicha función (es decir, $L_\mathcal{E} = \{\mathbf{XY} \mid \mathbf{X}, \mathbf{Y} \in \mathcal{E}\}$), obtenemos el siguiente diagrama comutativo:

$$\begin{array}{ccc} \mathcal{E} \times \mathcal{E} & \xrightarrow{\mathbf{X} \cdot \mathbf{Y}} & L_\mathcal{E} \\ & \searrow \langle \mathbf{X} | \mathbf{Y} \rangle_\eta & \downarrow E_\eta(\mathbf{X} \cdot \mathbf{Y}) \\ & & \mathbb{R} \end{array} \quad \begin{array}{ccc} L_2(\Omega, \mathcal{F}, P) \times L_2(\Omega, \mathcal{F}, P) & \xrightarrow{\mathbf{X} \cdot \mathbf{Y}} & L_1(\Omega, \mathcal{F}, P) \\ & \searrow \langle \mathbf{X} | \mathbf{Y} \rangle_{L_2} & \downarrow E(\mathbf{X} \cdot \mathbf{Y}) \\ & & \mathbb{R} \end{array}$$

Figure 5: Diagrama del semi producto escalar η entre vectores de \mathcal{E} (es decir, entre variables aleatorias con segundos momentos definidos). Ambos diagramas muestran lo mismo, pero el de la derecha emplea la notación usual en la literatura. Compárese con la Figura 2 en la página 16.

La función $E_\eta : L_\mathcal{E} \rightarrow \mathbb{R}$ se llama *esperanza del producto*. Por claridad, deberíamos referirnos a ella como la *esperanza del producto asociada al semiproducto interior* η ,⁴⁴ pero para no complicar la notación del curso omitiré el subíndice.

³⁸Es decir, \mathcal{E} es un subespacio y $\langle \cdot | \cdot \rangle_\eta : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}$ es un semi producto escalar, es decir, verifica los axiomas de

- Simetría: $\langle \mathbf{X} | \mathbf{Y} \rangle_\eta = \langle \mathbf{Y} | \mathbf{X} \rangle_\eta$.
- Linealidad respecto al primer argumento:
 1. $\langle a\mathbf{X} | \mathbf{Y} \rangle_\eta = a \langle \mathbf{Y} | \mathbf{X} \rangle_\eta$.
 2. $\langle (\mathbf{X} + \mathbf{Y}) | \mathbf{Z} \rangle_\eta = \langle \mathbf{X} | \mathbf{Z} \rangle_\eta + \langle \mathbf{Y} | \mathbf{Z} \rangle_\eta$.
- Positivo: $\langle \mathbf{X} | \mathbf{Y} \rangle_\eta \geq 0$.

La única diferencia entre un producto escalar y un semi producto escalar es que el semi producto escalar \mathbf{X} puede ser distinta de la función cero incluso cuando $\langle \mathbf{X} | \mathbf{X} \rangle_\eta = 0$.

³⁹Estos 5 axiomas caracterizan que (\mathcal{E}, η) es el “espacio de Hilbert” (realmente semi-espacio de Hilbert) $L_2(\Omega, \mathcal{F}, P)$ conocido como *conjunto de variables aleatorias con segundos momentos finitos*, ($E(\mathbf{X}^2) < \infty$); y donde (como veremos a continuación) el semi producto escalar es la esperanza del producto: $\langle \mathbf{X} | \mathbf{Y} \rangle_\eta = E(\mathbf{XY})$, siendo la esperanza matemática una integral de Lebesgue (véase la Figura 5).

⁴⁰Por una parte, el conjunto Ω se denomina *Conjunto de sucesos elementales*; y por otra, el conjunto \mathbb{R}^Ω es el conjunto de funciones que van de Ω a \mathbb{R} , es decir, $\mathbb{R}^\Omega = \{\mathbf{X} : \Omega \rightarrow \mathbb{R}\}$. El conjunto \mathbb{R}^Ω , dotado de la suma y el producto de escalas, es un espacio vectorial.

⁴¹Es decir, $\mathbf{X}_1 \mathbf{Y}_1 = \mathbf{X}_2 \mathbf{Y}_2 \Rightarrow E(\mathbf{X}_1 \mathbf{Y}_1) = E(\mathbf{X}_2 \mathbf{Y}_2)$.

⁴²Es decir, la sucesión \mathbf{X}_n converge en *media cuadrática* a \mathbf{X} ; que normalmente se denota como: $(\mathbf{X}_n) \xrightarrow{m.s.} \mathbf{X}$

⁴³Donde $\mathbf{1}$ es la función constante 1.

⁴⁴Consecuentemente, en la Figura 2, la media $\mu_{\mathbf{x} \odot \mathbf{y}}$ es la *esperanza del producto asociada al producto escalar de la estadística s*. Por otra parte, indicar a qué producto escalar corresponde la esperanza sería la forma de indicar qué función de densidad (o de cuantía) se está usando en cada momento.

Así pues, y dado que el producto escalar en \mathcal{E} es la *esperanza del producto*, tenemos que la *longitud de una variable aleatoria* $\mathbf{Y} \in \mathcal{E}$ es $\|\mathbf{Y}\|_\eta = \sqrt{\langle \mathbf{Y} | \mathbf{Y} \rangle_\eta} = \sqrt{\mathbb{E}(\mathbf{Y}^2)}$. Por el axioma A5, y como $\mathbf{1} = \mathbf{1} \cdot \mathbf{1}$, tenemos que⁴⁵

$$\mathbb{E}(\mathbf{1}) = \mathbb{E}(\mathbf{1} \cdot \mathbf{1}) = \langle \mathbf{1} | \mathbf{1} \rangle_\eta = \|\mathbf{1}\|_\eta^2 = \|\mathbf{1}\|_\eta \|\mathbf{1}\|_\eta = 1.$$

El axioma A4 tiene el efecto de maridar la convergencia en semi-norma con la convergencia puntual (cuestiones que no veremos en el curso). Por último, A5 es específico de la teoría de la probabilidad.

Aunque los cinco axiomas son necesarios para construir la probabilidad, en esta lección explotaremos principalmente A1 (i.e., las variables aleatorias son vectores de \mathbb{R}^Ω), A3 (pues el semi Producto escalar está asociado a una función que llamamos esperanza) y A5.

6.1 Esperanza matemática

La *esperanza matemática*, \mathbb{E} , es una función definida en un conjunto más grande que el espacio semi-euclídeo de probabilidad \mathcal{E} (las variables aleatorias con varianza definida). En concreto, el dominio de la *esperanza matemática* es el conjunto $L_{\mathcal{E}}$ de variables aleatorias que se pueden obtener mediante el producto (punto a punto) de dos variables aleatorias de \mathcal{E} :

$$L_{\mathcal{E}} = \{Z \mid Z = X \cdot Y; \text{ donde } X, Y \in \mathcal{E}\}.$$

Este conjunto⁴⁶ es, dado un Ω y un \mathcal{E} , el conjunto de variables aleatorias que tienen primer momento finito; es decir, aquellas variables aleatorias Z para las que existe la esperanza de su valor absoluto⁴⁷ ($\mathbb{E}(|Z|) < \infty$).

Puesto que $\mathbf{1}$ pertenece a \mathcal{E} , resulta que si $X \in \mathcal{E}$ entonces $X \in L_{\mathcal{E}}$, pues $X = X \cdot \mathbf{1}$. Dicho de otro modo $\mathcal{E} \subset L_{\mathcal{E}}$. Consecuentemente, toda variable con varianza definida, tiene esperanza definida (el recíproco no es cierto, hay variables aleatorias en $L_{\mathcal{E}}$ que no pertenecen a \mathcal{E}).

Como hemos indicado más arriba, la *esperanza matemática* está ligada al semi Producto escalar⁴⁸ $\langle - | - \rangle_\eta$ de \mathcal{E} pues, para toda Z tal que $Z = XY$, donde $X, Y \in \mathcal{E}$, su *esperanza matemática* es $\mathbb{E}(Z) = \mathbb{E}(XY) = \langle X | Y \rangle_\eta$.

Dado que $\mathcal{E} \subset L_{\mathcal{E}}$, para toda X de \mathcal{E} tenemos que

$$\mathbb{E}(X) = \mathbb{E}(X \cdot \mathbf{1}) = \langle X | \mathbf{1} \rangle_\eta.$$

Puesto que el producto escalar $\langle - | - \rangle_\eta$ es lineal en la primera componente, la *esperanza matemática* también es lineal:⁴⁹

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Ahora vuelva a la media aritmética (Página 15) y compruebe hasta qué punto son similares la esperanza matemática y la media aritmética. Más adelante veremos el motivo.

6.2 Subespacios probabilísticos

El Álgebra Lineal trata con subespacios vectoriales; que son subconjuntos de un espacio vectorial que son espacios vectoriales por si mismos (un espacio vectorial contenido dentro de otro). Aquí también podemos definir *subespacios probabilísticos* (que son un *espacio euclídeo probabilístico* contenido en otro). Siendo (\mathcal{E}, η) un *espacio euclídeo probabilístico*, diremos que $\mathcal{S} \subset \mathcal{E}$ es un *subespacio probabilístico* de \mathcal{E} si $(\mathcal{S}, \eta|_{\mathcal{S} \times \mathcal{S}})$ es un *espacio euclídeo probabilístico*.⁵⁰

En Álgebra Lineal los subespacios se generan a partir de un conjunto de vectores. En concreto, el subespacio vectorial $\mathcal{L}(\vec{x}_1, \dots, \vec{x}_k)$ generado por el conjunto de vectores $\{\vec{x}_1, \dots, \vec{x}_k\}$ es que “*es el más pequeño de los espacios vectoriales que los contiene*”; es decir, de todos los subespacios que contienen a los vectores $\vec{x}_1, \dots, \vec{x}_k$, el subespacio vectorial

⁴⁵De manera similar a lo que ocurría con los vectores $\mathbf{1}$: $\mu_1 = N^{-1}(\mathbf{1} \cdot \mathbf{1}) = \langle \mathbf{1} | \mathbf{1} \rangle_s = \|\mathbf{1}\|_s^2 = \|\mathbf{1}\|_s \|\mathbf{1}\|_s = 1$.

⁴⁶Que se denota en la literatura con $L_1(\Omega, \mathcal{F}, P)$.

⁴⁷Además $\|X\|_E = \mathbb{E}(|Z|)$ es la “vara de medir” en $L_{\mathcal{E}}$. (Si en algún momento meto temas de convergencia en el curso, debería extender esto para hablar de convergencia en media (la de $L_{\mathcal{E}}$ usando $\|X\|_E$) y convergencia en media cuadrática (la de \mathcal{E} usando $\|X\|_\eta$)).

⁴⁸Que se suele denominar *esperanza del producto*: $\langle X | Y \rangle_\eta = \mathbb{E}(XY)$;

⁴⁹Esto demuestra que la esperanza es lineal en \mathcal{E} ; realmente es lineal en todo su dominio $L_{\mathcal{E}}$ pero, como no vamos a estudiar la esperanza en detalle, con esto nos basta.

⁵⁰Con $\eta|_{\mathcal{S} \times \mathcal{S}}$ nos referimos al semi Producto escalar η restringido al subconjunto \mathcal{S} .

$\mathcal{L}(\vec{x}_1, \dots, \vec{x}_k)$ es el más pequeño, pues contiene las combinaciones lineales de ellos y ningún otro vector más. En consecuencia, y dado que la intersección de subespacios vectoriales es un subespacio vectorial, podemos caracterizar $\mathcal{L}(\vec{x}_1, \dots, \vec{x}_k)$ como la intersección de todos los subespacios vectoriales \mathcal{S}_i que contienen los vectores $\vec{x}_1, \dots, \vec{x}_k$

$$\mathcal{L}(\vec{x}_1, \dots, \vec{x}_k) = \bigcap \{\mathcal{S}_i \text{ es subespacio tal que } \{\vec{x}_1, \dots, \vec{x}_k\} \subset \mathcal{S}_i\}.$$

De manera análoga se puede definir un *subespacio probabilístico* $\mathcal{E}(\mathbf{X}_1, \dots, \mathbf{X}_k)$ a partir de un conjunto de variables aleatorias $\{\mathbf{X}_1, \dots, \mathbf{X}_k\}$. Dicho *subespacio probabilístico* es el más pequeño de todos los *espacios euclídeos probabilísticos* que contienen a dichas variables aleatorias. Es más, dado que la intersección de *subespacios probabilísticos* también resulta ser un *subespacio probabilístico*; podemos definir $\mathcal{E}(\mathbf{X}_1, \dots, \mathbf{X}_k)$ de manera similar

$$\mathcal{E}(\mathbf{X}_1, \dots, \mathbf{X}_k) = \bigcap \{\mathcal{S}_i \text{ es subespacio probabilístico tal que } \{\mathbf{X}_1, \dots, \mathbf{X}_k\} \subset \mathcal{S}_i\}.$$

Definición 14. Llamaremos sistema de variables aleatorias a una lista ordenada de k variables aleatorias:

$$\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k].$$

Empelando el operador selector “|” podemos denotar su j -ésima componente así: $\mathbf{X}_{|j}$. Es decir, $\mathbf{X}_{|j} \equiv \mathbf{X}_j$.

Así, con $\mathcal{L}(\mathbf{Z})$ denotamos al menor subespacio vectorial que contiene las variables aleatorias del sistema (de la lista) \mathbf{Z} ; y con $\mathcal{E}(\mathbf{Z})$ denotamos al menor subespacio probabilístico que contiene las variables aleatorias de \mathbf{Z} .

Pese a todas las similitudes entre $\mathcal{L}(\mathbf{Z})$ y $\mathcal{E}(\mathbf{Z})$, hay una importante diferencia. Un conjunto de k vectores linealmente independientes genera un subespacio de dimensión k (pues solo contiene las combinaciones lineales de esos k vectores); sin embargo, un conjunto de k variables aleatorias linealmente independientes puede generar un *subespacio probabilístico* mucho más grande, incluso de dimensión infinita (pues, además de las combinaciones lineales, también debe contener las variables aleatorias indicadas en los axiomas 2, 4 y 5 de la Definición 6 en la página 40... ¡y generalmente resultan ser muchísimas!). Así, si \mathbf{Z} es una lista de variables aleatorias

$$\mathcal{L}(\mathbf{Z}) \subset \mathcal{E}(\mathbf{Z}); \quad \text{pero en general} \quad \mathcal{E}(\mathbf{Z}) \not\subset \mathcal{L}(\mathbf{Z}).$$

Es decir, el subespacio probabilístico generado por un conjunto de variables aleatorias $\mathcal{E}(\mathbf{Z})$ normalmente es muchísimo más grande que el conjunto de combinaciones lineales $\mathcal{L}(\mathbf{Z})$. No obstante, hay casos en los que $\mathcal{E}(\mathbf{Z})$ y $\mathcal{L}(\mathbf{Z})$ son exactamente iguales.

Un ejemplo se da cuando las funciones indicatrices⁵¹ $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_N} \in \mathcal{E}$ son tales que $\mathbb{1}_{A_1} + \dots + \mathbb{1}_{A_N} = \mathbb{1}_\Omega = \mathbb{1}$ (es decir, cuando $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ con $A_i \cap A_j = \emptyset$ para $i \neq j$),⁵² entonces $\mathcal{E}(\{\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_N}\}) = \mathcal{L}(\{\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_N}\})$ y la dimensión de dicho subespacio es igual al número de funciones indicatrices no nulas. Por tanto $\mathcal{E}(\mathbb{1}) = \mathcal{L}(\mathbb{1})$ es una recta (dimensión 1).

7 Esperanza condicional

En esta exposición, tan alejada de la tradicional, estoy omitiendo los detalles (por ejemplo, he dicho que la intersección de dos subespacios de probabilidad es un subespacio de probabilidad pero no lo he demostrado). Pero todo lo expuesto es correcto y demostrable (incluido que los cinco axiomas de la definición de espacio euclídeo de probabilidad son una axiomatización alternativa de la teoría de la probabilidad). Sin embargo, en esta sección voy a cometer una incorrección en aras de no complicar las cosas. Es la misma incorrección que se comete en la inmensa mayoría de manuales, pues en ellos se dice que la esperanza condicional (o esperanza condicional estocástica) es “una” variable aleatoria. Esto no es del todo correcto y en un apéndice al final de la lección lo veremos.

Consideremos una variable aleatoria \mathbf{Y} del espacio euclídeo de probabilidad \mathcal{E} y un subespacio probabilístico $\mathcal{S} \subset \mathcal{E}$. Podemos preguntarnos ¿cuál de las variables aleatorias de \mathcal{S} es la que está más próxima a \mathbf{Y} ?⁵³

⁵¹Llamamos función indicatriz de un subconjunto A de Ω , que denotamos por $\mathbb{1}_A$, a la función que para todo ω que pertenece a A toma el valor 1 y para todo ω que NO pertenece a A toma el valor 0. Es decir $\begin{cases} \mathbb{1}_A(\omega) = 1 & \omega \in A \\ \mathbb{1}_A(\omega) = 0 & \omega \notin A \end{cases}$. Por tanto, la función indicatriz del conjunto vacío, $\mathbb{1}_\emptyset$, es la función constante cero; y la función indicatriz $\mathbb{1} = \mathbb{1}_\Omega$ es la función constante uno. Puesto que es fácil confundir a primera vista los símbolos $\mathbb{1}_\emptyset$ y $\mathbb{1}_\Omega$, denotaré $\mathbb{1}_\emptyset$ con $\mathbf{0}$ (es decir, $\mathbf{0}$ es la función que asigna 0 a todo elemento de Ω). Nótese que las funciones indicatrices son idempotentes: $(\mathbb{1}_A)^2 = \mathbb{1}_A \cdot \mathbb{1}_A = \mathbb{1}_A$ (ya que $0 \cdot 0 = 0$ y $1 \cdot 1 = 1$).

⁵²Entonces se dice que los subconjuntos A_1, \dots, A_n son una partición de Ω

⁵³Resulta que ésta es una pregunta con trampa ya que no necesariamente hay una única variable aleatoria que sea la que esté más

La respuesta es que dicha variable aleatoria⁵⁴ es la *proyección ortogonal de \mathbf{Y} sobre el subespacio de probabilidad \mathcal{S}* . Dicha variable aleatoria se denomina *esperanza condicional*:

Definición 15. Sea \mathcal{E} un espacio euclídeo probabilístico, sea $\mathbf{Y} \in \mathcal{E}$ y sea \mathbf{X} un sistema de k variables aleatorias de \mathcal{E} . Llamaremos *esperanza condicional de \mathbf{Y} condicionada a \mathbf{X}* , $\mathbb{E}(\mathbf{Y} | \mathbf{X})$, a la proyección ortogonal de \mathbf{Y} sobre $\mathcal{L}(\mathbf{X})$.

Por tanto, la esperanza condicional $\mathbb{E}(\mathbf{Y} | \mathbf{X})$ es la variable aleatoria⁵⁵ de $\mathcal{L}(\mathbf{X})$ que está a distancia mínima de \mathbf{Y} .

Fíjese que la definición de más arriba no es la definición habitual de la esperanza condicional en los cursos de estadística o econometría. Pero su interpretación resulta más sencilla, pues es análoga a lo visto en lecciones anteriores con los vectores de \mathbb{R}^N . Esto nos permitirá emplear los mismos argumentos geométricos de las lecciones anteriores.⁵⁶

Recordatorio sobre la proyección ortogonal. Recuerde (Definición 6 en la página 9)⁵⁷ que la proyección ortogonal de un vector \vec{y} sobre un subespacio \mathcal{V} es una función lineal que nos arroja aquel vector de \mathcal{V} tal que el vector $\vec{y} - \text{Pr}_{\mathcal{V}}(\vec{y})$ es ortogonal a \vec{y} . Y recuerde que dicha proyección $\text{Pr}_{\mathcal{V}}(\vec{y})$ resulta ser el vector de \mathcal{V} más próximo a \vec{y} . Los dos entornos que nos interesan en el curso son:

- en \mathbb{R}^N (i.e., con listas de números), el ajuste *mínimo cuadrático* $\hat{\mathbf{y}}$ es la proyección ortogonal del vector \mathbf{y} sobre $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^N$ (el subespacio generado por las columnas de la matriz de regresores \mathbf{X}) es el vector de $\mathcal{C}(\mathbf{X})$ tal que $(\mathbf{y} - \hat{\mathbf{y}})$ es ortogonal a \mathbf{y} . Por tanto, $\hat{\mathbf{y}}$ es el vector de $\mathcal{C}(\mathbf{X})$ más próximo a \mathbf{y} .
- en un espacio euclídeo de probabilidad \mathcal{E} (i.e., con variables aleatorias), la esperanza condicional $\mathbb{E}(\mathbf{Y} | \mathbf{X})$ es la proyección ortogonal de \mathbf{Y} sobre $\mathcal{L}(\mathbf{X}) \subset \mathcal{E}$; es decir, es la variable aleatoria de $\mathcal{L}(\mathbf{X})$ tal que $(\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X}))$ es ortogonal a \mathbf{Y} . Por tanto, $\mathbb{E}(\mathbf{Y} | \mathbf{X})$ es la variable aleatoria de $\mathcal{L}(\mathbf{X})$ más próxima a \mathbf{Y} .⁵⁸

Consecuentemente, y dado que las proyecciones ortogonales son funciones lineales (Definición 6 en la página 9), la esperanza condicionada de una combinación lineal es una combinación lineal de esperanzas condicionadas: $\mathbb{E}(a\mathbf{X} + b\mathbf{Y} | \mathbf{Z}) = a\mathbb{E}(\mathbf{X} | \mathbf{Z}) + b\mathbb{E}(\mathbf{Y} | \mathbf{Z})$.

7.1 Momentos de variables aleatorias.

Tal como hicimos en la Lección 2, vamos a exponer algunos momentos, pero ahora en el contexto genérico de los espacios euclídeos de probabilidad (es conveniente comparar ambas exposiciones, verá hasta qué punto son análogas... listas de números en las lecciones anteriores y variables aleatorias ahora).

Definición 16. La *esperanza* de la variable aleatoria $\mathbf{Y} \in \mathcal{E}$ es su producto escalar con el vector constante $\mathbf{1}$:

$$\mathbb{E}(\mathbf{Y}) = \langle \mathbf{Y} | \mathbf{1} \rangle_{\eta} = \mathbb{E}(\mathbf{Y} \cdot \mathbf{1}).$$

Véase la Figura 5 en la página 41 y compárese esta definición con la Definición 7 en la página 15 y verá que la media aritmética en \mathbb{R}^N es similar a la esperanza en \mathcal{E} .

Por tanto, una variable aleatoria tiene esperanza nula si, y solo si, es perpendicular a las variables aleatorias constantes:

$$\mathbb{E}(\mathbf{Y}) \Leftrightarrow \mathbf{Y} \perp \mathbf{1}.$$

Dado que el producto escalar es lineal en la primera componente, la esperanza también es lineal, y por tanto $\mathbb{E}(a\mathbf{X} + b\mathbf{Y}) = a\mathbb{E}(\mathbf{X}) + b\mathbb{E}(\mathbf{Y})$.

Recuerde que con la media ocurría lo mismo: $\mu_{(ax+by)} = a\mu_x + b\mu_y$

próxima... puede haber infinitas a la misma distancia mínima. Ahí radica la incorrección indicada en el párrafo anterior (lo veremos en el apéndice a esta lección).

⁵⁴En realidad es el conjunto (la clase de equivalencia) formado por todas las variables aleatorias que están a esa distancia mínima por estar entre ellas a distancia cero las unas respecto de las otras (lo veremos en el apéndice a esta lección).

⁵⁵En realidad la clase de equivalencia (como se explicará en el apéndice).

⁵⁶De esta manera seguiremos evitando “feas” expresiones con integrales (o sumatorios) y funciones de densidad (o de cuantía) conjunta, marginal o condicionada. Fíjese que tampoco hemos necesitado (ni falta que hace) distinguir entre variables discretas o continuas. Sólo nos centramos en qué tipo de objetos son aquellos con los que trabajamos y cuáles son sus interrelaciones. Este enfoque evita las definiciones basadas en cómo se realizan los cálculos si la variable es discreta o si es continua (en mi opinión, centrarse en el cálculo con integrales o sumatorios nos distrae la atención hacia cuestiones distintas de las relaciones fundamentales entre los objetos matemáticos).

⁵⁷O véase la Lección “Proyecciones sobre subespacios” del *Curso de Álgebra Lineal*.

⁵⁸En realidad la esperanza condicionada es la clase de equivalencia contenida en $\mathcal{L}(\mathbf{X})$ tal que la clase $\mathbb{E}(\mathbf{Y} \cdot (\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X})))$ es ortogonal a \mathbf{Y} , es decir, clase de equivalencia más próxima a \mathbf{Y} (véase el apéndice).

La siguiente definición se corresponde con el vector de medias $\bar{\mathbf{y}}$ de la Lección 2:

Definición 17. La esperanza de \mathbf{Y} condicionada a $\mathbf{1}$, $\mathbb{E}(\mathbf{Y}|\mathbf{1})$, es la proyección ortogonal de \mathbf{Y} sobre $\mathcal{L}(\mathbf{1})$.

Por tanto, $\mathbb{E}(\mathbf{Y}|\mathbf{1})$ es una variable aleatoria constante, es decir, $\mathbb{E}(\mathbf{Y}|\mathbf{1}) = a\mathbf{1}$. Para ver cuál es el valor de a , basta tener en cuenta que $\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})$ es perpendicular a $\mathbf{1}$:

$$\mathbb{E}(\mathbf{1} \cdot (\mathbf{Y} - a\mathbf{1})) = 0 \Leftrightarrow \mathbb{E}(\mathbf{Y} - a\mathbf{1}) = 0 \Leftrightarrow \mathbb{E}(\mathbf{Y}) - a\mathbb{E}(\mathbf{1}) = 0 \Leftrightarrow a = \mathbb{E}(\mathbf{Y}),$$

pues $\mathbb{E}(\mathbf{1}) = 1$. Por tanto, la esperanza de \mathbf{Y} es el valor por el que hay que multiplicar la variable aleatoria $\mathbf{1}$ para obtener la variable aleatoria constante más próxima a \mathbf{Y} : $\mathbb{E}(\mathbf{Y}|\mathbf{1}) = \mathbb{E}(\mathbf{Y}) \cdot \mathbf{1}$. Consecuentemente $\mathbb{E}(\mathbf{1}|\mathbf{1}) = \mathbf{1}$; y además $\mathbb{E}(a\mathbf{X} + b\mathbf{Y}|\mathbf{1}) = a\mathbb{E}(\mathbf{X}|\mathbf{1}) + b\mathbb{E}(\mathbf{Y}|\mathbf{1}) = (a\mathbb{E}(\mathbf{X}) + b\mathbb{E}(\mathbf{Y}))\mathbf{1}$.

Fíjese en la analogía con el vector de medias $\bar{\mathbf{y}}$, que es la proyección ortogonal de $\mathbf{y} \in \mathbb{R}^N$ sobre $\mathcal{L}(\mathbf{1})$, es decir, el vector constante $\bar{\mathbf{y}} = \mu_{\mathbf{y}}\mathbf{1}$, donde $\mu_{\mathbf{y}} = \langle \mathbf{y} | \mathbf{1} \rangle_s$ es el producto escalar entre $\mathbf{1}$ e \mathbf{y} . También en \mathbb{R}^N tenemos que $\bar{\mathbf{1}} = \mathbf{1}$; y que $\bar{ax + by} = a\bar{x} + b\bar{y} = (a\mu_x + b\mu_y)\mathbf{1}$.

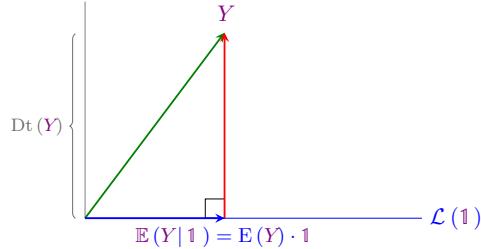
Puesto que $\mathbb{E}(\mathbf{Y}|\mathbf{1})$ es la proyección ortogonal de \mathbf{Y} sobre $\mathcal{L}(\mathbf{1})$, la diferencia $(\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1}))$ es la proyección de \mathbf{Y} sobre $\mathcal{L}(\mathbf{1})^\perp$ (sobre el complemento ortogonal del subespacio generado por $\mathbf{1}$). Así que la esperanza de dicha diferencia es necesariamente cero:

$$\mathbb{E}(\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})) = \mathbb{E}(\mathbf{Y} - \mathbb{E}(\mathbf{Y}) \cdot \mathbf{1}) = \mathbb{E}(\mathbf{Y}) - \mathbb{E}(\mathbf{Y})\mathbb{E}(\mathbf{1}) = \mathbb{E}(\mathbf{Y}) - \mathbb{E}(\mathbf{Y})\mathbf{1} = 0.$$

(de igual modo que $\mu_{(\mathbf{y} - \bar{\mathbf{y}})} = 0$)

(Lección 4)

T-3 Geometría de los momentos teóricos: Esperanza y varianza



$$\mathbb{E}(\mathbf{Y}|\mathbf{1}) = \mathbb{E}(\mathbf{Y}) \cdot \mathbf{1}$$

$$\text{Var}(\mathbf{Y}) = \mathbb{E}((\mathbf{Y} - \mathbb{E}(\mathbf{Y})\mathbf{1})^2) = \|(\mathbf{Y} - \mathbb{E}(\mathbf{Y})\mathbf{1})\|_\eta^2$$

Por Pitágoras:

$$\mathbb{E}(\mathbf{Y}^2) = \text{Var}(\mathbf{Y}) + \mathbb{E}(\mathbb{E}(\mathbf{Y})^2 \cdot (\mathbf{1})^2) = \text{Var}(\mathbf{Y}) + \mathbb{E}(\mathbf{Y})^2$$

y por tanto $\boxed{\text{Var}(\mathbf{Y}) = \mathbb{E}(\mathbf{Y}^2) - \mathbb{E}(\mathbf{Y})^2}$

F45

Resumiendo, al proyectar ortogonalmente una variable aleatoria \mathbf{Y} de \mathcal{E} sobre $\mathbf{1}$, la descomponemos en una componente constante $\mathbb{E}(\mathbf{Y}|\mathbf{1})$ (la proyección de \mathbf{Y} sobre $\mathcal{L}(\mathbf{1})$) y una segunda componente, $\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})$, que es perpendicular a la primera (la proyección de \mathbf{Y} sobre $\mathcal{L}(\mathbf{1})^\perp$) y que por contraposición voy a llamar *componente variable*.

La longitud de la componente constante es $\|\mathbb{E}(\mathbf{Y}|\mathbf{1})\|_\eta = \|\mathbb{E}(\mathbf{Y}) \cdot \mathbf{1}\|_\eta = |\mathbb{E}(\mathbf{Y})| \cdot 1$. Fijémonos ahora en la longitud de la componente variable $\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})$.

Definición 18. La desviación típica de \mathbf{Y} es la longitud (la norma) de la proyección $\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})$:

$$Ds(\mathbf{Y}) = \|\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})\|_\eta = \sqrt{\mathbb{E}([\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{1})]^2)} = \sqrt{\mathbb{E}([\mathbf{Y} - \mathbb{E}(\mathbf{Y})\mathbf{1}]^2)}.$$

Dado que las proyecciones ortogonales son funciones lineales, la proyección de $a\mathbf{Y}$ sobre $\mathcal{L}(\mathbf{1})^\perp$ es a veces la proyección de \mathbf{Y} , por tanto $\boxed{Ds(a\mathbf{Y}) = |a| Ds(\mathbf{Y})}$, pues la desviación típica es una longitud.

De manera análoga a lo descrito en la figura (F20), sumar a una variable aleatoria otra variable aleatoria constante (i.e., un múltiplo de $\mathbf{1}$) no cambia la desviación típica, pero sí su esperanza

$$\text{Dt}(\mathbf{Y} + a\mathbf{1}) = \text{Dt}(\mathbf{Y}) \quad \text{y} \quad \mathbb{E}(\mathbf{Y} + a\mathbf{1}) = \mathbb{E}(\mathbf{Y}) + a;$$

y al sumar una variable aleatoria con esperanza nula (i.e., perpendicular a $\mathbf{1}$) no cambia la esperanza

$$\mathbb{E}(\mathbf{X}) = 0 \Rightarrow \mathbb{E}(\mathbf{Y} + \mathbf{X}) = \mathbb{E}(\mathbf{Y}) + \mathbb{E}(\mathbf{X}) = \mathbb{E}(\mathbf{Y});$$

aunque de la desviación típica de la suma nada podemos decir, pues no podemos saber con seguridad si ha cambiado la longitud de la componente de $\mathbf{Y} + \mathbf{X}$ perpendicular a $\mathbf{1}$ (para visualizar esto quizás le ayude mirar la figura del recuadro (F20) considerando que los vectores representan variables aleatorias, pues en dicha representación $\sigma_{\mathbf{y}} = \sigma_{\mathbf{y} + \mathbf{z}}$).

Definición 19. La varianza de \mathbf{Y} es el cuadrado de la longitud (la norma) de la proyección $\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})$:

$$\text{Var}(\mathbf{Y}) = \|\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})\|_{\eta}^2 = (\text{Dt}(\mathbf{Y}))^2 = \mathbb{E}([\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})]^2) = \mathbb{E}([\mathbf{Y} - \mathbb{E}(\mathbf{Y}) \mathbf{1}]^2).$$

Como al multiplicar \mathbf{Y} por a su desviación típica se multiplica por $|a|$, tenemos que $\boxed{\text{Var}(a\mathbf{Y}) = a^2 \text{Var}(\mathbf{Y})}$. Sin embargo, como sumar una variable aleatoria constante no cambia la desviación típica, tampoco cambia la varianza $\boxed{\text{Var}(\mathbf{Y} + a\mathbf{1}) = \text{Var}(\mathbf{Y})}$;

Al proyectar ortogonalmente \mathbf{Y} sobre $\mathbf{1}$ descomponemos la variable aleatoria en una componente constante, $\mathbb{E}(\mathbf{Y} | \mathbf{1}) = \mathbb{E}(\mathbf{Y}) \cdot \mathbf{1}$, y una componente variable, $\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})$, perpendicular a la primera. Así, por el Teorema de Pitágoras, tenemos que $\|\mathbf{Y}\|_{\eta}^2 = \|\mathbb{E}(\mathbf{Y}) \cdot \mathbf{1}\|_{\eta}^2 + \|\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})\|_{\eta}^2$, es decir, $\mathbb{E}(\mathbf{Y}^2) = (\mathbb{E}(\mathbf{Y}))^2 + \text{Var}(\mathbf{Y})$; por tanto

$$\text{Var}(\mathbf{Y}) = \mathbb{E}(\mathbf{Y}^2) - (\mathbb{E}(\mathbf{Y}))^2.$$

Definición 20. La covarianza entre \mathbf{X} e \mathbf{Y} es el producto escalar de las proyecciones $\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})$ e $\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})$:

$$\begin{aligned} \text{Cov}(\mathbf{X}, \mathbf{Y}) &= \mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})) \cdot (\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1}))) \\ &= \mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})) \cdot \mathbf{Y}) \\ &= \mathbb{E}(\mathbf{XY}) - \mathbb{E}(\mathbf{X}) \mathbb{E}(\mathbf{Y}) \end{aligned} \tag{P-1}$$

Siguiendo argumentos similares a los anteriores: $\text{Cov}(a\mathbf{X} + b\mathbf{1}, c\mathbf{Y} + d\mathbf{1}) = ac \cdot \text{Cov}(\mathbf{X}, \mathbf{Y})$. (la derivación de las dos últimas expresiones se deja como ejercicio). (P-2)
(54)

Definición 21. Se denomina correlación entre dos variables aleatorias \mathbf{X} e \mathbf{Y} (no constantes) al coseno del ángulo formado por las proyecciones $(\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1}))$ e $(\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1}))$:

$$\text{Corr}(\mathbf{X}, \mathbf{Y}) = \frac{\langle (\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})) | (\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1})) \rangle}{\|(\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1}))\| \cdot \|(\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1}))\|} = \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\text{Dt}(\mathbf{X}) \cdot \text{Dt}(\mathbf{Y})}.$$

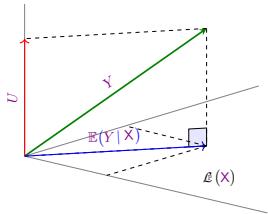
Fíjese que sumar una constante a cualquiera de las variables \mathbf{X} o \mathbf{Y} no cambia ni la covarianza y la desviación típica, por lo que tampoco cambia la correlación (algo natural, pues la correlación es el coseno del ángulo formado por las respectivas componentes variables, que no cambian al sumar variables constantes). Fíjese también que multiplicar cualquiera de las variables \mathbf{X} o \mathbf{Y} por un número no nulo, multiplica por dicho número tanto la covarianza como la desviación típica, por lo que tampoco cambia la correlación (pues tampoco dicha operación cambia el ángulo cuando el número es distinto de cero).

7.2 Momentos condicionados.

La esperanza de \mathbf{Y} condicionada al sistema de variables aleatorias \mathbf{X} , que denotamos con $\mathbb{E}(\mathbf{Y} | \mathbf{X})$, es la proyección ortogonal sobre $\mathcal{L}(\mathbf{X})$, que es el menor subespacio probabilístico que contiene a \mathbf{X} . Como todo subespacio probabilístico contiene la variable $\mathbf{1}$, la proyección $\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X})$ es perpendicular a $\mathbf{1}$; consecuentemente la diferencia $\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X})$ tiene esperanza nula pues: $\mathbb{E}(\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X})) = \mathbb{E}(\mathbf{1}(\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X}))) = \langle \mathbf{1} | \mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{X}) \rangle = 0$.

Este hecho da lugar al conocido Teorema de las Esperanzas Iteradas (véase la siguiente transparencia).

$$Y = \mathbb{E}(Y|\mathbf{X}) + U$$

**Teorema de las esperanzas iteradas**

Como $\mathbb{1} \perp (Y - \mathbb{E}(Y|\mathbf{X}))$, pues $\mathbb{1} \in L(\mathbf{X})$

$$\mathbb{E}(\mathbb{1} \cdot (Y - \mathbb{E}(Y|\mathbf{X}))) = 0$$

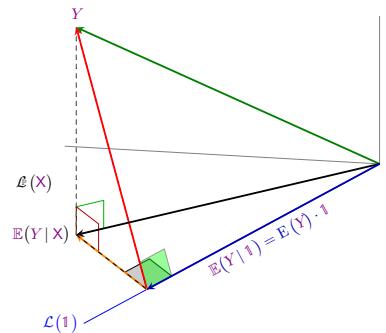
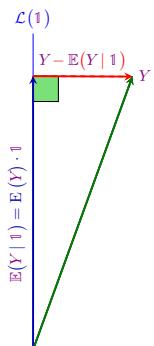
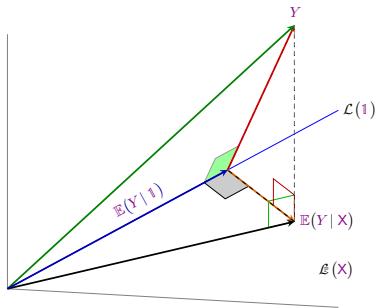
$$\mathbb{E}(Y - \mathbb{E}(Y|\mathbf{X})) = 0$$

$$\mathbb{E}(Y) - \mathbb{E}(\mathbb{E}(Y|\mathbf{X})) = 0$$

$$\Rightarrow \boxed{\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|\mathbf{X}))}$$

F46

Recuerde que en la Lección 2 también vimos que $\mu_{\mathbf{y}-\hat{\mathbf{y}}} = 0$ y que $\mu_{\mathbf{y}} = \mu_{\hat{\mathbf{y}}}$.



$$\mathbb{E}(\mathbb{E}(Y|\mathbf{X}) | \mathbb{1}) = \mathbb{E}(Y | \mathbb{1}) = \mathbb{E}(Y) \cdot \mathbb{1}$$

$$\text{Var}(Y) = \mathbb{E}((Y - \mathbb{E}(Y))^2)$$

$$\text{Var}(\mathbb{E}(Y|\mathbf{X})) = \mathbb{E}((\mathbb{E}(Y|\mathbf{X}) - \mathbb{E}(Y))^2)$$

$$\mathbb{E}(Y^2) = \text{Var}(Y) + \mathbb{E}(\mathbb{E}(Y)^2) = \text{Var}(Y) + \mathbb{E}(Y)^2$$

$$\mathbb{E}(\mathbb{E}(Y|\mathbf{X})^2) = \text{Var}(\mathbb{E}(Y|\mathbf{X})) + \mathbb{E}(Y)^2$$

F47

Recuerde que en la Lección 3, con el producto escalar de la estadística en \mathbb{R}^N , vimos que $\mu_{(\mathbf{y}^2)} = \sigma_y^2 + (\mu_y)^2$ y que $\mu_{(\hat{\mathbf{y}}^2)} = \sigma_{\hat{y}}^2 + (\mu_y)^2$; de donde obtuvimos las expresiones para la varianza de \mathbf{y} por una parte: $\sigma_y^2 = \mu_{(\mathbf{y}^2)} - (\mu_y)^2$; y de $\hat{\mathbf{y}}$ por otra: $\sigma_{\hat{y}}^2 = \mu_{(\hat{\mathbf{y}}^2)} - (\mu_y)^2$.

Algunas propiedades de la esperanza condicional (que no voy a demostrar).

- Si $\mathbf{Y} \in \mathcal{L}(\mathbf{X})$, entonces $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{Y}$.
- $\mathbb{E}(\mathbb{E}(\mathbf{Y}|\mathbf{X})) = \mathbb{E}(\mathbf{Y})$.
- Para $a, b \in \mathbb{R}$, tenemos que $\mathbb{E}(a\mathbf{X} + b\mathbf{Y}|\mathbf{Z}) = a\mathbb{E}(\mathbf{X}|\mathbf{Z}) + b\mathbb{E}(\mathbf{Y}|\mathbf{Z})$.
- Si $\mathcal{L}(\mathbf{X}) \subset \mathcal{L}(\mathbf{Z}) \subset \mathcal{E}$ (un Espacio Semi-Euclídeo Probabilístico), entonces $\mathbb{E}(\mathbb{E}(\mathbf{Y}|\mathbf{Z})|\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X})$.
- Si $f(\mathbf{X}) \in \mathcal{L}(\mathbf{Z})$ y el valor absoluto del producto de variables $\mathbf{Y} \cdot f(\mathbf{X})$ tienen definida la esperanza (si pertenece a $L_{\mathcal{E}}$) entonces $\mathbb{E}(f(\mathbf{X})\mathbf{Y}|\mathbf{Z}) = f(\mathbf{X}) \cdot \mathbb{E}(\mathbf{Y}|\mathbf{Z})$.
- Si $\mathcal{L}(\mathbf{X})$ y $\mathcal{L}(\mathbf{Y})$ son espacios probabilísticamente independientes, entonces $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbb{E}(\mathbf{Y})$.

(Lección 4)

T-6

Varianza condicional

Si $\mathbb{E}((\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}))^2) < \infty$, entonces:

$$\text{Var}(\mathbf{Y}|\mathbf{X}) = \mathbb{E}((\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}))^2 | \mathbf{X})$$

por tanto $\mathbb{E}(\text{Var}(\mathbf{Y}|\mathbf{X})) = \mathbb{E}((\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}))^2)$



$$\mathbb{E}(\mathbf{Y}^2) = \mathbb{E}(\mathbb{E}(\mathbf{Y}|\mathbf{X})^2) + \text{Var}(\mathbf{Y}|\mathbf{X})$$

Ley de la varianza total

$$\text{Var}(\mathbf{Y}) = \text{Var}(\mathbb{E}(\mathbf{Y}|\mathbf{X})) + \mathbb{E}(\text{Var}(\mathbf{Y}|\mathbf{X}))$$

F48

Recuerde que en la Lección 3 (aplicando el Teorema de Pitágoras en la descomposición ortogonal de \mathbf{y}) dedujimos que $\mu_{(\mathbf{y}^2)} = \mu_{(\hat{\mathbf{y}}^2)} + \sigma_e^2$, de donde obtuvimos que $\sigma_e^2 = \mu_{(\mathbf{y}^2)} - \mu_{(\hat{\mathbf{y}}^2)}$ (que corresponde a la primera de las dos últimas ecuaciones de la transparencia anterior); y por otra obtuvimos la expresión de la descomposición de la varianza $\sigma_y^2 = \sigma_{\hat{\mathbf{y}}}^2 + \sigma_e^2$ (que se corresponde con la Ley de la varianza total).

Hay una segunda manera de expresar la varianza condicional

$$\text{Var}(\mathbf{Y}|\mathbf{X}) = \mathbb{E}(\mathbf{Y}^2|\mathbf{X}) - (\mathbb{E}(\mathbf{Y}|\mathbf{X}))^2, \quad (P-3)$$

y por tanto $\mathbb{E}(\text{Var}(\mathbf{Y}|\mathbf{X})) = \mathbb{E}(\mathbb{E}(\mathbf{Y}^2|\mathbf{X})) - \mathbb{E}((\mathbb{E}(\mathbf{Y}|\mathbf{X}))^2)$; que por el Teorema de las Esperanzas Iteradas se reduce a $\mathbb{E}(\text{Var}(\mathbf{Y}|\mathbf{X})) = \mathbb{E}(\mathbf{Y}^2) - \mathbb{E}((\mathbb{E}(\mathbf{Y}|\mathbf{X}))^2)$.

Recuerde que en la Lección 3 también dedujimos que $\sigma_e^2 = \mu_{(\mathbf{y}^2)} - \mu_{(\hat{\mathbf{y}}^2)}$.

¡Nótese que todo proviene del álgebra lineal. Es una repetición de los pasos dados en el caso de los vectores de \mathbb{R}^N ; y ni una sola vez hemos necesitado emplear las funciones de densidad de las variables! (podemos necesitar las funciones de densidad y las integrales para calcular (el cálculo) de algunas esperanzas, pero no para derivar sus propiedades).

\mathcal{E}	\mathbb{R}^N
$\textcolor{violet}{Y}$	\mathbf{y}
$E(\textcolor{violet}{Y})$	$\langle \vec{y} \vec{1} \rangle_{\eta}$
$E(Y \textcolor{violet}{X})$	$\text{Prj}_{\mathcal{L}(\vec{1})}(\vec{y})$
$\text{Var}(Y)$	$\left\ \text{Prj}_{\mathcal{L}(\vec{1})^\perp}(\vec{y}) \right\ _{\eta}^2$
$E(Y X)$	$\text{Prj}_{\mathcal{L}(\vec{x}_1 \dots \vec{x}_k)}(\vec{y})$
$\text{Var}(E(Y X))$	$\left\ \text{Prj}_{\mathcal{L}(\vec{1})^\perp} \left(\text{Prj}_{\mathcal{L}(\vec{x}_1 \dots \vec{x}_k)}(\vec{y}) \right) \right\ _{\eta}^2$
$E(\text{Var}(Y X))$	$\left\ \left(\text{Prj}_{\mathcal{L}(\vec{x}_1 \dots \vec{x}_k)}(\vec{y}) \right)^2 \right\ _{\eta}$

Table 5: Dado que STC, SEC y SRC son nombres que corresponden a distancias medidas con la norma usual en \mathbb{R}^N , creo que sería más claro y sencillo usar exclusivamente σ_y^2 , $\sigma_{\hat{y}}^2$ y σ_e^2 . Lamentablemente los primeros términos son de uso generalizado y aparecen en la mayoría de manuales y programas informáticos estadísticos. Nótese también que aunque $\mathbf{y}^2 \in \mathbb{R}^N$, en general \vec{y}^2 no pertenece al espacio semi-euclídeo de probabilidad (tan solo pertenece al dominio de la función esperanza), por ello la última linea de la tabla no tiene una interpretación geométrica como el resto, pues la varianza condicional no es una proyección ortogonal en general (aunque sí lo sea en \mathbb{R}^N).

Nota técnica: la esperanza condicional $E(Y | X)$ pertenece a \mathcal{E} .⁵⁹ Por tanto su cuadrado pertenece a $L_{\mathcal{E}}$. Más arriba definimos la esperanza condicional como una proyección ortogonal; pero las proyecciones ortogonales están definidas en \mathcal{E} , no en $L_{\mathcal{E}}$. No obstante podemos definir la varianza condicional porque es posible extender la esperanza condicional fuera de \mathcal{E} . Con dicha extensión (cuya formalización queda fuera del curso) podemos definir $\text{Var}(Y | X)$; pero la extensión de la esperanza condicional fuera de \mathcal{E} pierde la interpretación geométrica como una proyección ortogonal.

7.3 Relación con las lecciones anteriores (Parte II).

Sin duda ya habrá advertido que los resultados expuestos en esta lección se parecen mucho a los de las lecciones 2 y 3. ¿Cuál es la razón? ¿Por qué el ajuste MCO de vectores en \mathbb{R}^N se asemeja a la esperanza condicional de una variable aleatoria? El motivo es que lo visto en las tres primeras lecciones es un caso particular de lo expuesto en esta: \mathbb{R}^N verifica⁶⁰ trivialmente los axiomas A1, A2 y A4 de la Definición 6 en la página 40 y $\mathbf{1} \in \mathbb{R}^N$; además vimos en la Sección 2.3 (Lección 2) que el producto escalar de la estadística $\langle \cdot | \cdot \rangle_s$ es tal que $\|\mathbf{1}\|_s = \langle \mathbf{1} | \mathbf{1} \rangle_s = 1$ (por tanto verifica A5) y además verifica A3, pues $\mathbf{x}_1 \odot \mathbf{y}_1 = \mathbf{x}_2 \odot \mathbf{y}_2 \implies \langle \mathbf{x}_1 | \mathbf{y}_1 \rangle_s = \langle \mathbf{x}_2 | \mathbf{y}_2 \rangle_s$, es decir, que el conjunto de pares $\{(x \odot y, \langle x | y \rangle_s) \mid x, y \in \mathbb{R}^N\}$ es una función, que en ese momento denominamos *media aritmética del producto*, pero que en este contexto más general es la *esperanza del producto*. Por tanto...

7.3.1 (\mathbb{R}^N, s) es un Espacio Euclídeo Probabilístico

Si considera \mathbb{R}^N junto al producto escalar habitual de la estadística $\langle \cdot | \cdot \rangle_s$; es decir, si Ω es el conjunto de índices $\{1, \dots, N\}$ de manera que \mathcal{E} es \mathbb{R}^N y si $\langle \cdot | \cdot \rangle_{\eta} = \langle \cdot | \cdot \rangle_s$; entonces se verifican los cinco axiomas de la Definición 6 y, por tanto, (\mathbb{R}^N, s) es un *Espacio Euclídeo de Probabilidad*: sus vectores son *variables aleatorias* y la *media aritmética* de un vector es su *esperanza matemática* en \mathbb{R}^N .

Dicho de otro modo, ¡al sustituir el producto escalar usual en \mathbb{R}^N por otro que garantice que la norma del vector constante $\mathbf{1}$ es 1 y que verifique A3 “probabilizamos” el Espacio Euclídeo \mathbb{R}^N ; así pues, la Figura 2 en la página 16 es un caso particular de la Figura 5 donde las variables aleatorias son los vectores de \mathbb{R}^N y la esperanza es la media (Tabla 5).

7.3.2 La regresión o ajuste MCO es un caso particular de esperanza condicional

En las lecciones anteriores exigimos que el vector constante $\mathbf{1}$ perteneciera al subespacio generado por los regresores, es decir, que $\mathbf{1} \in \mathcal{C}(\mathbf{X})$. Esta condición viene impuesta por el axioma A5 de la Definición 6. El axioma A5 impone que la función (o vector) constante *uno* debe pertenecer a cualquier subespacio de probabilidad. Consecuentemente, $(\mathcal{C}(\mathbf{X}), s|_{\mathcal{C}(\mathbf{X}) \times \mathcal{C}(\mathbf{X})})$ solo puede ser un subespacio probabilístico de \mathbb{R}^N si $\mathbf{1} \in \mathcal{C}(\mathbf{X})$. Y solo en ese caso la proyección de \mathbf{y} sobre $\mathcal{C}(\mathbf{X})$ es una “esperanza condicional”.

⁵⁹Mejor sería decir que es una clase de equivalencia contenida en \mathcal{E} .

⁶⁰En este contexto se entiende que si las componentes de \mathbf{x} son positivas, $\sqrt{\mathbf{x}}$ es aquel vector de \mathbb{R}^N cuyas componentes son la raíz cuadrada de las componentes de \mathbf{x} .

Es decir, sólo cuando el ajuste MCO (cuando la proyección) se realiza empleando un conjunto de regresores tales que $\mathbf{1} \in \mathcal{C}(\mathbf{X})$ (donde las columnas de \mathbf{X} son dichos regresores) podemos hablar propiamente de *regresión*. En caso contrario estaremos realizando una proyección ortogonal, pero no deberemos hablar de regresión. Normalmente esta condición se asegura haciendo que la primera columna de \mathbf{X} sea el vector $\mathbf{1}$, pero esto no es estrictamente necesario (lo veremos al estudiar algunos modelos con variables ficticias).

7.3.3 Y si \mathbb{R}^N es un subespacio de probabilidad ¿dónde está el azar?

Cuando nos referimos a juegos de *azar* (lanzamiento de monedas, dados, o reparto de naipes de una baraja) nos referimos a juegos en los que no tenemos información suficiente para conocer el resultado antes de cada partida. Esta situación de ignorancia (o falta de información) habitualmente se modeliza con variables aleatorias y modelos estocásticos. Por ello los términos *aleatorio* o *estocástico* han cobrado un aire un tanto “esotérico”⁶¹ pese a que la realidad es que tan solo son adjetivos para modelos que tratan con algo tan cotidiano como la ignorancia. Piense que cuando nos reparten cartas de una baraja, el resultado es incierto solo porque no sabemos cuál es la disposición de los naipes. Lo que convierte el resultado en “aleatorio” es precisamente eso, que no sabemos con antelación cómo están dispuestas las cartas del mazo.



Figure 6: El lenguaje estadístico a menudo resulta un tanto esotérico.

En el espacio euclídeo probabilístico \mathbb{R}^3 , el vector $\mathbf{x} = (0, 0, \sqrt{2},)$ es una variable aleatoria. No obstante, aquí todo es conocido: el conjunto de sucesos elementales es $\Omega = \{1, 2, 3\}$ y la *variable aleatoria* \mathbf{x} asigna al índice 1 el número 0, al índice 2 el número 0 y al índice 3 le asigna $\sqrt{2}$. Como todo es conocido nos resulta chocante denominar *variable aleatoria* al vector $(0, 0, \sqrt{2},)$. Pero la Definición 6 en la página 40 es clara: si consideramos el par (\mathbb{R}^3, s) , sus vectores son variables aleatorias. Así pues, el azar no está en las variables aleatorias, aunque las empleemos para modelizarlo. El uso de variables aleatorias para modelizar el azar consiste en asociar cada valor observado en la naturaleza con el valor tomado por *una variable aleatoria*, pero desconociendo en qué elemento $\omega \in \Omega$ ha sido evaluada la función.

Para entenderlo, imaginemos que solicito a una tercera persona que se fije en una de las tres componentes del vector $\mathbf{x} = (0, 0, \sqrt{2},)$ y que la diga en voz alta (en la jerga estadística, llamaremos al número escuchado una *realización* de la variable aleatoria). Aunque yo conozca perfectamente la función \mathbf{x} , desconozco qué numero va a decir el sujeto. Pero incluso tras escuchar el número quizás no lo sepa todo sobre el experimento: si me dice “ $\sqrt{2}$ ” sabré que se fijó en la última componente... pero si dice “0” no sabré si se fijó en la primera o en la segunda.

Ahora imagine que el juego se repite con el vector de \mathbb{R}^{1555} cuyas componentes son los primeros 1555 decimales de π . Aunque la variable aleatoria es conocida (véase más abajo los 1555 decimales), si usted escucha que el sujeto dice “1” no sabrá cuál de los 1555 decimales escogió (pese a que pueda descartar todos los que son distintos de 1).

$\pi \approx 3.1415926535897932384626433832795028841971693993751058209749445923078164062862089986280348253421170679821480865132823$
 $066470938446095505822317253594081284811174502841027019385211055596446229489549303819644288109756659334461284756482337867$
 $83165271201909145648566923460348610454326648213393607260249141273724587006606315588174881520920962892540917153643678925$
 $903600113305305488204665213841469519415116094330572703657595919530921861173819326117931051185480744623799627495673518857$
 $52724891227938183011949129833673362440656643086021394946395224737190721798609437027705392171762931767523846748184676694$
 $051320005681271452635608277857713427577896091736371787214684409012249534301465495853710507922796892589235420199561121290$
 $219608640344181598136297747713099605187072113499999983729780499510597317328160963185950244594553469083026425223082533446$
 $850352619311881710100031378387528865875332083814206171776691473035982534904287554687311595628638823537875937519577818577$
 $805321712268066130019278766111959092164201989380952572010654858632788659361533818279682303019520353018529689957736225994$
 $138912497217752834791315155748572424541506959508295331168617278558890750983817546374649393192550604009277016711390098488$
 $24012853616035637076601047101819429555961989467678374494482553797747268471040475346462080466842590694912933136770289891$
 $521047521620569660240580381501935112533824300355876402474964732639141992726042699227967823547816360093417216412199245863$
 $150302861829745557067498385054945885869269956909272107975093029553211653449872027559602364806654991198818347977535663698$

⁶¹Y el modo habitual de exponer la probabilidad creo que contribuye a ello.

La situación en la práctica es aún más incierta. Normalmente ni siquiera Ω queda explícitamente especificado. Al modelizar una magnitud económica con una variable aleatoria, (por ejemplo *el volumen de ventas de El Corte Inglés*) nunca se describe el contenido del conjunto Ω , así que tampoco podemos saber qué valor toma la variable aleatoria en cada uno de los elementos de su dominio (sucesos elementales). Pero el azar tampoco proviene de esa habitual ausencia de descripción explícita del conjunto Ω (en los ejemplos en \mathbb{R}^3 y \mathbb{R}^{1555} de más arriba, Ω está perfectamente descrito).

El azar no queda recogido en la definición de las variables aleatorias. El azar queda recogido en el modo en que se emplean las variables aleatorias: interpretamos que los datos que observamos son valores tomados por una función (variable aleatoria) en un valor desconocido de su dominio (en un suceso elemental desconocido).

7.4 Diferencias con las lecciones previas.

En la sección anterior hemos visto que lo expuesto en las tres primeras lecciones es un caso particular de lo expuesto en esta, es decir, que si consideramos \mathbb{R}^N junto al producto escalar de la estadística $\langle \cdot | \cdot \rangle_s$, los vectores de \mathbb{R}^N son variables aleatorias. Pero los espacios semi-euclídeos de probabilidad son más generales que el caso particular (\mathbb{R}^N, s) estudiado en las tres primeras lecciones.

Una primera diferencia radica en que $\langle \cdot | \cdot \rangle_s$ es un producto escalar, y por tanto

$$\|\mathbf{x}\|_s = \langle \vec{x} | \vec{x} \rangle_s = 0 \Leftrightarrow \mathbf{x} = 0.$$

Esto no es cierto con los semi-productos escalares. En un espacio semi-euclídeo de probabilidad (\mathcal{E}, η) puede ocurrir que

$$\|\mathbf{X}\|_\eta = \langle \mathbf{X} | \mathbf{X} \rangle_\eta = 0 \quad \text{con} \quad \mathbf{X} \neq \mathbf{0};$$

donde $\mathbf{0}$ es la función constante cero, i.e., $\mathbf{0} = 0 \cdot \mathbf{1}$.

Una consecuencia que se deriva de esto es que dos variables aleatorias distintas pueden estar a distancia cero entre ellas, y consecuentemente ambas estarán a la misma distancia de una tercera. Por dicho motivo, el conjunto de variables aleatorias que están a distancia mínima de una variable aleatoria dada puede contener más de un elemento. Así que decir que la esperanza condicional $\mathbb{E}(\mathbf{Y} | \mathbf{X})$ es “una” variable aleatoria de $\mathcal{L}(\mathbf{X})$ es incorrecto en general. En realidad es el *conjunto* de variables aleatorias de $\mathcal{L}(\mathbf{X})$ que están a mínima distancia de \mathbf{Y} . Dicho conjunto constituye una clase de equivalencia de *variables aleatorias que son iguales salvo en un conjunto de probabilidad (o de medida) nula*. En el apéndice se dedicara algo de atención a esto, aunque sin profundizar demasiado.

En cualquier caso, para el curso no tiene mayor relevancia práctica considerar si la esperanza condicional es una variable aleatoria o si es una clase de equivalencia.

Sin embargo, la segunda diferencia si tiene importancia, pues condiciona fuertemente los supuestos necesarios para poder estimar la esperanza condicional por MCO. Veamos el motivo.

Si tenemos una matriz de regresores tal que $\mathbf{1}$ es combinación lineal de sus columnas, el subespacio euclídeo probabilístico de (\mathbb{R}^N, s) generado por las k columnas (los regresores) de una matriz \mathbf{X} es el espacio columna de la matriz, es decir, el subespacio vectorial formado por el conjunto de combinaciones lineales de las columnas: $\mathcal{L}(\mathbf{X}) = \mathcal{L}(\mathbf{X})$. Consecuentemente $\mathcal{L}(\mathbf{X})$ tiene dimensión k . Esto es así porque $\mathcal{L}(\mathbf{X})$ cumple de manera trivial los supuestos A2, A4 y A5 de la Definición 6 en la página 40.

En la Lección 1 ya hemos visto que, si \mathbf{X} tiene columnas linealmente independientes, el vector de $\mathcal{L}(\mathbf{X})$ más próximo a \mathbf{y} es

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}, \quad \text{donde} \quad (\mathbf{X}^\top \mathbf{X})\hat{\beta} = \mathbf{X}^\top \mathbf{y};$$

es decir, la *esperanza condicional* $\hat{\mathbf{y}} \equiv \mathbb{E}(\mathbf{y} | \mathbf{X})$ es un vector de $\mathcal{L}(\mathbf{X})$ (i.e., una combinación lineal de los regresores).

Pero si disponemos de un sistema de k variables aleatorias \mathbf{X} , para cumplir los supuestos A2, A4 y A5 es necesario incorporar a $\mathcal{L}(\mathbf{X})$ tantos vectores (tantas variables aleatorias) que la dimensión del subespacio vectorial resultante es, en general, infinita; es decir, generalmente el conjunto de vectores $\mathcal{L}(\mathbf{X})$ es muchísimo más grande que $\mathcal{L}(\mathbf{X})$:

$$\mathcal{L}(\mathbf{X}) \subset \mathcal{L}(\mathbf{X}).$$

Por este motivo, dados $\mathbf{Y} \in \mathcal{E}$ y $\mathbf{X} \subset \mathcal{E}$, generalmente el vector de $\mathcal{L}(\mathbf{X})$ más próximo a \mathbf{Y} no es el vector de $\mathcal{L}(\mathbf{X})$ más próximo a \mathbf{Y} , ya que al añadir a $\mathcal{L}(\mathbf{X})$ las variables aleatorias necesarias para satisfacer los 5 axiomas y así obtener $\mathcal{L}(\mathbf{X})$, algunas de las variables aleatorias añadidas pudieran estar más próximas a \mathbf{Y} que cualquiera de las que ya había

en $\mathcal{L}(\mathbf{X})$. Por tanto, generalmente la esperanza condicional no es una combinación lineal de los regresores; es decir,

$$\text{aunque } \mathbb{E}(Y|\mathbf{X}) \subset \mathcal{L}(\mathbf{X}), \quad \text{generalmente } \mathbb{E}(Y|\mathbf{X}) \notin \mathcal{L}(\mathbf{X}).$$

A fin de lograr modelos en los que exista una relación simple entre los regresores \mathbf{X} y la esperanza condicional, en la próxima lección se impondrán condiciones suficientes como para lograr que la esperanza condicional sí sea combinación lineal de los regresores \mathbf{X} . Dicho modelo simplificado se denomina *modelo clásico de regresión lineal*.

Una tercera e importante diferencia es que \mathbb{R}^N es cerrado para el producto punto a punto (o producto Hadamard), es decir, para todo \mathbf{x} y \mathbf{y} de \mathbb{R}^N , su producto $\mathbf{x} \odot \mathbf{y}$ también es un vector de \mathbb{R}^N . Pero esto es generalmente falso en los espacios semi-euclídeos de probabilidad, pues si \mathbf{X} e \mathbf{Y} son variables aleatorias de un espacio semi-euclídeo de probabilidad \mathcal{E} (es decir, variables con varianza definida), en general el producto $\mathbf{XY} \notin \mathcal{E}$; dicho producto pertenece al dominio de la función esperanza matemática $L_{\mathcal{E}}$ (es decir, es una variable cuya esperanza del valor absoluto está definida, pero no su varianza). Además, el espacio $L_{\mathcal{E}}$ (que contiene a \mathcal{E}) tampoco es cerrado para el producto punto pues el producto de dos variables aleatorias de $L_{\mathcal{E}}$ no pertenece a $L_{\mathcal{E}}$ en general (por tanto, ni siquiera la esperanza está definida para dicho producto). Es decir, la estructura en \mathbb{R}^N es muchísimo más simple que la de un espacio semi-euclídeo de probabilidad genérico \mathcal{E} .

7.5 La regresión como descomposición ortogonal

Si tanto \mathbf{Y} como las componentes de $\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k]$ variables aleatorias pertenecen a un mismo Espacio Semi-Euclídeo de Probabilidad \mathcal{E} , podemos descomponer \mathbf{Y} en dos componentes ortogonales: la esperanza condicional

$$\mathbb{E}(Y|\mathbf{X})$$

que es la proyección ortogonal de \mathbf{Y} sobre el subespacio probabilístico $\mathcal{L}(\mathbf{X})$ generado por \mathbf{X} ; más otro componente,

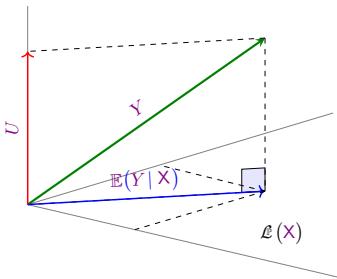
$$\mathbf{U} = \mathbf{Y} - \mathbb{E}(Y|\mathbf{X})$$

que es ortogonal al anterior. Esta descomposición se conoce como *regresión*. Lo destacable de la interpretación geométrica de la regresión que acabamos de ver es que es análoga al ajuste MCO en \mathbb{R}^N estudiado en las lecciones anteriores (en realidad lo visto en las lecciones anteriores es un caso particular).

(Lección 4)

T-7 La regresión es una descomposición ortogonal (que no implica causalidad)

$$\mathbf{Y} = \mathbb{E}(Y|\mathbf{X}) + \mathbf{U}$$



donde $\mathbb{E}(Y|\mathbf{X})$ es la proyección ortogonal sobre $\mathcal{L}(\mathbf{X})$.

- **como relación estadística:** siempre es cierta. No implica causalidad ni conclusiones teóricas
- **como lectura teórica:** su interpretación puede carecer de sentido (regresiones espurias)

F49

La regresión como modelo probabilístico. Únicamente dice que si disponemos un sistema de variables aleatorias \mathbf{X} , podemos descomponer la variable aleatoria \mathbf{Y} en dos componentes ortogonales entre si: $\mathbf{Y} = \mathbb{E}(Y|\mathbf{X}) + \mathbf{U}$. Es una descomposición algebraica sin ninguna *teoría económica* (o modelo explicativo) detrás. La descomposición ortogonal

$$\mathbf{Y} = \mathbb{E}(Y|\mathbf{X}) + \mathbf{U}$$

solo dice que “puedo descomponer \mathbf{Y} en las dos partes de la derecha de la ecuación”, y de esta mera descomposición algebraica no se puede (ni se debe) extrapolar una *teoría explicativa* de nada, ni tampoco relaciones de causalidad.

Un anhelo o pretensión como analistas: la regresión como modelo explicativo Al tratar de interpretar la regresión (la descomposición ortogonal de más arriba) como modelización de algún aspecto del mundo real, el analista podría pretender contemplar distintos e hipotéticos (aunque probablemente falsos) escenarios. Entre otros:

1. las variables \mathbf{X} generan (o causan) parcialmente a \mathbf{Y} . Una aplicación de dicha relación de causalidad, es que si disponemos de una muestra \mathbf{X} de \mathbf{X} quizá podemos tratar de prever parcialmente \mathbf{Y} (pues su componente $\mathbb{E}(\mathbf{Y} | \mathbf{X})$ es función de \mathbf{X}).
2. \mathbf{Y} causa (o genera) parcialmente las variables \mathbf{X} ; por tanto, al observar \mathbf{X} podemos tratar de inferir qué ha ocurrido con la variable causante \mathbf{Y} ; como cuando vemos lluvia cayendo sobre el suelo y deducimos que hay nubes en el cielo (que son las causantes de la lluvia).
3. hay alguna otra causa común que genera conjuntamente tanto \mathbf{Y} como \mathbf{X} .
4. la descomposición correspondiente es completamente inútil, pues las variables involucradas no tienen nada que ver las unas con las otras (como por ejemplo el numero de coches rojos circulando por la M-30 durante el sorteo de Navidad con el número premiado en la Lotería).

Como economistas querríamos que la descomposición algebraica de la regresión fuera reflejo de relaciones teóricas entre \mathbf{X} e \mathbf{Y} (donde \mathbf{X} e \mathbf{Y} son modelos estadísticos para magnitudes del mundo real). En este sentido queremos leer que \mathbf{Y} (por ejemplo el consumo) está generado por una función de las variables \mathbf{X} (por ejemplo la renta disponible) además de otras causas \mathbf{U} (“ortogonales” a la renta). Esta interpretación sugiere algunos de los nombres dados en estadística a \mathbf{X} e \mathbf{Y} .

Nombres dados en estadística a las variables de un modelo de regresión. Suponiendo que \mathbf{Y} es función del sistema de variables aleatorias \mathbf{X} y de \mathbf{U} :

- Llamamos a \mathbf{Y} *variable endógena* (cuando consideramos que se determina su valor o características a través del modelo), *variable objetivo* (cuando es una magnitud que deseamos controlar, por ejemplo la inflación si somos la autoridad monetaria), o *variable explicada*. Un nombre menos pretencioso es *regresando*.
- Las k variables \mathbf{X}_j se denominan *variables exógenas*, (cuando consideramos que vienen dadas de manera externa al modelo), o *vbles. de control* (cuando tenemos capacidad de alterar su valor para, a través del modelo, controlar \mathbf{Y} ; por ejemplo para controlar la inflación fijando la oferta monetaria o los tipos de interés si somos la autoridad monetaria), o *variables explicativas*. Un nombre menos pretencioso es el de *regresores*.
- \mathbf{U} es el efecto conjunto de otras variables o circunstancias que influyen sobre \mathbf{Y} , y que por alguna razón no contemplamos de manera explícita en el modelo (... dificultad o imposibilidad de observarlas, porque las desconocemos, etc.).⁶² La variable \mathbf{U} se suele denominar *perturbación* o *error* (en el contexto de la Econometría, a mi me gusta llamarla “*otras cosas*”).

(Lección 4)

T-8 Modelo de regresión: Nombres de las variables

En la expresión

$$\mathbf{Y} = \mathbb{E}(\mathbf{Y} | \mathbf{X}) + \mathbf{U}$$

usamos los siguientes nombres:

- \mathbf{Y} : vble. endógena, objetivo, explicada (o *regresando*)
- $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \dots \mathbf{X}_k]$: vbles. exógenas, de control, explicativas (o *regresores*)
- \mathbf{U} : factor desconocido o *perturbación* (a mi me gusta llamarlo “*otras cosas*”)

F50

Pero recuerde que a pesar de lo que algunos de estos “*sugerentes*” nombres puedan inducir a pensar, la descomposición algebraica en componentes ortogonales es una relación que siempre podemos encontrar entre cualesquiera \mathbf{Y} y \mathbf{X} ⁶³; y que por sí misma no permite extraer ni relaciones de causalidad, ni conclusiones teóricas.

No obstante, imagine un contexto en el que dispongamos *previamente* de un modelo teórico bien fundamentado y que nos indica claramente en el modelo estadístico qué variables \mathbf{X}_j son causantes (y por ello las situamos a derecha

⁶²En ciencias experimentales suele considerarse que \mathbf{U} es el *error* cometido al medir la variable \mathbf{Y} .

⁶³Siempre y cuando \mathbf{Y} y \mathbf{X}_j pertenezcan al mismo Espacio Euclídeo de Probabilidad \mathcal{E} .

como regresores) y qué variable Y es causada (y que consecuentemente situamos a la izquierda de la ecuación como regresando); pues bien, en tal contexto es donde *quizá* podamos sacar conclusiones *teóricas* de la regresión⁶⁴ (al menos para corroborar, o para descartar, el modelo teórico previo).

Empleo la palabra *quizá* porque con frecuencia los datos disponibles no miden aquellos conceptos empleados en los modelos teóricos (consumo permanente, preferencias, nivel de precios, utilidades, aversión al riesgo, etc.) y porque en ocasiones el modelo teórico empleado tampoco representa adecuadamente los hechos.

(Lección 4)

T-9

Modelo Clásico de Regresión Lineal

Modelo especial en el que la descomposición ortogonal

$$Y = \mathbb{E}(Y | X) + U$$

es tal que

- $\mathbb{E}(Y | X) = X\beta \in \mathcal{L}(X)$ (función lineal)
- $\text{Var}(Y | X)$ está definida y es cte.

¿QUÉ HACE FALTA PARA QUE ESTO SE CUMPLA?

¿En qué condiciones es la recta de regresión una estimación insesgada de la esperanza condicional $\mathbb{E}(Y | X)$?

F51

Problemas de la Lección 4

Momentos teóricos de una variable aleatoria

(L-4) PROBLEMA 1. Demuestre que $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

(L-4) PROBLEMA 2. Demuestre que $\text{Cov}(aX + b\mathbb{I}, cY + d\mathbb{I}) = ac \cdot \text{Cov}(X, Y)$.

(L-4) PROBLEMA 3. Demuestre que $\text{Var}(Y | X) = \mathbb{E}(Y^2 | X) - (\mathbb{E}(Y | X))^2$.

Fin de los Problemas de la Lección 4

8 Apéndice: interpretación geométrica de la Probabilidad

Este apéndice es un intento apresurado de esbozar algunas propiedades geométricas que hay detrás de la Probabilidad y que generalmente son desconocidas (pese a que son de gran ayuda a la hora de pensar). No es parte del temario, pero es interesante conocer este enfoque geométrico, que es el que está detrás de lo que aparece en las transparencias de clase, y que sí es parte del temario.

En Econometría estamos interesados en usar un conjunto reducido de variables aleatorias: específicamente nos interesan aquellas que tienen varianza (y eventualmente también algunas que, aún sin varianza, al menos tienen definida la esperanza).⁶⁵ Además, si desde el principio nos pudiéramos centrar en aquellas variables aleatorias con varianza, estaríamos trabajando desde el principio en un marco análogo al de las lecciones anteriores (pues dispondríamos de las proyecciones ortogonales al ceñirnos a un subespacio que además es un espacio semi-euclídeo).

Afortunadamente es posible arrancar de este modo. Los detalles técnicos están fuera del ámbito de este curso pero, a modo de ilustración, sepa que todo se puede desarrollar formalmente a partir de la Definición 6 de *Espacio Euclídeo*

⁶⁴ Nótese que solo en este contexto algunos de los nombres dados a X e Y podrían estar justificados.

⁶⁵ Para aclarar... todas las variables aleatorias que tienen varianza también tienen esperanza, pero no al revés. Es decir, hay variables aleatorias que tienen esperanza pero no tienen varianza. Si recapitula sobre lo que le han contado en los cursos de estadística y probabilidad, se dará cuenta de que, aunque quizás le hayan escrito los espacios probabilísticos, luego el curso se habrá limitado a tratar exclusivamente con variables aleatorias que tenían tanto esperanza como varianza (es probable que algunos estudiantes ni sean conscientes de la existencia de variables aleatorias sin esperanza). Así pues, aunque los cursos de estadística y probabilidad plantean un marco teórico muy general, luego se limitan a tratar con un grupo muy reducido de variables aleatorias: aquellas con esperanza y varianza. (Véase el video *The Two Envelope Problem - a Mystifying Probability Paradox*).

Probabilístico en la página 40. Partiendo de los *espacios semi-euclídeos de probabilidad* se pueden definir los conceptos de: suceso, probabilidad, independencia, etc. Son los mismos conceptos que usted ya ha visto en las asignaturas de estadística y probabilidad, pero definidos de manera distinta. Por ello es conveniente indicar algunas de sus propiedades en el contexto de los *espacios semi-euclídeos probabilísticos*. A continuación, voy a apuntar algunas ideas importantes; pero tenga en cuenta que la demostración y justificación de lo que indicaré a continuación requeriría de un curso completo de Probabilidad visto desde los *espacios semi-euclídeos probabilísticos*, es decir, como una extensión a otro curso de Álgebra Lineal (desgraciadamente no hay tiempo para ver todo eso dentro de uno de Econometría).

Espacios Euclídeos Probabilísticos

- Un *Espacio Semi-Euclídeo Probabilístico*⁶⁶ (\mathcal{E}, η) es un espacio vectorial \mathcal{E} con un semi producto escalar η (véase la Definición 6 en la página 40). Los vectores de \mathcal{E} son funciones que van de un conjunto Ω a \mathbb{R} (a dichos vectores los llamamos *variables aleatorias*). Al semi producto escalar lo llamamos *esperanza del producto (de variables aleatorias)*. Todo *Espacio Euclídeo Probabilístico* cumple los cinco axiomas de la Definición 6 en la página 40.

Dado que el semi producto escalar de dos variables aleatorias \mathbf{X} e \mathbf{Y} es $\langle \mathbf{X} | \mathbf{Y} \rangle_{\eta} = E(\mathbf{XY})$, tenemos que

- como dos vectores son ortogonales cuando su producto escalar es nulo: $\mathbf{X} \perp \mathbf{Y} \Leftrightarrow E(\mathbf{XY}) = 0$.
- como el cuadrado de la norma es el semi producto escalar de un vector por si mismo, el cuadrado de la norma de una variable aleatoria es $\|\mathbf{X}\|_{\eta}^2 = E(\mathbf{X}^2)$, o si se prefiere: su longitud o *norma* es $\|\mathbf{X}\|_{\eta} = \sqrt{E(\mathbf{X}^2)}$.

En la literatura el *Espacio Euclídeo Probabilístico* $(\mathcal{E}, \eta, \Omega)$ se conoce como $L_2(\Omega, \mathcal{F}, P)$, es decir, el conjunto de variables aleatorias definidas sobre Ω con *segundos momentos finitos*, ($E(\mathbf{X}^2) < \infty$); o dicho de otra forma: *con varianza definida*.

Variables aleatorias constantes Una *variable aleatoria constante* es una función constante, es decir, asigna un mismo valor c a todo *suceso elemental* ω de Ω . Dado que $\mathbf{1}(\omega) = 1$ para todo $\omega \in \Omega$, tenemos que toda variable aleatoria constante es un múltiplo de la función constante uno $\mathbf{1}$; es decir, las variables aleatorias constantes son los vectores de la recta $\mathcal{L}(\mathbf{1})$:

$$\{\text{variables aleatorias constantes}\} = \mathcal{L}(\mathbf{1}) = \{\mathbf{X} \mid \text{existe un } c \in \mathbb{R} \text{ tal que } \mathbf{X} = c\mathbf{1}\}.$$

Variables aleatorias ortogonales a las constantes Dado que los subespacios probabilísticos de \mathcal{E} contienen el vector constante $\mathbf{1}$, todos ellos comparten la recta común $\mathcal{L}(\mathbf{1})$. Así, dentro de cualquier subespacio probabilístico \mathcal{S} podemos considerar el conjunto de variables aleatorias (de vectores) que son perpendiculares a la recta $\mathcal{L}(\mathbf{1})$. Si denotamos con $\mathcal{L}(\mathbf{1})^\perp$ al conjunto de variables aleatorias de \mathcal{E} que son perpendiculares a $\mathbf{1}$ y si \mathcal{S} es un subespacio probabilístico de \mathcal{E} , entonces $\mathcal{S} \cap \mathcal{L}(\mathbf{1})^\perp$ es el subconjunto de variables aleatorias de \mathcal{S} que son perpendiculares a $\mathbf{1}$.

Los subespacios probabilísticos $\mathcal{L}(\mathbf{1})$ y $\mathcal{L}(\mathbf{1})^\perp$ son complementos ortogonales, es decir, que cualquier *espacio euclídeo probabilístico* \mathcal{E} se puede expresar mediante la siguiente suma directa: $\mathcal{E} = \mathcal{L}(\mathbf{1}) \oplus \mathcal{L}(\mathbf{1})^\perp$. Así que cualquier variable aleatoria \mathbf{X} de \mathcal{E} se puede descomponer en dos componentes ortogonales: una de ellas es la proyección de \mathbf{X} sobre $\mathcal{L}(\mathbf{1})$ y la otra la proyección de \mathbf{X} sobre $\mathcal{L}(\mathbf{1})^\perp$.

Sucesos y probabilidad

- Llamamos función indicatriz de un subconjunto A , que denotamos por $\mathbf{1}_A$, a la función que para todo ω que pertenece a A toma el valor 1 y para todo ω que NO pertenece a A toma el valor 0. Es decir

$$\begin{cases} \mathbf{1}_A(\omega) = 1 & \omega \in A \\ \mathbf{1}_A(\omega) = 0 & \omega \notin A \end{cases}.$$

Por tanto, la función indicatriz del conjunto vacío, $\mathbf{1}_{\emptyset}$, es la función constante *cero*, $\mathbf{0}$; y la función indicatriz $\mathbf{1}_{\Omega}$ es la función constante *uno*, $\mathbf{1}$.

Las funciones indicatrices son *idempotentes*: $(\mathbf{1}_A)^2 = \mathbf{1}_A \cdot \mathbf{1}_A = \mathbf{1}_A$ (ya que $0 \cdot 0 = 0$ y $1 \cdot 1 = 1$).

- Como ya hemos visto, todo *Espacio Euclídeo Probabilístico* \mathcal{E} contiene la función constante *uno*, $\mathbf{1} \in \mathcal{E}$; y además su norma es uno, $\|\mathbf{1}\|_{\eta} = 1$ (es el quinto axioma de la Definición 6 en la página 40).

⁶⁶Sería más correcto decir *Espacio Semi-Euclídeo Probabilístico*, ya que está dotado de un semi producto escalar.

- Se dice que un subconjunto A de Ω es un *suceso* si, y solo si, su función indicatriz es una variable aleatoria de \mathcal{E} , es decir, si y solo si $\mathbb{1}_A \in \mathcal{E}$.
- La *probabilidad* de un suceso A es el cuadrado de la norma de su función indicatriz: $\mathbb{P}(A) = \|\mathbb{1}_A\|_\eta^2 = \langle \mathbb{1}_A | \mathbb{1}_A \rangle_\eta = \mathbb{E}((\mathbb{1}_A)^2) = \mathbb{E}(\mathbb{1}_A)$; pues $\mathbb{1}_A \cdot \mathbb{1}_A = \mathbb{1}_A$. Pero entonces... recordando la interpretación geométrica de la esperanza también tenemos que la *probabilidad* de un suceso A es el valor por el que hay que multiplicar el vector constante $\mathbb{1}$ para obtener la proyección ortogonal de $\mathbb{1}_A$. Véase la Figura 7.

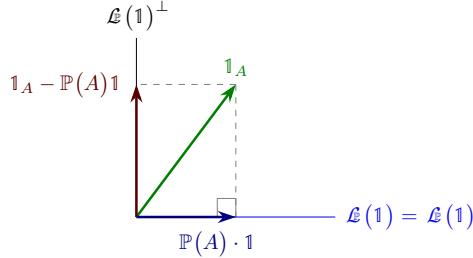


Figure 7: Descomposición ortogonal de la función indicatriz correspondiente al suceso A .

Sucesos de probabilidad nula y distancia entre vectores medida con la norma de \mathcal{E} . Consecuentemente, aquellas funciones indicatrices que pertenecen a $L(\mathbb{1})^\perp$, es decir, aquellas que son ortogonales a las variables aleatorias constantes, corresponden a sucesos de probabilidad nula, $\mathbb{1}_A \in L(\mathbb{1})^\perp \iff \mathbb{P}(A) = 0$.

Y como $\mathbb{P}(A) = \|\mathbb{1}_A\|_\eta^2$, se deduce que las funciones indicatrices de sucesos de probabilidad nula tienen longitud nula según la norma $\|\cdot\|_\eta$ (la “vara de medir”) de \mathcal{E} .

Por otra parte, en todo espacio euclídeo la distancia entre dos vectores es la raíz cuadrada del semi producto escalar:

$$\text{distancia}(\mathbf{X}, \mathbf{Y})_\varphi = \|\mathbf{Y} - \mathbf{X}\|_\eta = \sqrt{\langle (\mathbf{Y} - \mathbf{X}) | (\mathbf{Y} - \mathbf{X}) \rangle_\eta} = \sqrt{\mathbb{E}((\mathbf{Y} - \mathbf{X})^2)}.$$

Pues bien, imagine que $\mathbf{Y} = \mathbf{X} + \mathbb{1}_A$ donde $\mathbb{1}_A$ es perpendicular a $L(\mathbb{1})$ (es decir, donde $\mathbb{P}(A) = 0$). Si $\mathbb{1}_A$ no es la función constante cero (si $A \neq \emptyset$), entonces $\mathbf{Y} \neq \mathbf{X}$... pero ¿cuál es la distancia entre \mathbf{X} e \mathbf{Y} en ese caso? pues...

$$\text{distancia}(\mathbf{X}, \mathbf{Y})_\varphi = \|\mathbf{Y} - \mathbf{X}\|_\eta = \|\mathbb{1}_A\|_\eta = 0.$$

¡Uy!... por lo que se ve el mundo de las variables aleatorias es curioso. Es posible que dos variables que son distintas estén a distancia cero la una de la otra.⁶⁷ En esta situación, a las distancias se las denomina *semi-distancias*.

La dimensión de $L(\mathbb{1})^\perp$ puede ser enorme (incluso infinita). Eso quiere decir que puede haber infinitas funciones indicatrices (todas ellas perpendiculares a las variables aleatorias constantes) que podemos sumar a una variable aleatoria \mathbf{X} , el resultado son nuevas variables aleatorias que son distintas, pero que todas ellas están a distancia cero de \mathbf{X} (infinitas variables aleatoria que son distintas las unas de las otras, pero que distan cero las unas de las otras!).

8.1 Interpretación geométrica en un modelo para el lanzamiento de una moneda

Como esta exposición (basada en una interpretación geométrica de la probabilidad) es distinta del habitual, creo que puede ayudar al lector ver su aplicación en un ejemplo conocido: el lanzamiento de una moneda.

Para ello basta considerar que existe un *espacio semi-euclídeo probabilístico* tal que H y T sean dos subconjuntos disjuntos de un conjunto Ω (que asociamos respectivamente con los sucesos “cara” y “cruz”) y tales que $H \cup T = \Omega$; es decir, tales que $\mathbb{1}_H + \mathbb{1}_T = \mathbb{1}$. ¡Eso es todo! No necesitamos mayor especificación.

Dos representaciones esquemáticas de un modelo así aparecen en la Figura 8. En ambas aparecen la recta $L(\mathbb{1})$, la función indicatriz $\mathbb{1}_H$ que toma el valor 1 en H (“cara”) y la función indicatriz $\mathbb{1}_T$ que toma el valor 1 en T (“cruz”).

Puesto que la suma de ambas funciones indicatrices (de ambos vectores) es $\mathbb{1}$, en este caso $L(\mathbb{1}_H, \mathbb{1}_T) = L(\mathbb{1}_H, \mathbb{1}_T)$; es decir, este subespacio probabilístico es de dimensión 2 (un plano). Dicho plano contiene las infinitas variables aleatorias que son combinación lineal de los vectores $\mathbb{1}_H$ y $\mathbb{1}_T$.

⁶⁷En ese caso se dice que las variables aleatorias *son iguales salvo en un conjunto de probabilidad (o medida) nula*.

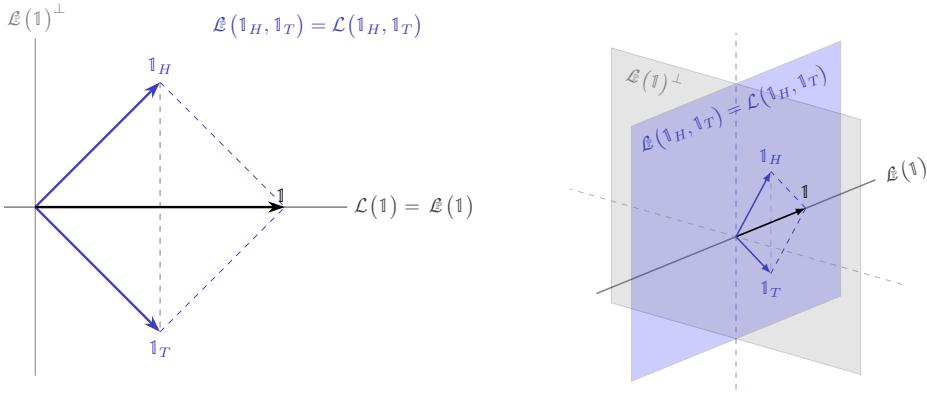


Figure 8: Dos representaciones esquemáticas del lanzamiento de una moneda “justa”.

En ambas representaciones aparece el conjunto de variables aleatorias perpendiculares a las constantes. Es aquí donde radica la diferencia entre la representación de la izquierda y la de la derecha. A la izquierda, $\mathcal{L}(1)^\perp$ tiene dimensión 1 (es la recta vertical) y a la derecha tiene dimensión 2 (es el plano ortogonal a la recta $\mathcal{L}(1)$). En realidad, $\mathcal{L}(1)^\perp$ podría tener cualquier dimensión: desde dimensión 1 hasta dimensión infinita. Cada uno de esos casos sería un modelo alternativo para el lanzamiento de una moneda... y todos ellos serían igualmente válidos.

Fíjese que nada hemos dicho del contenido de Ω . Es irrelevante para la modelización. Ω podría tener solo dos números, $\Omega = \{\pi, 17\}$, de manera que $H = \{\pi\}$ y $T = \{17\}$, por tanto $\mathbb{1}_H(\pi) = 1$ (“cara”) y $\mathbb{1}_T(17) = 1$ (“cruz”). Pero podría ser que Ω fuera el conjunto de números naturales y que $\mathbb{1}_H$ solo tomara el valor 1 con los números pares y $\mathbb{1}_T$ con los impares (ó el valor 1 para el número 1492 y cero para todos los demás). O quizás Ω contiene otro tipo de objetos matemáticos (vectores, matrices, funciones, conjuntos, etc.) que se puedan relacionar con las condiciones físicas del lanzamiento de la moneda. Todo eso queda abstraído al decir que nos fijamos en dos subconjuntos disjuntos H y T cuya unión es Ω (*lo que haya en dichos conjuntos es irrelevante en la modelización*).

Probabilidad de un suceso y proyección ortogonal Normalmente deseamos que un modelo probabilístico describa la frecuencia con la que cabe esperar cada resultado, pero nada de eso se ha indicado en la modelización anterior. No obstante, las representaciones particulares que se muestran en la Figura 8 corresponden a modelos del lanzamiento de una moneda en la que los resultados de cara o cruz son equiprobables puesto que las funciones indicadoras $\mathbb{1}_H$ y $\mathbb{1}_T$ tienen idéntica proyección ortogonal sobre $\mathcal{L}(1)$. Compárese con la Figura 9, que muestra la representación de un modelo de lanzamiento de una moneda “trucada”.

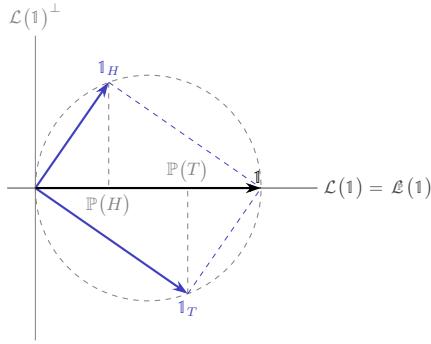


Figure 9: Representación esquemática del lanzamiento de una moneda para la que es más probable el resultado “cruz” (T) que el resultado “cara” (H).

Independencia Dado que cada *subespacio probabilístico* de (\mathcal{E}, φ) es por sí mismo un *espacio euclídeo probabilístico*, por el axioma A5 de la Definición 6, sabemos que todos ellos contienen a la variable aleatoria 1 que es la función constante igual a 1. Por tanto, dichos subespacios tienen en común la recta $\mathcal{L}(1)$. Consecuentemente, dos *subespacios probabilísticos* de un *espacio euclídeo probabilístico* \mathcal{E} no pueden ser ortogonales entre si.

Sin embargo, os *subespacios probabilísticos* pueden estar “a escuadra” (como dos paredes que se unen en una esquina for-

mando un ángulo recto). Cuando dos *subespacios probabilísticos* están “a escuadra” decimos que son *probabilísticamente independientes*. Véase la Figura 10. Así pues,

- decimos que dos *subespacios probabilísticos* \mathcal{S}, \mathcal{T} son *probabilísticamente independientes* si

$$(\mathcal{S} \cap \mathcal{L}(\mathbf{1})^\perp) \perp (\mathcal{T} \cap \mathcal{L}(\mathbf{1})^\perp)$$

donde $\mathcal{S} \cap \mathcal{L}(\mathbf{1})^\perp = \{\mathbf{X} \in \mathcal{S} \mid \langle \mathbf{X} | \mathbf{1} \rangle_\eta = 0\}$

- decimos que los *subespacios probabilísticos* de una familia $\{\mathcal{S}_i\}_{i \in I}$ son *p. independientes* si cada \mathcal{S}_i es p. independiente de $\mathcal{L}(\bigcup_{j \neq i} S_j)$
- decimos que las *funciones (variables aleatorias)* de una familia $\{\mathbf{X}_i\}_{i \in I}$ son *p. independientes* si lo son los subespacios de la familia $\{\mathcal{L}(\mathbf{X}_i)\}_{i \in I}$
- decimos que los *sucesos* de una familia $\{A_i\}_{i \in I}$ son *p. independientes* si lo son los subespacios de la familia $\{\mathcal{L}(\mathbb{1}_{A_i})\}_{i \in I}$

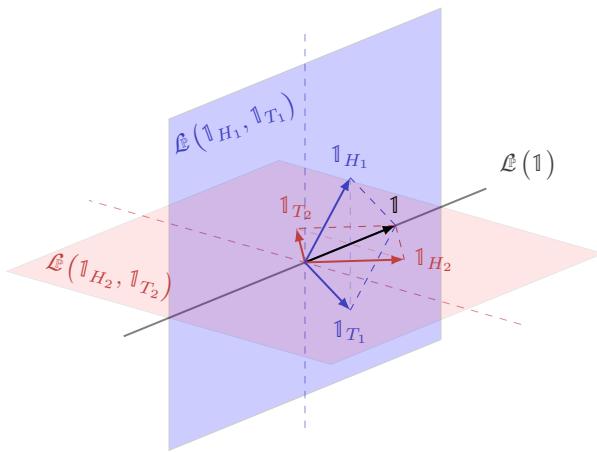


Figure 10: Representación esquemática de dos subespacios probabilísticos independientes (uno en horizontal y otro en vertical) correspondientes al lanzamiento de dos monedas “justas”. Esta representación esquemática es muy incompleta: dado que aparecen cuatro funciones indicatrices (necesarias para cubrir todos los posibles resultados $\{HH, HT, TH, TT\}$), el menor subespacio probabilístico que las contiene ¡debe ser de dimensión cuatro... algo imposible de representar en un dibujo!

8.2 “Clases de equivalencia” de variables aleatorias y la esperanza condicional

8.2.1 “Clases de equivalencia” de variables aleatorias

Decimos que una *variable aleatoria es constante casi seguro* si toma un valor constante c “con probabilidad 1”. Esto quiere decir que puede haber sucesos elementales ω para los que la variable NO toma el valor c , pero que el subconjunto de sucesos elementales que da lugar a un valor distinto de c tiene probabilidad nula. Estas variables se obtienen sumando a una variable aleatoria constante, cualquier vector ortogonal a \mathcal{E}^\perp (recuerde lo dicho sobre los sucesos de probabilidad nula en la página 56).

A una variable aleatoria genérica *igual a uno casi seguro* la denotaremos con $\mathbf{1}$ y a una variable aleatoria genérica *igual a cero casi seguro* con $\mathbf{0}$. Así por ejemplo, los vectores constantes *iguales a 1 casi seguro* son de la forma⁶⁸

$$\{\mathbf{1} \in \mathcal{E} \mid \text{existe } \mathbf{N} \in \mathcal{E}^\perp \text{ tal que } \mathbf{1} = \mathbf{1} + \mathbf{N}\}, \quad (29)$$

y todos ellos están a distancia cero de $\mathbf{1}$ (y a distancia cero los unos de los otros). Evidentemente, el tamaño de cada clase (el número de vectores que contiene) depende del tamaño del \mathcal{E}^\perp .⁶⁹

⁶⁸En el espacio euclídeo usual \mathbb{R}^N , el subconjunto \mathbb{R}^{N^\perp} ortogonal a \mathbb{R}^N solo contiene el vector nulo, $\{\mathbf{0}\}$. Sin embargo, en un \mathcal{E} genérico, el subconjunto de \mathcal{E}^\perp ortogonal a \mathcal{E} contiene las variables aleatorias de \mathcal{E} con norma cero. La dimensión de dicho subespacio puede ser cero, puede ser finita o incluso puede ser de dimensión infinita.

⁶⁹Fíjese cómo se parece esta caracterización al conjunto de soluciones de un sistema de ecuaciones lineales compatible indeterminado, donde la solución general es una particular más cualquier vector solución del sistema homogéneo (o núcleo). Por ese motivo, podemos

Lo mismo se puede hacer con cualquier variable aleatoria \mathbf{X} . Cuando sumamos a \mathbf{X} una variable aleatoria no nula y perteneciente al subespacio $\mathcal{L}(\mathbf{1})^\perp$, obtenemos una nueva variable aleatoria que toma los mismos valores que \mathbf{X} salvo en un conjunto de puntos de probabilidad cero. Al conjunto de variables aleatorias que son iguales a \mathbf{X} salvo por un conjunto de sucesos de probabilidad nula lo llamaremos la *clase de equivalencia* de la variable aleatoria \mathbf{X} (o sencillamente la clase de \mathbf{X}) que denotaré con $[\![\mathbf{X}]\!]$

$$[\![\mathbf{X}]\!] = \{\mathbf{Z} \in \mathcal{E} \mid \text{existe } \mathbf{N} \in \mathcal{E}^\perp \text{ tal que } \mathbf{Z} = \mathbf{X} + \mathbf{N}\}.$$

Así, a la clase de la Ecuación 29 la denotaremos con $[\![\mathbf{1}]\!]$; y de manera análoga tenemos la clase $[\![\mathbf{0}]\!]$.

Finalmente, decimos que dos vectores pertenecen a la misma clase si su diferencia pertenece a \mathcal{E}^\perp .

$$\mathbf{Y}, \mathbf{Z} \in [\![\mathbf{X}]\!] \text{ si y solo si } \mathbf{Y} - \mathbf{Z} \in \mathcal{E}^\perp;$$

en este caso también se dice que dichas variables aleatorias son *iguales casi seguro* (que en la literatura suele expresarse con $P(\mathbf{Y} = \mathbf{Z}) = 1$). Otra forma de decirlo es que dos variables aleatorias pertenecen a la misma clase (son iguales casi seguro) si la distancia entre ellas es cero. Evidentemente, todos los vectores de una clase tienen la misma longitud.

Operaciones con clases de equivalencia Para operar con las clases usaremos cualquiera de sus elementos como representante con el que operar, ya que con las clases de equivalencia se verifica que

$$[\![\mathbf{X}]\!] + [\![\mathbf{Y}]\!] = [\![\mathbf{X} + \mathbf{Y}]\!]; \quad \text{y también} \quad [a\mathbf{X}] = a[\![\mathbf{X}]\!] \text{ para cualquier } a \in \mathbb{R}.$$

Es más, *el conjunto de clases de equivalencia de \mathcal{E} es un espacio de Hilbert* (espacio euclídeo de dimensión infinita) cuyo vector nulo, $[\![\mathbf{0}]\!]$, es la clase de las variables aleatorias constantes cero *casi seguro*.

Las clases se pueden multiplicar: $[\![\mathbf{X}]\!] \cdot [\![\mathbf{Y}]\!] = [\![\mathbf{X} \cdot \mathbf{Y}]\!]$; y además $E([\![\mathbf{X}]\!]) = E(\mathbf{X})$. En este subespacio la esperanza del producto entre clases $E([\![\mathbf{X}]\!][\![\mathbf{Y}]\!])$ es definido, puesto que si $E([\![\mathbf{X}]\!]^2) = 0$, entonces necesariamente $[\![\mathbf{X}]\!] = [\![\mathbf{0}]\!]$. Así, este subespacio de clases junto la esperanza del producto entre clases es un espacio euclídeo. Además, $E([\![\mathbf{1}]\!] \cdot [\![\mathbf{1}]\!]) = 1$, por lo que la norma de $[\![\mathbf{1}]\!]$ es uno, $\|[\![\mathbf{1}]\!]\|_\eta = 1$. De igual manera $\|[\![\mathbf{0}]\!]\|_\eta = 0$.

8.2.2 La esperanza condicional

Esperanza condicional. Consideremos una variable aleatoria \mathbf{Y} de \mathcal{E} y un subespacio de probabilidad $\mathcal{S} \subset \mathcal{E}$. Podemos preguntarnos ¿cuál de las variables aleatorias de \mathcal{S} es la que está más próxima a \mathbf{Y} ?

Resulta que ésta es una pregunta con trampa ya que no necesariamente hay una única variable aleatoria que sea la que esté más próxima... puede haber infinitas a la misma distancia mínima. Si $\mathbf{Z} \in \mathcal{S}$ está a distancia mínima de \mathbf{Y} , siempre le podremos sumar otra variable de la intersección $\mathcal{E}^\perp \cap \mathcal{S}$ obteniendo una variable aleatoria de \mathcal{S} que estará a distancia cero de \mathbf{Z} y, por tanto, a la misma distancia de \mathbf{Y} que \mathbf{Z} . Evidentemente, solo si la intersección $\mathcal{E}^\perp \cap \mathcal{S} = \{\mathbf{0}\}$ (i.e., si solo contiene a la indicatriz nula) “la” variable aleatoria más próxima es \mathbf{Z} .

Así que debemos cambiar la pregunta a ¿cuál de las *clases* de variables aleatorias contenidas en \mathcal{S} es la que está más próxima a \mathbf{Y} ? Donde las clases son de forma

$$[\![\mathbf{Z}]\!]_{\mathcal{S}} = \{\mathbf{W} \in \mathcal{S} \mid \text{existe } \mathbf{N} \in \mathcal{S} \cap \mathcal{E}^\perp \text{ tal que } \mathbf{W} = \mathbf{Z} + \mathbf{N}\}.$$

La respuesta es que dicha clase es la *proyección ortogonal de \mathbf{Y} sobre el subespacio de probabilidad \mathcal{S}* . Es decir, en este contexto de los espacios (semi-)euclídeos de probabilidad, la proyección ortogonal es una *clase de equivalencia* (es decir, puede contener muchas variables aleatorias) a la que llamamos *esperanza condicional*. Por tanto, la esperanza condicional $E(\mathbf{Y} | \mathcal{S})$ es la clase de equivalencia $[\![\mathbf{Z}]\!]$ de vectores de $\mathcal{L}(\mathbf{X})$ que están a distancia mínima de \mathbf{Y} ; es decir,

$$E(\mathbf{Y} | \mathcal{S}) = [\![\mathbf{Z}]\!] \subset \mathcal{L}(\mathbf{X}) \text{ tal que } [\![\mathbf{Y}]\!] - [\![\mathbf{Z}]\!] \text{ es ortogonal a } \mathcal{L}(\mathbf{X}).$$

Sin embargo, en la literatura se simplifica y se dice que “la” esperanza condicional es una *variable aleatoria*. Si se consideran las esperanzas condicionales como variables aleatorias, se debería omitir el artículo determinado “la” y en cada caso referirse a “una” esperanza condicional (a “una” de las muchas variables aleatorias de $\mathcal{L}(\mathbf{X})$ que están a distancia mínima de \mathbf{Y}). Si, por el contrario, queremos que la esperanza condicional sea *única*, entonces es necesario considerarla como “la” clase de equivalencia de las variables aleatorias de $\mathcal{L}(\mathbf{X})$ que están a distancia mínima de \mathbf{Y} .

Fin de la lección

denominar al subespacio probabilístico \mathcal{E}^\perp el “núcleo” del producto escalar.

LECCIÓN 5: Especificación y Estimación del Modelo Lineal General

9 Modelo Clásico de Regresión Lineal

En la lección anterior hemos indicado que la descomposición ortogonal $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$ en \mathbb{R}^N es un caso particular de la descomposición ortogonal $\mathbf{Y} = \mathbb{E}(\mathbf{Y}|\mathbf{X}) + \mathbf{U}$ en un Espacio Euclídeo Probabilístico \mathcal{E} .

- $\hat{\mathbf{y}}$ es la proyección ortogonal de \mathbf{y} (vector de \mathbb{R}^N) sobre el subespacio vectorial $\mathcal{C}(\mathbf{X})$ generado por las k columnas de $\mathbf{X} = [\mathbf{X}_{|1}; \mathbf{X}_{|2}; \dots; \mathbf{X}_{|k}]$ (los regresores); y por tanto $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathcal{C}(\mathbf{X})$. Consecuentemente $\hat{\mathbf{e}} \perp \mathcal{C}(\mathbf{X})$.
- $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ es la proyección ortogonal de \mathbf{Y} (variable aleatoria) sobre el subespacio probabilístico $\mathcal{L}(\mathbf{X})$ generado por las k variables aleatorias del sistema $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_k]$ (los regresores); y por tanto $\mathbb{E}(\mathbf{Y}|\mathbf{X}) \in \mathcal{L}(\mathbf{X})$. Consecuentemente $\mathbf{U} \perp \mathcal{L}(\mathbf{X})$.

Los subespacios vectoriales $\mathcal{C}(\mathbf{X})$ y $\mathcal{L}(\mathbf{X})$ están formados por el conjunto de combinaciones lineales de los respectivos k regresores (\mathbf{X} en el primer caso y \mathbf{X} en el segundo) y consecuentemente, si los regresores son linealmente independientes, ambos subespacios tienen de dimensión k . Pero los Espacios Euclídeos de Probabilidad son “generalmente” de dimensión *infinita*. Es decir, $\mathcal{L}(\mathbf{X})$ suele ser un conjunto “muy muy grande” y siempre contiene a $\mathcal{C}(\mathbf{X})$.

Puesto que $\mathcal{L}(\mathbf{X})$, además de contener todas las variables aleatorias de $\mathcal{L}(\mathbf{X})$, generalmente contiene muchísimas más, es habitual que la variable aleatoria de $\mathcal{L}(\mathbf{X})$ más próxima a \mathbf{Y} no sea la más próxima de $\mathcal{L}(\mathbf{X})$ y por tanto $\mathbb{E}(\mathbf{Y}|\mathbf{X}) \neq \mathbf{X}\boldsymbol{\beta} \subset \mathcal{L}(\mathbf{X})$. Dicho de otro modo, en general la proyección $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ “cae fuera” del subespacio de dimensión finita $\mathcal{L}(\mathbf{X})$.

El *Modelo Clásico de Regresión* consiste en asumir ciertas hipótesis que garanticen que

1. la proyección ortogonal de \mathbf{Y} “caiga dentro de $\mathcal{L}(\mathbf{X})$ ”, es decir, que $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, y
2. la varianza condicional de \mathbf{Y} es una variable aleatoria constante $\text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$.

Así, bajo los supuestos del modelo clásico de regresión, la esperanza condicional de \mathbf{Y} coincide con la proyección ortogonal de \mathbf{Y} sobre el subespacio de dimensión finita $\mathcal{L}(\mathbf{X})$, formado por las combinaciones lineales de los regresores \mathbf{X} . Gracias a ello, en la siguiente lección podremos interpretar el ajuste MCO, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, como una estimación de la esperanza condicional $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ cuando \mathbf{y} y \mathbf{X} sean muestras de \mathbf{Y} y de \mathbf{X} respectivamente.

(Lección 5) T-1 Modelo Clásico de Regresión Lineal

Modelo especial en el que la descomposición ortogonal

$$\mathbf{Y} = \mathbb{E}(\mathbf{Y}|\mathbf{X}) + \mathbf{U}$$

es tal que

- $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ (Comb. lin. regresores) (Sup. 1 y 2)
- $\text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$ (v.a. cte.) (Sup. 2 y 3)

¿QUÉ CONDICIÓN ES SUFICIENTE PARA ESTO?

F52

9.1 Los cuatro primeros supuestos en el Modelo Clásico de Regresión Lineal

(Wooldridge, 2006, Capítulos 2 y 3, Sección 6.2 y Apéndice E1)

Definición 22. El producto de un sistema $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_k]$ por un vector $\boldsymbol{\beta} \in \mathbb{R}^k$ es la combinación lineal de las variables aleatorias:

$$\mathbf{X}\boldsymbol{\beta} = \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_3 + \dots + \beta_k\mathbf{X}_k.$$

Por tanto $\mathbf{X}\boldsymbol{\beta}$ es una variable aleatoria.

9.2 Primer supuesto

El primer supuesto es que la variable aleatoria \mathbf{Y} es una combinación de las variables $\mathbf{1}, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_k$ más la perturbación \mathbf{U} . Es decir: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$.

Supuesto 1

$$Y = \mathbf{X}\beta + U$$

donde $\mathbf{X} = [\mathbf{1}; \mathbf{X}_2; \mathbf{X}_3; \dots; \mathbf{X}_k]$ y $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$

es decir,

$$Y = \underbrace{\beta_1 \mathbf{1} + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \dots + \beta_k \mathbf{X}_k}_{\mathbf{X}\beta} + U$$

F53

9.3 Segundo supuesto

El segundo supuesto consiste en que la esperanza de la perturbación condicionada a los regresores es cero

$$\mathbb{E}(U|\mathbf{X}) = \mathbf{0};$$

que implica que U es $(Y - \mathbb{E}(Y|\mathbf{X}))$, es decir, que U es ortogonal a los regresores \mathbf{X} .

Cuando se verifica este segundo supuesto se dice que *los regresores son estrictamente exógenos*

Definición 23. Definimos la *esperanza* de un sistema de k variables aleatorias $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_k]$ como el vector de \mathbb{R}^k con los valores esperados de dichas variables, es decir:

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(\mathbf{X}_1), \mathbb{E}(\mathbf{X}_2), \dots, \mathbb{E}(\mathbf{X}_k)) \in \mathbb{R}^k;$$

o con el operador selector⁷⁰: $\mathbb{E}_{jl}(\mathbf{X}) = \mathbb{E}(\mathbf{X}_{lj}) = \mathbb{E}(\mathbf{X}_j)$, que también podemos escribir como $\mathbb{E}(\mathbf{X})_{lj} = \mathbb{E}(\mathbf{X}_{lj})$.

Pues bien, el segundo supuesto del modelo clásico de regresión lineal implica los siguientes resultados:

Proposición 9.1. Si $\mathbb{E}(U|\mathbf{X}) = \mathbf{0}$, entonces $\mathbb{E}(\mathbf{X}_j U) = 0$ para $j = 1 : k$. P-1
(70)

Corolario 9.2. Si $\mathbb{E}(U|\mathbf{X}) = \mathbf{0}$, entonces $\mathbb{E}(U) = 0$. P-2
(70)

Corolario 9.3. Si $\mathbb{E}(U|\mathbf{X}) = \mathbf{0}$, entonces $\text{Cov}(U, \mathbf{X}_j) = 0$ para $j = 1, \dots, k$. P-3
(70)

Supuesto 2 y sus implicaciones

$$\boxed{\mathbb{E}(U|\mathbf{X}) = \mathbf{0}} \Rightarrow \begin{cases} \mathbb{E}(\mathbf{X}_j U) = 0 \text{ para } j = 1 : k. & \boxed{U \perp \mathbf{X}_j} \\ \mathbb{E}(U) = 0 \\ \text{Cov}(U, \mathbf{X}_j) = 0 \text{ para } j = 1 : k. \end{cases}$$

Implicación conjunta de los supuestos 1 y 2

$$\left. \begin{array}{l} Y = \mathbf{X}\beta + U \\ \mathbb{E}(U|\mathbf{X}) = \mathbf{0} \end{array} \right\} \Rightarrow \mathbb{E}(Y|\mathbf{X}) = \mathbf{X}\beta \quad (F52)$$

F54

⁷⁰nótese la asociatividad del operador selector “|” por ambos lados.

Ejemplo 7. Función de consumo (continuación) : En el modelo $C = a\mathbf{1} + bR + U$, donde C es el consumo y R la renta disponible, la estricta exogeneidad

$$\mathbb{E}(U|R) = \mathbf{0},$$

es decir, que la esperanza de la perturbación condicionada a la renta disponible R es cero, implica que si asociamos esas “otras cosas” que afectan al consumo de un individuo (características personales o culturales, obligaciones familiares, etc) con U , entonces asumimos que esas “otras cosas” son ortogonales a la renta.

Ejemplo 8. Ecuación de salarios (continuación) : la estricta exogeneidad

$$\mathbb{E}(U|\mathbf{X}) = \mathbf{0}, \quad \text{con } \mathbf{X} = [\mathit{EDUC}; \mathit{ANTIG}; \mathit{EXPER}]$$

significa que la esperanza de la perturbación condicionada a los años de educación, antigüedad y experiencia es cero. Por tanto, si asociamos esas “otras cosas” que afectan al salario de un individuo (habilidades, coeficiente intelectual, etc.) con U , entonces estaremos asumiendo que son ortogonales a los años de estudio, la antigüedad o la experiencia.

Advertencia! La estricta exogeneidad es una hipótesis muy muy restrictiva; es una asunción tan fuerte que lo más probable es que no sea realista en la inmensa mayoría de modelos empíricos.

Nótese que los supuestos 1 y 2 son suficientes para satisfacer la primera de las dos condiciones del *Modelo Clásico de Regresión Lineal* (Página 61). P-4
(70)

9.4 Tercer supuesto

El Supuesto 3 se denomina supuesto de *homocedasticidad*:

(Lección 5)

T-4

Supuesto 3: Homocedasticidad

Supuesto 3

$$\mathbb{E}(U^2|\mathbf{X}) = \sigma^2 \mathbf{1}$$

Junto con $\mathbb{E}(U|\mathbf{X}) = \mathbf{0}$ es equivalente a: $\text{Var}(U|\mathbf{X}) = \sigma^2 \mathbf{1}$

Implicación de los supuestos 2 y 3

$$\begin{aligned} \sigma^2 \mathbf{1} &= \mathbb{E}(U^2|\mathbf{X}) \\ &= \mathbb{E}\left((Y - \mathbb{E}(Y|\mathbf{X}))^2 \mid \mathbf{X}\right) \\ &= \text{Var}(Y|\mathbf{X}). \end{aligned} \quad (\text{F52})$$

F55

Nótese que los supuestos 2 y 3 son suficientes para satisfacer la segunda de las dos condiciones del *Modelo Clásico de Regresión Lineal* (Página 61).

Antes de enunciar el cuarto supuesto, veamos algunas definiciones:

Definición 24. El producto de un sistema \mathbf{X} de k variables aleatorias $[X_1; X_2; \dots; X_k]$ por una variable aleatoria U (todas pertenecientes a un espacio semi-euclídeo de probabilidad (ESP) \mathcal{E}) es el sistema de variables aleatorias de $L_{\mathcal{E}}$ (es decir, con esperanza definida) que contiene los productos de cada una de las k variables por U :

$$\mathbf{X}U = [X_1U; \dots; X_kU] = [UX_1; \dots; UX_k] = UX.$$

de manera que

$$\mathbf{X}^T U = \begin{bmatrix} X_1U \\ \vdots \\ X_kU \end{bmatrix} = \begin{bmatrix} UX_1 \\ \vdots \\ UX_k \end{bmatrix} = UX^T.$$

Definición 25. Sean dos sistemas $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_m]$ y $\mathbf{U} = [\mathbf{U}_1; \mathbf{U}_2; \dots; \mathbf{U}_n]$, de variables aleatorias de un ESP \mathcal{E} . El producto $\mathbf{X}^\top \mathbf{U}$ es la matriz de orden $m \times n$ cuyo elemento (i, j) -ésimo es la variable aleatoria de $L_{\mathcal{E}}$ resultante de multiplicar la i -ésima componente de \mathbf{X} por la j -ésima componente de \mathbf{U} .

$$\mathbf{X}^\top \mathbf{U} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} [\mathbf{U}_1; \mathbf{U}_2; \dots; \mathbf{U}_n] = \begin{bmatrix} \mathbf{X}_1 \mathbf{U}_1 & \mathbf{X}_1 \mathbf{U}_2 & \dots & \mathbf{X}_1 \mathbf{U}_n \\ \mathbf{X}_2 \mathbf{U}_1 & \mathbf{X}_2 \mathbf{U}_2 & \dots & \mathbf{X}_2 \mathbf{U}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_m \mathbf{U}_1 & \mathbf{X}_m \mathbf{U}_2 & \dots & \mathbf{X}_m \mathbf{U}_n \end{bmatrix}; \text{ es decir, } {}_{ij}(\mathbf{X}^\top \mathbf{U})_{ij} = \mathbf{X}_i \mathbf{U}_j.$$

Definición 26. Definimos la *esperanza* de $\mathbf{X}^\top \mathbf{X}$, donde $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_k]$ como la matriz simétrica de orden k tal que su elemento (i, j) -ésimo es la esperanza del producto entre la variable i -ésima, \mathbf{X}_i , y la variable j -ésima, \mathbf{X}_j :

$$E(\mathbf{X}^\top \mathbf{X}) = \begin{bmatrix} E(\mathbf{X}_1^2) & E(\mathbf{X}_1 \mathbf{X}_2) & \dots & E(\mathbf{X}_1 \mathbf{X}_k) \\ E(\mathbf{X}_2 \mathbf{X}_1) & E(\mathbf{X}_2^2) & \dots & E(\mathbf{X}_2 \mathbf{X}_k) \\ \vdots & \vdots & \ddots & \vdots \\ E(\mathbf{X}_k \mathbf{X}_1) & E(\mathbf{X}_k \mathbf{X}_2) & \dots & E(\mathbf{X}_k^2) \end{bmatrix}; \text{ es decir, } {}_{ij}(E(\mathbf{X}^\top \mathbf{X}))_{ij} = E(\mathbf{X}_i \mathbf{X}_j).$$

Nótese que ${}_{ij}E(\mathbf{X}^\top \mathbf{X}) \mathbf{c} = (E(\mathbf{X}_i \mathbf{X}_1), \dots, E(\mathbf{X}_i \mathbf{X}_k)) \mathbf{c} = \sum_{j=1}^k E(\mathbf{X}_i \mathbf{X}_j) c_j$

$$= E\left(\sum_{j=1}^k \mathbf{X}_i \mathbf{X}_j c_j\right) = E\left(\mathbf{X}_i \sum_{j=1}^k \mathbf{X}_j c_j\right) = E(\mathbf{X}_i (\mathbf{X} \mathbf{c})) = E\left({}_{ij}(\mathbf{X}^\top \mathbf{X}) (\mathbf{X} \mathbf{c})\right) = E({}_{ij}(\mathbf{X}^\top \mathbf{X}) \mathbf{c}) = {}_{ij}E(\mathbf{X}^\top \mathbf{X} \mathbf{c}).$$

9.5 Cuarto supuesto

(Lección 5)

T-5 Supuesto 4 y la identificación de los parámetros β

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{U} && \text{Por Sup. 1} \\ \mathbf{X}^\top \mathbf{Y} &= \mathbf{X}^\top \mathbf{X}\beta + \mathbf{X}^\top \mathbf{U} && \text{premultiplicando por } \mathbf{X}^\top \\ E(\mathbf{X}^\top \mathbf{Y}) &= E(\mathbf{X}^\top \mathbf{X})\beta + E(\mathbf{X}^\top \mathbf{U}) && \text{tomando esperanzas} \\ E(\mathbf{X}^\top \mathbf{Y}) &= E(\mathbf{X}^\top \mathbf{X})\beta && E(\mathbf{X}^\top \mathbf{U}) = \mathbf{0} \text{ (Sup. 2)} \end{aligned}$$

donde ${}_{ij}(E(\mathbf{X}^\top \mathbf{X}))_{ij}$ es $E(\mathbf{X}_i \mathbf{X}_j)$.

Supuesto 4

La matriz $E(\mathbf{X}^\top \mathbf{X})$ es de rango completo
entonces β está identificado: $\beta = (E(\mathbf{X}^\top \mathbf{X}))^{-1} E(\mathbf{X}^\top \mathbf{Y})$

F56

Este supuesto implica que el vector de parámetros β es único y se puede expresar en función de los momentos de las variables \mathbf{X} e \mathbf{Y} . Se dice que existe *multicolinealidad exacta o perfecta* cuando el Supuesto 4 NO se satisface.

Definición 27. Un sistema de k variables aleatorias \mathbf{X} es *linealmente dependiente* si existe un vector no nulo $\mathbf{c} \in \mathbb{R}^k$ tal que $\mathbf{X} \mathbf{c} = \mathbf{0}$. Si no existe tal vector el sistema \mathbf{X} es *linealmente independiente*.

Proposición 9.4. Si la matriz $E(\mathbf{X}^\top \mathbf{X})$ es de rango completo entonces \mathbf{X} es *linealmente independiente*. P-5
(70)

Fíjese que dado que trabajamos con semi-productos escalares y semi-normas, el recíproco no es cierto. Si θ es una variable aleatoria *constante cero casi seguro* que no es la constante nula ($\theta \neq 0$), y si $\mathbf{X} = [\theta; \mathbf{0}]$, entonces $E(\mathbf{X}^\top \mathbf{X}) = \mathbf{0}$ aunque \mathbf{X} es linealmente independiente.⁷¹

⁷¹Una condición suficiente para asegurar que $E(\mathbf{X}^\top \mathbf{X})$ es invertible es exigir que las *clases* $[\mathbf{X}_1], [\mathbf{X}_2], \dots, [\mathbf{X}_k]$ sean linealmente independientes, ya que el espacio euclídeo formado por las clases de variables aleatorias si es un Espacio de Hilbert con producto escalar (y norma). Otra condición suficiente (y alternativa a la anterior) es que $\mathcal{E}^\perp \cap \mathcal{L}(\mathbf{X}) = \{\mathbf{0}\}$, es decir, que solo contenga a la indicatriz nula.

Los supuestos 1, 2 y 4 garantizan la [unicidad de los parámetros \$\beta\$](#) ; es decir, que el vector β está *identificado*.

10 Regresión cuando \mathcal{E} es \mathbb{R}^N

Cuando \mathcal{E} es \mathbb{R}^N junto al producto escalar $\langle \cdot | \cdot \rangle_s$; es decir, cuando los vectores de \mathbb{R}^N son considerados como variables aleatorias (y no como una muestra de variables aleatorias) el sistema de ecuaciones para un \mathcal{E} genérico

$$E(\mathbf{X}^\top \mathbf{X})\beta = E(\mathbf{X}^\top \mathbf{Y})$$

se reduce a

$$\left(\frac{1}{N} \mathbf{X}^\top \mathbf{X} \right) \beta = \frac{1}{N} \mathbf{X}^\top \mathbf{y},$$

donde he escrito explícitamente la división por N correspondiente al producto escalar $\langle \cdot | \cdot \rangle_s$.

La ausencia de multicolinealidad exacta implica que la matriz $\mathbf{X}^\top \mathbf{X}$ es invertible; por tanto la solución β es

$$\beta = \left(\frac{1}{N} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{N} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

(Lección 5)

T-6 Regresión cuando \mathcal{E} es \mathbb{R}^N

$$E(\mathbf{X}^\top \mathbf{X})\beta = E(\mathbf{X}^\top \mathbf{Y}) \quad \text{se reduce a} \quad \left(\frac{1}{N} \mathbf{X}^\top \mathbf{X} \right) \beta = \frac{1}{N} \mathbf{X}^\top \mathbf{y}$$

Ausencia de multicolinealidad exacta implica que

$$\beta = \left(\frac{1}{N} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{N} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

donde

- $\left(\frac{1}{N} (\mathbf{X}^\top \mathbf{X}) \right)_{ij} = \mu_{(\mathbf{x}_{|i} \odot \mathbf{x}_{|j})}$
- $\left(\frac{1}{N} (\mathbf{X}^\top \mathbf{y}) \right) = \mu_{(\mathbf{x}_{|i} \odot \mathbf{y})}$

F57

es decir, $\frac{1}{N} (\mathbf{X}^\top \mathbf{X})$ es una matriz con las medias de los productos Hadamard (componente a componente) entre los regresores; $\mu_{(\mathbf{x}_i \odot \mathbf{x}_j)}$; y por otra parte $\frac{1}{N} \mathbf{X}^\top \mathbf{y}$ es un vector con las medias de los productos Hadamard (componente a componente) entre los regresores y el regresando ($\mu_{(\mathbf{x}_i \odot \mathbf{y})}$). Si el primer regresor es el vector **1**, entonces tenemos

$$N^{-1}(\mathbf{X}^\top \mathbf{X}) = \begin{bmatrix} \mu(\mathbf{1}) & \mu(\mathbf{x}_2) & \cdots & \mu(\mathbf{x}_k) \\ \mu(\mathbf{x}_2) & \mu(\mathbf{x}_2^2) & \cdots & \mu(\mathbf{x}_2 \odot \mathbf{x}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mu(\mathbf{x}_k) & \mu(\mathbf{x}_k \odot \mathbf{x}_2) & \cdots & \mu(\mathbf{x}_k^2) \end{bmatrix} \quad \text{y} \quad N^{-1} \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} \mu(\mathbf{y}) \\ \mu(\mathbf{x}_2 \odot \mathbf{y}) \\ \vdots \\ \mu(\mathbf{x}_k \odot \mathbf{y}) \end{pmatrix}.$$

10.1 Dos casos particulares de MLG

10.1.1 Modelo con una constante como único regresor

(Lección 5) T-7 Cte. como único regresor

$$\mathbf{X} = [\mathbf{1};] \quad \rightarrow \quad \mathbf{Y} = \mathbb{E}(\mathbf{Y}|\mathbf{1}) + \mathbf{U};$$

$$\begin{aligned} \mathbb{E}(\mathbf{X}^\top \mathbf{X}) \boldsymbol{\beta} &= \mathbb{E}(\mathbf{X}^\top \mathbf{Y}) && (\text{donde } \mathbf{X} = [\mathbf{1};]); \\ \mathbb{E}(\mathbf{1} \cdot \mathbf{1}) \boldsymbol{\beta} &= \mathbb{E}(\mathbf{1} \cdot \mathbf{Y}) \Rightarrow \boldsymbol{\beta} = \mathbb{E}(\mathbf{Y}) \end{aligned}$$

Cuando el EEP es \mathbb{R}^N con $\langle \cdot | \cdot \rangle_s$ tenemos

$$\frac{1}{N} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \frac{1}{N} \mathbf{X}^\top \mathbf{y} \quad (\text{donde } \mathbf{X} = [\mathbf{1};]);$$

así

$$\frac{1}{N} [\mathbf{1} \cdot \mathbf{1}] \boldsymbol{\beta} = \frac{1}{N} (\mathbf{1} \cdot \mathbf{y},) \Rightarrow \boldsymbol{\beta} = \mu_{\mathbf{y}}$$

F58

10.1.2 Modelo Lineal Simple

(Lección 5) T-8 Modelo Lineal Simple (Modelo teórico)

$$\mathbf{X} = [\mathbf{1}; \mathbf{X};] \quad \rightarrow \quad \mathbf{Y} = \mathbb{E}(\mathbf{Y}|\mathbf{X}) + \mathbf{U};$$

$$\begin{aligned} \mathbb{E}(\mathbf{X}^\top \mathbf{X}) \boldsymbol{\beta} &= \mathbb{E}(\mathbf{X}^\top \mathbf{Y}) \\ \begin{bmatrix} \mathbb{E}(\mathbf{1}) & \mathbb{E}(\mathbf{X}) \\ \mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{X}^2) \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} &= \begin{pmatrix} \mathbb{E}(\mathbf{Y}) \\ \mathbb{E}(\mathbf{XY}) \end{pmatrix} \end{aligned}$$

cuya solución es

$$\beta_1 = \mathbb{E}(\mathbf{Y}) - \beta_2 \mathbb{E}(\mathbf{X}) \quad y \quad \beta_2 = \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\text{Var}(\mathbf{X})} \quad (30)$$

En \mathbb{R}^N con $\langle \cdot | \cdot \rangle_s$:

$$\beta_1 = \mu_{\mathbf{y}} - \beta_2 \mu_{\mathbf{x}} \quad y \quad \beta_2 = \frac{\sigma_{\mathbf{xy}}}{\sigma_x^2}$$

Supuesto 4 (indep. lineal de regresores) garantiza $\rightarrow \sigma_x^2 \neq 0$

F59

Resolviendo $\mathbb{E}(\mathbf{X}^\top \mathbf{X}) \boldsymbol{\beta} = \mathbb{E}(\mathbf{X}^\top \mathbf{Y})$ por eliminación Gaussiana (por columnas):

$$\begin{array}{c|cc|c} 1 & \mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{X}^2) & -\mathbb{E}(\mathbf{Y}) \\ \hline \mathbb{E}(\mathbf{X}) & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 \end{array} \xrightarrow{\left[\begin{array}{c|cc|c} (-\mathbb{E}(\mathbf{X}))\mathbf{1}+2 & \mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{X}^2) & -\mathbb{E}(\mathbf{Y}) \\ [(\mathbb{E}(\mathbf{Y}))\mathbf{1}+3] & 1 & -\mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{Y}) \end{array} \right]} \begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline \mathbb{E}(\mathbf{X}) & 1 & -\mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{Y}) \\ \hline 0 & 1 & 0 & 0 \end{array} \xrightarrow{\left[\begin{array}{c|cc|c} (\frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\text{Var}(\mathbf{X})})\mathbf{2}+3 & \mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{X}^2) & 0 \\ \hline 1 & -\mathbb{E}(\mathbf{X}) & \mathbb{E}(\mathbf{Y}) & 0 \end{array} \right]} \begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline \mathbb{E}(\mathbf{X}) & 1 & -\mathbb{E}(\mathbf{X}) & -\frac{\text{Cov}(\mathbf{X}, \mathbf{Y})\mathbb{E}(\mathbf{X})}{\text{Var}(\mathbf{X})} + \mathbb{E}(\mathbf{Y}) \\ \hline 0 & 1 & \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\text{Var}(\mathbf{X})} & 0 \\ \hline 0 & 0 & 1 & 0 \end{array}.$$

Por tanto, $\beta_1 = \mathbb{E}(\mathbf{Y}) - \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\text{Var}(\mathbf{X})} \mathbb{E}(\mathbf{X})$ y $\beta_2 = \frac{\text{Cov}(\mathbf{X}, \mathbf{Y})}{\text{Var}(\mathbf{X})}$.

11 Estimación del Modelo Clásico de Regresión Lineal

Consideremos el siguiente modelo clásico de regresión lineal

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}, \quad \text{donde } \mathbf{X} = [\mathbf{1}; \mathbf{X}_2; \dots; \mathbf{X}_k;],$$

que cumple los cuatro supuestos vistos anteriormente; y donde las correspondientes variables aleatorias del modelo pertenecen a un mismo *Espacio Semi-euclídeo de Probabilidad* (ESP) \mathcal{E} ; es decir, todas las variables aleatorias son funciones que van de un mismo conjunto de sucesos elementales Ω a \mathbb{R} .

Suponga que el vector $\boldsymbol{\beta}$ es desconocido y que lo queremos calcular. Sabemos que $\boldsymbol{\beta}$ está identificado, pues se cumple el Supuesto 4 (independencia lineal entre regresores); y por tanto, sabemos que $\boldsymbol{\beta}$ es la única solución del sistema

$$\mathbf{E}(\mathbf{X}^T \mathbf{X}) \boldsymbol{\beta} = \mathbf{E}(\mathbf{X}^T \mathbf{Y}). \quad (31)$$

¿Qué problema puede haber entonces para resolver dicho sistema y calcular $\boldsymbol{\beta}$?

Pues sencillamente que las variables aleatorias (funciones que van de Ω a \mathbb{R}) del modelo son desconocidas en general, y también son desconocidas la matriz de coeficientes $\mathbf{E}(\mathbf{X}^T \mathbf{X})$ y el vector del lado derecho $\mathbf{E}(\mathbf{X}^T \mathbf{Y})$ del sistema de más arriba. Así pues, al no poder resolver el sistema, tampoco podemos conocer el vector $\boldsymbol{\beta}$.

Como alternativa, podemos tratar de estimar dicho vector $\boldsymbol{\beta}$, pero para ello necesitamos dar ¡un *salto!* Y es un salto un tanto *audaz*. Veámoslo con un ejemplo que ya hemos empleado anteriormente.

Suponga que asociamos el precio de la vivienda con la función \mathbf{Y} y el tamaño de la vivienda con la función \mathbf{X} . Además suponga que asumimos (quizá erróneamente) el siguiente *modelo de referencia* que relaciona ambas variables aleatorias:

$$\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \mathbf{U}$$

Si además asumimos que este modelo cumple los cuatro supuestos del modelo clásico general (que es mucho asumir), estaremos asumiendo que el precio de una vivienda es una cantidad constante $\beta_1 \mathbf{1}$; más un múltiplo de su superficie: $\beta_2 \mathbf{X}$; más algo que es ortogonal al subespacio de probabilidad generado por los regresores y que denotamos con \mathbf{U} .

Pese a todas estas asunciones (quizá todas erróneas) seguiremos sin conocer el valor de β_1 y β_2 ; ni tampoco el detalle de las funciones \mathbf{Y} , \mathbf{X} y \mathbf{U} .

¿Y qué podemos hacer?... pues haremos lo siguiente: cuando observemos tamaño y precio de venta de un conjunto (*una muestra*) de viviendas que se hayan intercambiado en el mercado inmobiliario, actuaremos como si los datos de cada una de las viviendas fuera la *realización* de una “copia” del modelo de referencia. Es decir, asumiremos que el precio de la vivienda n -ésima sigue un modelo $\mathbf{Y}_n = \beta_1 \mathbf{1} + \beta_2 \mathbf{X}_n + \mathbf{U}_n$ con *idéntica distribución* al modelo de referencia. Además asumiremos que todas las variables aleatorias asociadas a todas las “copias” pertenecen a un mismo y único ESP \mathcal{E} . Por tanto estaremos asumiendo que los datos son los valores tomados por dichas variables aleatorias al ser evaluadas (todas ellas) en un mismo y único punto $\omega \in \Omega$. También asumiremos que las variables aleatorias (no constantes) correspondientes a cada vivienda son *independientes* de las variables aleatorias (no constantes) correspondientes a las otras viviendas. Es decir (recordando la interpretación geométrica de la independencia probabilística), asumiremos que cada “copia” está contenida en un subespacio de probabilidad que está “a escuadra” respecto a los demás subespacios (recuérdese la Figura 10 en la página 58).

Consecuentemente, si disponemos de datos de 14 viviendas, diremos que son una *muestra aleatoria simple* de las variables aleatorias \mathbf{X} (superficie) e \mathbf{Y} (precio) del modelo de referencia. Con dicha expresión estaremos indicando que asumimos que los datos x_n e y_n son valores tomados por 14 copias independientes e idénticamente distribuidas; es decir, por 14 funciones cuyos rangos (el conjunto de valores que toman dichas funciones) son iguales a los de las variables \mathbf{X} e \mathbf{Y} del modelo de referencia y con distribuciones de probabilidad iguales a las de las variables del modelo de referencia.

Y ahora viene el “audaz salto” al que aludimos más arriba: en lugar de trabajar con las variables aleatorias \mathbf{X} e \mathbf{Y} del modelo de referencia, emplearemos los vectores de datos de la muestra $\mathbf{x} = (x_1, \dots, x_N)$ e $\mathbf{y} = (y_1, \dots, y_N)$. Por tanto, en lugar de trabajar con las variables aleatorias \mathbf{X} e \mathbf{Y} , que son funciones que van de un conjunto Ω a \mathbb{R} , usaremos listas de números, que asumimos que son los valores de que toman las variables aleatorias \mathbf{X}_j e \mathbf{Y}_j ⁷² al ser evaluadas en un único punto ω del dominio Ω común a todas ellas.

⁷²que respectivamente son independientes y con idéntica distribución a \mathbf{X} e \mathbf{Y}

Evidentemente, si \mathbf{y} es solo una *muestra* de una variable aleatoria \mathbf{Y} , entonces la *media* de \mathbf{y} NO es la *esperanza* de $E(\mathbf{Y})$. Consecuentemente, no es lo mismo la media de un vector \mathbf{y} cuando se considera que \mathbf{y} es una *variable aleatoria* del espacio euclídeo de probabilidad (\mathbb{R}^N, s) que cuando \mathbf{y} es una *muestra* de una variable aleatoria. Usaremos el término *media muestral* (que denotaremos con $m_{\mathbf{y}}$) para la media de una *muestra* \mathbf{y} de una variable aleatoria \mathbf{Y} .

Lo mismo ocurre con otros estadísticos calculados a partir de vectores de datos que son interpretados como muestras de variables aleatorias. Consecuentemente, al tratar con muestras de **datos** hablaremos de *media muestral*, *varianza muestral*, , *correlación muestral*, etc. Y sustituiremos las letras griegas μ , σ y ρ por las letras latinas m , s y r . (véase la Tabla 6).

\mathcal{E} genérico	$\mathcal{E} = \mathbb{R}^N$	Muestra
$E(\mathbf{Y})$	$\mu_{\mathbf{y}}$	$m_{\mathbf{y}}$
$\text{Var}(\mathbf{Y})$	$\sigma_{\mathbf{y}}^2$	$s_{\mathbf{y}}^2$
$\text{Cov}(\mathbf{X}, \mathbf{Y})$	$\sigma_{\mathbf{x}\mathbf{y}}$	$s_{\mathbf{x}\mathbf{y}}$
$\text{Corr}(\mathbf{X}, \mathbf{Y})$	$\rho_{\mathbf{x}\mathbf{y}}$	$r_{\mathbf{x}\mathbf{y}}$

Table 6: Notación de momentos teóricos o poblacionales (mediante abreviaturas o con letras griegas cuando $\mathcal{E} = \mathbb{R}^N$) y de momentos muestrales (con letras latinas)

Retomando la discusión de más arriba, queríamos calcular β en la ecuación

$$E(\mathbf{X}^T \mathbf{X}) \beta = E(\mathbf{X}^T \mathbf{Y}).$$

pero no es posible, pues $E(\mathbf{X}^T \mathbf{X})$ y $E(\mathbf{X}^T \mathbf{Y})$ son desconocidos en general. Como las componentes de la matriz y del vector anteriores son momentos teóricos de productos de las variables aleatorias

$${}_{ij}(E(\mathbf{X}^T \mathbf{X}))_{ij} = E(\mathbf{X}_i \mathbf{X}_j) \quad \text{y} \quad {}_{ij}(E(\mathbf{X}^T \mathbf{Y})) = E(\mathbf{X}_i \mathbf{Y})$$

vamos a sustituir la matriz y el vector anteriores por $\frac{1}{N}(\mathbf{X}^T \mathbf{X})$ y $\frac{1}{N}(\mathbf{X}^T \mathbf{y})$, donde

$${}_{ij}\left(\frac{1}{N}(\mathbf{X}^T \mathbf{X})\right)_{ij} = m_{(\mathbf{X}_{ij} \odot \mathbf{X}_{ij})} \quad \text{y} \quad {}_{ij}\left(\frac{1}{N}(\mathbf{X}^T \mathbf{y})\right) = m_{(\mathbf{X}_{ij} \odot \mathbf{y})}.$$

Así, haciendo la sustitución en la Ecuación 31 y multiplicando ambos lados por N , obtenemos las ecuaciones normales

$$(\mathbf{X}^T \mathbf{X}) \hat{\beta} = \mathbf{X}^T \mathbf{y}.$$

Y si las columnas de \mathbf{X} son linealmente independientes, la solución $\hat{\beta}$ es una estimación de β .

Esta sustitución de los momentos de las variables aleatorias de un modelo por momentos muestrales de un conjunto de datos para estimar los parámetros del modelo se denomina *Estimación por Método de los Momentos*.

En el caso descrito más arriba, el *Método de los Momentos* resulta ser una estimación por MCO, pues los parámetros estimados son la solución a un sistema de ecuaciones normales.

11.1 Estimación de un modelo clásico de regresión con una muestra de datos

La estimación MCO de un modelo clásico de regresión $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, donde $\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k]$, consiste en ajustar por MCO una muestra \mathbf{y} del regresando \mathbf{Y} con una muestra $\mathbf{X} = [\mathbf{X}_{11}; \dots; \mathbf{X}_{1k}]$ de los regresores \mathbf{X} . Al realizar el ajuste MCO, y dado que en este caso los vectores son *muestras*, en lugar de escribir los momentos teóricos o poblacionales debemos indicar con la notación que los momentos son *muestrales*.

En tal caso, el vector $\hat{\beta}$ obtenido al resolver el sistema de ecuaciones normales es “una”⁷³ estimación por mínimos cuadrados ordinarios (MCO) del vector β de parámetros del modelo. Análogamente, el vector ajustado por MCO $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ se interpreta como una estimación de $E(\mathbf{Y} | \mathbf{X})$. Y dado un vector \mathbf{z} , se dice que $\mathbf{z}\hat{\beta}$ es la predicción puntual del modelo estimado para el caso en que los regresores toman los valores \mathbf{z} .

⁷³Se dice que es “una” estimación ya que con otra muestra distinta posiblemente obtendríamos resultados diferentes, es decir, “otra” estimación.

Si \mathbf{y} es una *muestra* de \mathbf{Y} y \mathbf{X} una *muestra* de \mathbf{X} ; y si se asume que $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$ es un modelo clásico de regresión que cumple los supuestos (y si $\mathbf{X}^\top \mathbf{X}$ es invertible)

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

donde

- $\left(\frac{1}{N}(\mathbf{X}^\top \mathbf{X})\right)_{ij} = m_{(\mathbf{X}_{|i} \odot \mathbf{X}_{|j})}$
- $\left(\frac{1}{N}(\mathbf{X}^\top \mathbf{y})\right) = m_{(\mathbf{X}_{|i} \odot \mathbf{y})}$

Método de los Momentos

F56

F60

Sea $\mathbf{Y} = a\mathbf{1} + b\mathbf{X} + \mathbf{U}$; si disponemos de una muestra

$$\mathbf{y} \in \mathbb{R}^N, \quad \mathbf{X} = [\mathbf{1}; \mathbf{x}]_{N \times 2}$$

resolviendo $\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X} \hat{\beta}$ con $\hat{\beta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$, obtenemos (Compárese con Ecuación 15, página 22)

$$\hat{b} = \frac{s_{xy}}{s_x^2} \quad \text{y} \quad \hat{a} = m_y - \hat{b} m_x$$

La estimación MCO sustituye los momentos teóricos por los muestrales (método de los momentos)
La indep. lineal de regresores garantiza $\rightarrow s_x^2 \neq 0$

F61

Nótese como las estimaciones MCO consisten en sustituir los momentos teóricos de la Ecuación (30) por sus análogos muestrales.

Ejemplo 9. Precio de las viviendas (continuación): Planteamos el modelo $\mathbf{Y} = a\mathbf{1} + b\mathbf{X} + \mathbf{U}$, donde \mathbf{Y} es el precio de la vivienda, \mathbf{X} es la superficie, y \mathbf{U} son otros factores que influyen en el precio de la vivienda, que asumimos “ortogonales” a la superficie del mismo (ubicación de la vivienda, estado de mantenimiento, servicios, etc.) Deseamos saber cuál es el efecto *marginal* del incremento de la superficie de un piso sobre el valor esperado del precio. Por lo tanto necesitamos estimar el valor del parámetro b .

Dada la muestra de las columnas *Precio* y *Superficie* de la Tabla 7, y estimando por MCO tenemos que

$$\hat{a} = m_y - m_x \frac{s_{xy}}{s_x^2} = 52.3509 \quad \hat{b} = \frac{s_{xy}}{s_x^2} = 0.13875$$

y por lo tanto la ecuación estimada es

$$\widehat{\text{price}} = \widehat{52,3509}_{(1,404)} + \widehat{0,13875}_{(7,407)} \times \text{sqft}$$

$$N = 14 \quad \bar{R}^2 = 0,8056 \quad F(1, 12) = 54,861 \quad \hat{\sigma} = 39,023$$

(entre paréntesis, los estadísticos t)

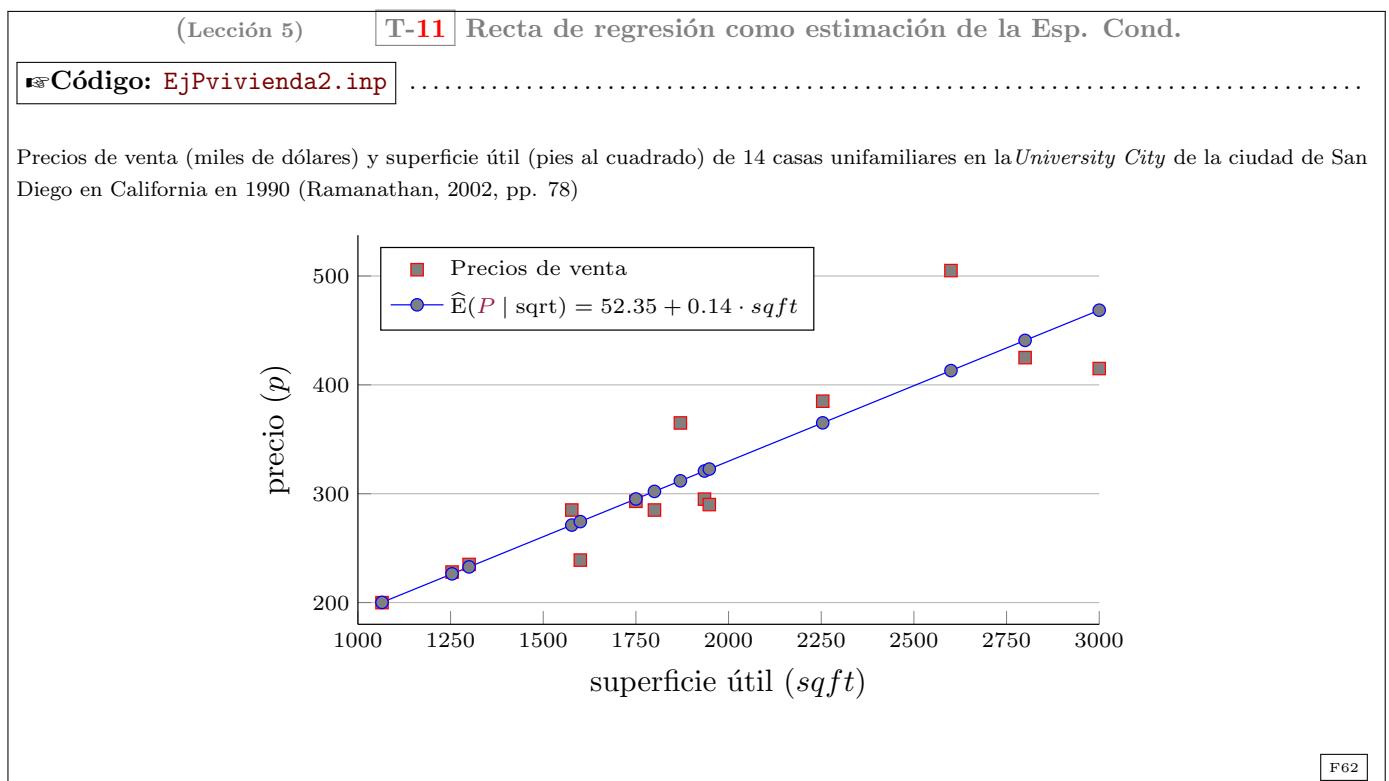
Así, la *previsión* del precio de venta de un piso con una superficie de 1800 pies cuadrados ($\text{sqft} = 1800$) es

$$\widehat{y}_7 = \widehat{52,3509} + \widehat{0,13875} \times 1800 = 302101.5.$$

La siguiente tabla muestra los resultados obtenidos para cada tamaño de casa de la muestra (no obstante, la línea punteada de la recta de regresión en la transparencia de más abajo indica que la relación estimada permite calcular el precio esperado para casa con superficies distintas de las que están presentes en la muestra empleada, basta aplicar la ecuación estimada más arriba):

1	Precio	Superficie	Precio estimado $\hat{E}(P \text{superficie})$	Error Estimación de U
1	199.9	1065	200.1200	-0.22000
2	228.0	1254	226.3438	1.65619
3	235.0	1300	232.7263	2.27368
4	285.0	1577	271.1602	13.83984
5	239.0	1600	274.3514	-35.35142
6	293.0	1750	295.1640	-2.16397
7	285.0	1800	302.1015	-17.10148
8	365.0	1870	311.8140	53.18600
9	295.0	1935	320.8328	-25.83278
10	290.0	1948	322.6365	-32.63653
11	385.0	2254	365.0941	19.90587
12	505.0	2600	413.1017	91.89826
13	425.0	2800	440.8518	-15.85180
14	415.0	3000	468.6019	-53.60187

Table 7: Superficie (en pies al cuadrado), precio de venta (en miles de dólares), estimación del precio esperado (en miles de dólares), y estimación del efecto de las perturbaciones (en miles de dólares).



En la siguiente lección se verán las propiedades estadísticas de este procedimiento. Es decir, las propiedades estadísticas de los estimadores MCO.

Problemas de la Lección 5

Los tres primeros supuestos

(L-5) PROBLEMA 1. Demuestre la Proposición 9.1 en la página 62, es decir, si $E(U|X) = \mathbf{0}$, entonces $E(X_j U) = 0$, para $j = 1 : k$.

(L-5) PROBLEMA 2. Demuestre el Corolario 9.2 en la página 62, es decir, si $E(U|X) = \mathbf{0}$, entonces $E(U) = 0$.

(L-5) PROBLEMA 3. Demuestre el Corolario 9.3 en la página 62, es decir, si $E(U|X) = \mathbf{0}$, entonces $\text{Cov}(U, X_j) = 0$.

(L-5) PROBLEMA 4. Demuestre que los supuestos 1 y 2 implican la primera condición del *Modelo Clásico de Regresión Lineal*, esto es, que la función de regresión de Y sobre los regresores es lineal: $E(Y|X) = X\beta$.

(L-5) PROBLEMA 5. Demuestre que si $E(X^T X)$ es de rango completo entonces X es *linealmente independiente*.

Regresión con muestras para estimar β

(L-5) PROBLEMA 6. Para $k = 2$ la matriz $N^{-1}(\mathbf{X}^\top)\mathbf{X}$ y vector $N^{-1}(\mathbf{X}^\top)\mathbf{y}$ resultan ser

$$\begin{bmatrix} 1 & m_{\mathbf{x}} \\ m_{\mathbf{x}} & m_{\mathbf{x}^2} \end{bmatrix}; \quad \mathbf{Y} = \begin{bmatrix} m_{\mathbf{y}} \\ m_{(\mathbf{x} \odot \mathbf{y})} \end{bmatrix}$$

- (a) Verifique las dos igualdades anteriores
- (b) Obtenga las expresiones de las ecuaciones (15) y (16) de la página 22.

(L-5) PROBLEMA 7. Intente repetir lo mismo para $k = 3$. Comprobará dos cosas: I) el cálculo es espantoso; y II) efectivamente la solución emplea los momentos muestrales entre los regresores y el regresando.

(L-5) PROBLEMA 8. Calcule la matriz de varianzas y covarianzas del estimador $\hat{\beta}$ en el modelo lineal simple.

(L-5) PROBLEMA 9.

- (a) Calcule la matriz de varianzas y covarianzas del estimador $\hat{\beta}$ en el modelo $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \beta_3 \mathbf{Z} + \mathbf{U}$.
(es un cálculo bastante penoso, si no lo logra consulte directamente la solución).
- (b) Verifique que la varianza del estimador de β_2 es $\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{(1 - R_{\mathbf{X}\mathbf{Z}}^2) \text{Var}(\mathbf{X})}$ donde $R_{\mathbf{X}\mathbf{Z}}^2$ es el coeficiente de determinación de la regresión $\mathbf{X} = \mathbf{1}\beta_1 + \mathbf{Z}\beta_2 + \mathbf{U}$, es decir, el cuadrado de la correlación entre \mathbf{X} y \mathbf{Z} ,

Fin de los Problemas de la Lección 5

Enlace a algunas **prácticas de la Lección 5**

Fin de la lección

LECCIÓN 6: Propiedades estadísticas de los estimadores MCO

Bibliografía:

Básica: Wooldridge (2006, Capítulos 1, 2 y 3 y secciones 4.1, 4.2, 6.2 y 6.3. Apéndices E1, E2 y E3)

Complementaria: Hayashi (2000, Capítulo 1)

12 Espacio Euclídeo de Probabilidad para un Muestreo Aleatorio Simple

Considere un modelo definido en un subespacio semi-euclídeo de probabilidad $\mathcal{E}(\mathbf{Z})$ de un subespacio semi-euclídeo de probabilidad $(\mathcal{E}, \eta, \Omega)$. Por ejemplo $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \mathbf{U}$ donde $\mathbf{X}, \mathbf{U} \in \mathcal{E}(\mathbf{Z})$. Llamaremos *muestreo aleatorio simple* (*m.a.s.*) de tamaño N a una familia de N copias de dicho modelo que sean independientes y con identica distribución. Es decir, N copias $\mathbf{Y}_n = \beta_1 \mathbf{1} + \beta_2 \mathbf{X}_n + \mathbf{U}_n$ cuyas variables aleatorias tienen la misma distribución que la de las respectivas variables aleatorias del modelo original, pero donde cada una de las copias está contenida en un subespacio semi-euclídeo de probabilidad $\mathcal{E}(\mathbf{Z}_n)$ de $(\mathcal{E}, \eta, \Omega)$ ⁷⁴ que corresponde a un giro de $\mathcal{E}(\mathbf{Z})$ alrededor de la recta $\mathcal{E}(\{\mathbf{1}\})$, de manera que cada copia queda “a escuadra” respecto a las demás; es decir, que sus variables aleatorias son probabilísticamente independientes respecto de las variables aleatorias de las otras copias. Un ejemplo de representación de 2 subespacios independientes se puede ver en la Figura 10 en la página 58. Evidentemente, la dimensión del espacio que contiene todos los giros crece con el número de copias, y esto ocasiona que sea imposible representar en un dibujo el corte a escuadra de más de dos subespacios. Un intento de realizar dicho dibujo se puede ver en la Figura 11:

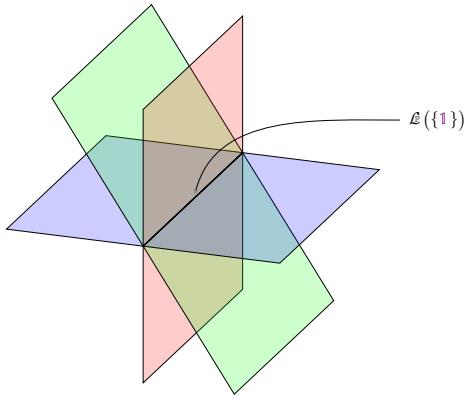


Figure 11: Intento de representación esquemática de tres subespacios probabilísticos independientes que se cortan perpendicularmente en la recta $\mathcal{E}(\{\mathbf{1}\})$ que contiene las variables aleatorias constantes (evidentemente es imposible representar el corte perpendicular entre más de dos subespacios en una figura plana como esta. Para representar fielmente el caso de tres planos sería necesaria una figura con un mínimo de $2^3 = 8$ dimensiones).

Llamaremos *muestra aleatoria simple* a la *realización* (en un suceso ω de Ω) de las variables aleatorias de un muestreo aleatorio simple. Es decir, al conjunto de valores que toman las funciones del muestreo al ser evaluadas (todas ellas) en un mismo punto ω . Así, una muestra aleatoria simple \mathbf{y} de una variable aleatoria \mathbf{Y} es un vector donde cada elemento (cada número) y_n corresponde a la realización de una variable aleatoria \mathbf{Y}_n con idéntica distribución a las demás, y donde todas ellas son probabilísticamente independientes entre si (pertencen a subespacios probabilísticos que están “a escuadra” entre ellos; véase Página 57). Consecuentemente, los valores que componen una muestra aleatoria simple de una variable \mathbf{Y} son realizaciones de variables aleatorias \mathbf{Y}_n *independientes e idénticamente distribuidas* (*i.i.d.*). En este curso siempre asumiremos que las muestras de datos son realizaciones de un *m.a.s.*⁷⁵

Las funciones definidas sobre muestras se denominan *estadísticos* (por ejemplo, la media muestral, la mediana, la varianza muestral, etc.).

Como ilustración, volvamos al ejemplo del lanzamiento de una moneda de la Página 56. Para modelizar el lanzamiento de una moneda dividimos Ω en dos subconjuntos disjuntos H y T tales que sus funciones indicatrices verifican que $\mathbf{1}_H + \mathbf{1}_T = \mathbf{1}$. Dichas funciones indicatrices generan el subespacio probabilístico $\mathcal{E}(\{\mathbf{1}_H\})$, que en este caso tan simple es de dimensión 2 (las figuras 8 y 9 muestran distintas representaciones esquemáticas de modelos de como este).

⁷⁴Donde cada $\mathcal{E}(\mathbf{Z}_n)$ es el menor subespacio probabilístico que contiene a \mathbf{X}_n y \mathbf{U}_n , para $n = 1 : N$.

⁷⁵Con datos de series temporales esta asunción será difícilmente aceptable.

Para modelizar dos lanzamientos independientes de una moneda necesitamos un espacio semi-euclídeo de probabilidad $(\mathcal{E}, \eta, \Omega)$ que contenga dos variables aleatorias independientes $\mathbb{1}_{H_1}$, $\mathbb{1}_{H_2}$ que estén distribuidas del mismo modo que $\mathbb{1}_H$. La dimensión del espacio semi-euclídeo de probabilidad \mathcal{E} más pequeño que contiene ambas copias es igual al producto de las dimensiones de las copias contenidas. Es decir, para contener las variables aleatorias del muestreo aleatorio de tamaño dos su dimensión debe ser como mínimo

$$\dim(\mathcal{L}(\{\mathbb{1}_{H_1}\})) \cdot \dim(\mathcal{L}(\{\mathbb{1}_{H_2}\})) = 2 \cdot 2.$$

Consecuentemente es imposible representar fielmente la figura en un papel. Pero como $\mathcal{L}(\{\mathbb{1}\})$ está contenido tanto en $\mathcal{L}(\{\mathbb{1}_{H_1}\})$ como $\mathcal{L}(\{\mathbb{1}_{H_2}\})$, podemos realizar una representación *esquemática* de los dos subespacios al modo de la Figura 10 en la página 58.⁷⁶

Como para un *m.a.s.* de tamaño N dispondremos de N “copia” del modelo descrito más arriba, el espacio semi-euclídeo de probabilidad completo tendrá, como mínimo, dimensión 2^N , pues debe contener N subespacios $\mathcal{L}(\{\mathbb{1}_{H_N}\})$ que se cortan “a escuadra” en la recta común de variables aleatorias constantes $\mathcal{L}(\{\mathbb{1}\})$.

Vectores y matrices estocásticas como representación de un muestreo En este contexto un muestreo de tamaño N de una variable aleatoria \mathbf{Y} es una N -tupla de la forma $(Y_1(\omega), Y_2(\omega), \dots, Y_N(\omega))$ donde las variables aleatorias Y_n son *i.i.d.*. Vamos a denotar estas n -tuplas con \mathbf{Y} (que llamaremos *vector aleatorio o estocástico* por estar formado por variables aleatorias):

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}.$$

(análogamente, llamaremos *matriz aleatoria o estocástica* a una matriz cuyas columnas son vectores aleatorios).

De esta manera, cuando digamos que el vector aleatorio \mathbf{Y} es un *m.a.s.* de \mathbf{Y} estaremos diciendo que las componentes del vector son probabilísticamente independientes y que todas se distribuyen identicamente a como lo hace \mathbf{Y} (es decir, que son *i.i.d.*); y que todas están contenidas en un mismo espacio semi-euclídeo de probabilidad $(\mathcal{E}, \eta, \Omega)$. Y cuando digamos que el vector \mathbf{y} es una muestra de \mathbf{Y} , estaremos indicando cada componente y_n es el valor que toma la correspondiente la variable aleatoria Y_n en un punto ω del conjunto de sucesos elementales Ω ; es decir que

$$\mathbf{y} = \begin{pmatrix} Y_1(\omega) \\ Y_2(\omega) \\ \vdots \\ Y_N(\omega) \end{pmatrix} \quad \text{para algún } \omega \in \Omega.$$

Así, para representar el muestreo en un modelo clásico de regresión $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$, denotaremos con \mathbf{X} a la matriz estocástica $[\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_k]$ cuya columna j -ésima $\mathbf{X}_{|j}$ es el vector aleatorio correspondiente a un *m.a.s.* del j -ésimo regresor $\mathbf{X}_{|j}$:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix}; \quad \mathbf{X}_{|j} = \mathbf{X}_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{Nj} \end{pmatrix}; \quad {}_{i|}\mathbf{X} = (X_{i1}, \quad X_{i2}, \quad \dots \quad X_{ik}),$$

y donde ${}_{i|}\mathbf{X}_{|j} = X_{ij}$ corresponde al elemento i -ésimo del *m.a.s.* del regresor j -ésimo. Denotaremos respectivamente con \mathbf{Y} y con \mathbf{U} a los muestreos de \mathbf{Y} y de \mathbf{U} . Asumiremos que todas las variables aleatorias contenidas en \mathbf{Y} , en \mathbf{X} y en \mathbf{U} pertenecen a un mismo ESP \mathcal{E} .⁷⁷ Recuerde que para que los muestreos sean *m.a.s.* los subespacios probabilísticos correspondientes al muestreo de cada “copia” del modelo ${}_n|\mathbf{Y} = {}_n|\mathbf{X} + {}_n|\mathbf{U}$ deberán ser probabilísticamente independientes entre si (es decir, estarán “a escuadra” los unos respecto a los otros).

⁷⁶Aunque el dibujo no representa fielmente a \mathcal{E} , que es de dimensión 4.

⁷⁷Piense que cada uno de esos subespacios probabilísticos de \mathcal{E} puede ser de dimensión infinita, por lo que ¡el espacio \mathcal{E} que los engloba puede ser enorme!

12.1 Momentos muestrales

12.1.1 Media muestral

Sea \mathbf{Y} un *m.a.s.* de \mathbf{Y} . Denotaremos con $m_{\mathbf{Y}}$ a la media del vector \mathbf{Y}

$$m_{\mathbf{Y}} = \frac{\sum \mathbf{Y}_n}{N}.$$

¿Cuál es su valor esperado de $m_{\mathbf{Y}}$? Dado que \mathbf{Y} es un *m.a.s.* de \mathbf{Y} , cada \mathbf{Y}_n tiene esperanza igual a la de \mathbf{Y} . Así,

$$\mathrm{E}(m_{\mathbf{Y}}) = \mathrm{E}\left(\frac{1}{N} \sum \mathbf{Y}_n\right) = \frac{1}{N} \sum \mathrm{E}(\mathbf{Y}_n) = \frac{1}{N} N \mathrm{E}(\mathbf{Y}) = \mathrm{E}(\mathbf{Y}) = \mu_{\mathbf{Y}}.$$

Por tanto, $m_{\mathbf{Y}}$ e \mathbf{Y} tienen el mismo valor esperado, y por ello diremos que *$m_{\mathbf{Y}}$ es un estimador insesgado de $\mathrm{E}(\mathbf{Y})$* . En cuanto a la varianza de $m_{\mathbf{Y}}$:

$$\begin{aligned} \mathrm{Var}(m_{\mathbf{Y}}) &= \mathrm{Var}\left(\frac{1}{N} \sum_{n=1}^N \mathbf{Y}_n\right) = \frac{1}{N^2} \mathrm{Var}\left(\sum_{n=1}^N \mathbf{Y}_n\right) \\ &= N^{-2} \sum_{n=1}^N \mathrm{Var}(\mathbf{Y}_n) \quad \text{dado que por ser independientes, sus covarianzas son nulas} \\ &= N^{-1} \mathrm{Var}(\mathbf{Y}) \quad \text{dado que tienen idéntica distribución} \end{aligned}$$

Nótese que cuando el tamaño de la muestra tiende a infinito ($N \rightarrow \infty$) la varianza del estimador $m_{\mathbf{Y}}$ tiende a cero. Por tanto, la media muestral es un *estimador consistente*.

12.1.2 La varianza muestral

El estadístico $s_{\mathbf{XY}} = \frac{\sum_{n=1}^N (\mathbf{X}_n - m_{\mathbf{X}})(\mathbf{Y}_n - m_{\mathbf{Y}})}{N}$ es la *covarianza muestral* entre \mathbf{X} e \mathbf{Y} . Es fácil comprobar que

$$s_{\mathbf{XY}} = m_{\mathbf{X} \odot \mathbf{Y}} - m_{\mathbf{X}} m_{\mathbf{Y}},$$

donde $\mathbf{X} \odot \mathbf{Y}$ es el vector de variables aleatorias que resulta de aplicar el producto componente a componente (o Hadamard), es decir, donde ${}_{n|}(\mathbf{X} \odot \mathbf{Y}) = ({}_{n|}\mathbf{X})({}_{n|}\mathbf{Y})$. Como caso particular tenemos la *varianza muestral* de \mathbf{Y} .

$$s_{\mathbf{Y}}^2 = s_{\mathbf{YY}} = \frac{\sum_{n=1}^N (\mathbf{Y}_n - m_{\mathbf{Y}})^2}{N} = m_{\mathbf{Y}^2} - (m_{\mathbf{Y}})^2$$

12.1.3 La cuasi-varianza muestral

La *cuasi-varianza* de \mathbf{Y} es

$$\mathfrak{s}_{\mathbf{Y}}^2 = \frac{\sum_{n=1}^N (\mathbf{Y}_n - m_{\mathbf{Y}})^2}{N-1} = \frac{N}{N-1} s_{\mathbf{Y}}^2$$

(nótese que se divide por $N-1$). *La cuasi-varianza es un estimador insesgado de la varianza.*

P-1
(86)

Consecuentemente la *varianza muestral* es un estimador *sesgado*. No obstante el sesgo tiene a cero cuando el tamaño muestral tiende a infinito, ya que $\frac{N}{N-1} \rightarrow 1$ cuando $N \rightarrow \infty$.

13 Propiedades estadísticas de los estimadores MCO

En adelante interpretaremos que \mathbf{X} e \mathbf{y} son muestras de tamaño N del sistema \mathbf{X} y de la variable \mathbf{Y} respectivamente.⁷⁸

Para poder deducir las propiedades de estimador MCO, debemos imponer algún modelo para el muestreo. En este caso vamos a imponer un modelo muy simple, pero también muy restrictivo. De hecho no suele ser apropiado para series temporales, aunque si se suele considerar aceptable con datos de sección cruzada (aunque no siempre). El modelo es que nuestros datos provienen de una *muestra aleatoria simple*. En particular, considere una matriz de variables aleatorias, \mathbf{X} , cuyas N filas tienen idéntica distribución que el sistema de regresores \mathbf{X} y, además, cada fila

⁷⁸Donde \mathbf{X} es la matriz de k columnas $[\mathbf{X}_{|1}; \mathbf{X}_{|2}; \dots; \mathbf{X}_{|k}]$. Cada columna $\mathbf{X}_{|j} \in \mathbb{R}^N$ es una muestra de tamaño N del regresor $\mathbf{X}_{|j} = \mathbf{x}_j$. Es decir, \mathbf{X} es una matriz cuyas componentes vienen determinadas por los valores que las componentes de \mathbf{X} toman en algún suceso elemental ω de Ω . Y exactamente lo mismo con \mathbf{y} respecto a \mathbf{Y} . Dicho de otro modo, \mathbf{X} es una realización de \mathbf{X} e \mathbf{y} una realización de \mathbf{Y} .

es probabilísticamente independiente del resto de filas:

$${}_{i|} \mathbf{X} \sim \text{iid. } \mathbf{X}$$

Y considere también un vector \mathbf{Y} de N variables que es un *m.a.s.* del regresando \mathbf{Y} .

$${}_{i|} \mathbf{Y} \sim \text{iid. } \mathbf{Y};$$

así como el vector \mathbf{U} donde ${}_{i|} \mathbf{U} \sim \text{iid. } \mathbf{U}$.

Así pues, asumiremos que tenemos N copias del Modelo Clásico de Regresión $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$ (con sus cuatro supuestos). Por tanto, para cada elemento ${}_{i|} \mathbf{Y}$ de la muestra, tendremos que (Sup. I)

$${}_{i|} \mathbf{Y} = {}_{i|} \mathbf{X} \beta + {}_{i|} \mathbf{U} \quad \text{donde} \quad [{}_{i|} \mathbf{Y}; {}_{i|} \mathbf{X}; {}_{i|} \mathbf{U};] \sim \text{iid. } [\mathbf{Y}; \mathbf{X}; \mathbf{U};];$$

donde (Sup. II)

$$\mathbb{E}({}_{i|} \mathbf{U} | \mathbf{X}) = \mathbb{E}({}_{i|} \mathbf{U} | {}_{i|} \mathbf{X}) = \mathbf{0},$$

(puesto $\mathbb{E}({}_{i|} \mathbf{X})$ es independiente de $\mathbb{E}({}_{j|} \mathbf{X})$ cuando $i \neq j$ por ser \mathbf{X} un *m.a.s.*); donde (Sup. III) $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2 \mathbf{I}$. (pues \mathbf{I} es una matriz estocástica cuadrada cuyas componentes en la diagonal son 1 y fuera de la diagonal son 0); y donde (Sup. IV) $E(\mathbf{X}^\top \mathbf{X})$ es invertible;⁷⁹ y esto último implica que $\mathbf{X}^\top \mathbf{X}$ es no singular con probabilidad 1.⁸⁰

13.1 Estimador MCO de β

En el ajuste MCO teníamos que $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ donde

- ${}_{i|}(N^{-1}(\mathbf{X}^\top \mathbf{X}))_{ij} = \mu_{\mathbf{x}_i \odot \mathbf{x}_j}$
- ${}_{i|}(N^{-1}(\mathbf{X}^\top \mathbf{y})) = \mu_{\mathbf{x}_i \odot \mathbf{y}}$

Sustituyendo $\mu_{\mathbf{x}_i \odot \mathbf{x}_j}$ por el estimador $m_{\mathbf{x}_i \odot \mathbf{x}_j}$ y sustituyendo $\mu_{\mathbf{x}_i \odot \mathbf{y}}$ por el estimador $m_{\mathbf{x}_i \odot \mathbf{y}}$, obtenemos el *estimador MCO* $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ donde

- ${}_{i|}(N^{-1}(\mathbf{X}^\top \mathbf{X}))_{ij} = m_{\mathbf{x}_i \odot \mathbf{x}_j}$
- ${}_{i|}(N^{-1}(\mathbf{X}^\top \mathbf{Y})) = m_{\mathbf{x}_i \odot \mathbf{y}}$

(Lección 6) T-1 Estimador MCO $\hat{\beta}$

Sean \mathbf{Y} (vector) y \mathbf{X} (matriz); *muestreos aleatorios simples* (*m.a.s.*) del modelo $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$ que cumple todos los supuestos. Entonces

$$[{}_{i|} \mathbf{Y}; {}_{i|} \mathbf{X}] \sim \text{iid. } [\mathbf{Y}; \mathbf{X}]; \quad \text{donde (Sup I)} \mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$$

y (Sup IV) $E(\mathbf{X}^\top \mathbf{X})$ es invertible. El *estimador MCO de β* es

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Además, el modelo muestral $\boxed{\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}}$ verifica que

(Sup II) $\mathbb{E}(\mathbf{U} | \mathbf{X}) = \mathbf{0}$

(Sup III) $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2 \mathbf{I}$ (*homocedasticidad*, NO *autocorrelación*)

Dadas las muestras \mathbf{X} (rango k) e \mathbf{y} , la *estimación MCO de β* es:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

F64

⁷⁹Como ${}_{n|} \mathbf{X} = \mathbf{X}_n \sim \text{iid. } \mathbf{X}$, el supuesto de que $E(\mathbf{X}^\top \mathbf{X})$ es invertible (del Modelo Clásico de Regresión) implica $E(\mathbf{X}^\top \mathbf{X})$ es invertible pues expresando el producto de matrices como suma de matrices, tenemos que $\mathbf{X}^\top \mathbf{X} = \sum_{n=1}^N \mathbf{X}_n^\top \mathbf{X}_n$. Consecuentemente $E(\mathbf{X}^\top \mathbf{X}) = N E(\mathbf{X}^\top \mathbf{X})$ donde $N \geq 1$.

⁸⁰(Véase Wooldridge, 2010)

Nota 2. Sea una muestra \mathbf{X} de \mathbf{Y} de rango k . Si denotamos con \mathbf{A} a la matriz

$$\underset{k \times N}{\mathbf{A}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top;$$

es evidente que cada $\hat{\beta}_j$, con $j = 1 : k$, es una transformación lineal de los N datos del vector \mathbf{y} , donde los coeficientes específicos de cada combinación son los elementos de cada una de las k filas de \mathbf{A}

$$\hat{\beta}_j = {}_{j|} \hat{\beta} = {}_{j|} \mathbf{A} \mathbf{y}.$$

Si $\underset{N \times k}{\mathbf{X}}$ es un m.a.s. de $\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k]$,⁸¹ y si denotamos con \mathbf{A} a la matriz estocástica

$$\underset{k \times N}{\mathbf{A}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top;$$

entonces cada elemento $\hat{\beta}_j$ del vector aleatorio $\hat{\beta}$ es de la forma

$$\hat{\beta}_j = {}_{j|} \hat{\beta} = {}_{j|} \mathbf{A} \mathbf{Y}.$$

No solo eso. Nótese que $\hat{\beta} = \mathbf{A} \mathbf{Y} = \mathbf{A}[\mathbf{X}\beta + \mathbf{U}] = \mathbf{I}\beta + \mathbf{A}\mathbf{U}$ es decir:

$\hat{\beta}$ es igual al verdadero vector constante de parámetros $\mathbf{I}\beta$ más el vector aleatorio $\mathbf{A}\mathbf{U}$.

13.2 Esperanza del estimador MCO $\hat{\beta}$

(Véase la Sección 16 en la página 84 para repasar las definiciones y resultados referentes a la esperanza y la esperanza condicional de varias variables.)

Nótese que $\mathbb{E}(\mathbf{U} | \mathbf{X})$ es un vector de variables aleatorias nulas (Sup II):

$$\underset{i|}{\mathbb{E}}(\mathbf{U} | \mathbf{X}) = \mathbb{E}(U_i | \mathbf{X}) = \mathbb{E}(U_i | {}_{i|} \mathbf{X}) = \mathbf{0},$$

que de manera compacta expresaremos así

$$\mathbb{E}(\mathbf{U} | \mathbf{X}) = \mathbf{0}.$$

donde $\mathbf{0}$ es un vector cuyas componentes son todas iguales a 0 . Recuerde también que si $f(\mathbf{X}) \in \mathcal{L}(\mathbf{X})$ entonces $\mathbb{E}(f(\mathbf{X})\mathbf{Y} | \mathbf{X}) = f(\mathbf{X})\mathbb{E}(\mathbf{Y} | \mathbf{X})$; y en particular $\mathbb{E}(f(\mathbf{X}) | \mathbf{X})$ es igual a $f(\mathbf{X})$.

(Lección 6) T-2 Esperanza del estimador MCO $\hat{\beta}$

En el m.a.s., $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, si $\mathbb{E}(\mathbf{X}^\top \mathbf{X})$ es invertible y denotamos $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ por $\underset{k \times N}{\mathbf{A}}$:

$$\hat{\beta} = \mathbf{A} \mathbf{Y} = \mathbf{A}(\mathbf{X}\beta + \mathbf{U}) = \mathbf{I}\beta + \mathbf{A}\mathbf{U}$$

$$\begin{aligned} \text{Así, } \mathbb{E}(\hat{\beta} | \mathbf{X}) &= \mathbb{E}(\mathbf{I}\beta + \mathbf{A}\mathbf{U} | \mathbf{X}) \\ &= \mathbf{I}\beta + \mathbf{A}\mathbb{E}(\mathbf{U} | \mathbf{X}) \\ &= \mathbf{I}\beta + \mathbf{A}\mathbf{0} = \mathbf{I}\beta \end{aligned} \quad \begin{matrix} \mathbf{I}\beta, \mathbf{A} \in \mathcal{L}(\mathbf{X}) \\ (\text{Sup II}). \end{matrix}$$

Por T^a Esperanzas iteradas: $\mathbb{E}(\hat{\beta}) = \mathbb{E}(\mathbb{E}(\hat{\beta} | \mathbf{X})) = \mathbb{E}(\mathbf{I}\beta) = \beta$.

Por tanto $\hat{\beta}$ es un estimador insesgado.

F65

⁸¹Y si $\mathbb{E}(\mathbf{X}^\top \mathbf{X})$ es invertible, lo que implica que también $\mathbb{E}(\mathbf{X}^\top \mathbf{X})$ es invertible, y por tanto $\mathbf{X}^\top \mathbf{X}$ es no singular con probabilidad 1.

13.3 Varianza del estimador MCO $\hat{\beta}$

Proposición 13.1 (Matriz de varianzas del estimador MCO). $\text{Var}(\hat{\beta} | \mathbf{X}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

Demostración. Dado el Sup IV (es decir, dado que $E(\mathbf{X}^\top \mathbf{X})$ es invertible), denotemos $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ por $\mathbf{A}_{k \times N}$. Sabiendo que $\hat{\beta} = \mathbf{I}\beta + \mathbf{AU}$, y empleando las propiedades de la esperanza y la varianza de vectores tenemos:

$$\begin{aligned}\text{Var}(\hat{\beta} | \mathbf{X}) &= \text{Var}(\hat{\beta} - \mathbf{I}\beta | \mathbf{X}) && \text{ya que } \mathbf{I}\beta \in \mathcal{L}(\mathbf{X}) \text{ por ser cte.} \\ &= \text{Var}(\mathbf{AU} | \mathbf{X}) && \text{ya que } \hat{\beta} = \mathbf{I}\beta + \mathbf{AU} \\ &= \mathbf{A} \text{Var}(\mathbf{U} | \mathbf{X}) \mathbf{A}^\top && \text{pues } \text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2 \mathbf{I} \text{ (Sup III)} \\ &= \mathbf{A} \sigma^2 \mathbf{I} \mathbf{A}^\top \\ &= \sigma^2 \mathbf{A} \mathbf{A}^\top = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.\end{aligned}$$

puesto que $\mathbf{A} \mathbf{A}^\top = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1}$. □

P-3
(87)

(Lección 6) T-3 Varianza del estimador MCO $\hat{\beta}$

Por los supuestos I, III y IV:

$$\begin{aligned}\text{Var}(\hat{\beta} | \mathbf{X}) &= \text{Var}(\hat{\beta} - \mathbf{I}\beta | \mathbf{X}) && = \text{Var}(\mathbf{AU} | \mathbf{X}) \\ &= \mathbf{A} \text{Var}(\mathbf{U} | \mathbf{X}) \mathbf{A}^\top && = \mathbf{A} \sigma^2 \mathbf{I} \mathbf{A}^\top \quad (\text{Sup III}) \\ &= \sigma^2 \mathbf{A} \mathbf{A}^\top && = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.\end{aligned}$$

donde $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.

$$\text{Var}(\hat{\beta}) = E(\text{Var}(\hat{\beta} | \mathbf{X})) + \text{Var}\left(\underbrace{E(\hat{\beta} | \mathbf{X})}_{\mathbf{I}\beta \text{ (Sup II)}}\right) = E\left(\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)$$

Por tanto: Var(\hat{\beta}) = \sigma^2 E(\mathbf{X}^\top \mathbf{X})^{-1}.

Así, $\text{Var}(\hat{\beta}_j) = \sigma^2 \left(\sum_{ij} E(\mathbf{X}^\top \mathbf{X})^{-1} \right)_{jj}$.

F66

Por la Ecuación 36 en la página 85 sabemos que $\text{Var}(\mathbf{I}\beta)$ es una matriz nula.

Ejemplo 10. Continuación de “precio de las viviendas”:

Podemos calcular la inversa de $\mathbf{X}^\top \mathbf{X}$:

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 9.1293e-01 & -4.4036e-04 \\ -4.4036e-04 & 2.3044e-07 \end{bmatrix};$$

por tanto, las desviaciones típicas estimadas de \hat{a} y \hat{b} son

$$\begin{aligned}\widehat{\text{Dt}}(\hat{a}) &= \sqrt{\sigma^2 \cdot (9.1293e-01)} = \sqrt{\frac{\sigma^2 m_{(\mathbf{x}^2)}}{N \cdot s_x^2}} \\ \widehat{\text{Dt}}(\hat{b}) &= \sqrt{\sigma^2 \cdot (2.3044e-07)} = \sqrt{\frac{\sigma^2}{N \cdot s_x^2}}.\end{aligned}$$

¡Pero nos falta el valor σ^2 (la varianza de \mathbf{U} es desconocida)!

Práctica 2. Continuación del ejemplo “Precio de las viviendas”:

Observe la matriz $(\mathbf{X}^\top \mathbf{X})^{-1}$, del ejemplo del “precio de las viviendas”.

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 9.1293e-01 & -4.4036e-04 \\ -4.4036e-04 & 2.3044e-07 \end{bmatrix};$$

¿Qué estimación es más fiable, la pendiente o la constante?

Repita la regresión para “precio de las viviendas” con las siguientes modificaciones en la muestra:

1. con todos los datos excepto los de la última vivienda
2. con todos los datos excepto los de las últimas dos viviendas
3. con todos los datos excepto los de la primera y la última vivienda

¿Confirman estos resultados su respuesta a la primera pregunta?

(Lección 6) T-4 Eficiencia del estimador MCO $\hat{\beta}$: T^a de Gauss-Markov

Gracias a los supuestos I a IV,

$\hat{\beta}$ eficiente entre estimadores lineales e insesgados
es decir, para cualquier estimador lineal^a insesgado $\tilde{\beta}$

$$\text{Var}(\tilde{\beta} | \mathbf{X}) \geq \text{Var}(\hat{\beta} | \mathbf{X}).$$

Entonces se dice ELIO (BLUE en inglés).

F68

^ade la forma $\tilde{\beta} = \mathbf{F}\mathbf{Y}$ donde \mathbf{F} es una matriz aleatoria.

Teorema 13.2 (Gauss-Markov). Sean $\hat{\beta}$ el estimador MCO de β , y $\tilde{\beta}$ otro estimador (función del muestreo \mathbf{X}) lineal e insesgado de β ; entonces bajo los supuestos I a IV, para cualquier $\mathbf{v} \in \mathbb{R}^k$ se verifica que $\text{Var}(\mathbf{v} \cdot \tilde{\beta} | \mathbf{X}) \geq \text{Var}(\mathbf{v} \cdot \hat{\beta} | \mathbf{X})$

Demostración. Como $\tilde{\beta} = \mathbf{F}\mathbf{Y}$ (con $\mathbf{F} = f(\mathbf{X})$) es un estimador insesgado, entonces $\mathbb{E}(\tilde{\beta} | \mathbf{X}) = \mathbf{F}\mathbb{E}(\mathbf{Y} | \mathbf{X}) = \mathbf{F}\mathbf{X}\beta = \mathbf{I}\beta$. Por tanto la insesgadez implica necesariamente que $\mathbf{F}\mathbf{X} = \mathbf{I}$. Sea $\mathbf{G} = \mathbf{F} - \mathbf{A}$, donde $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$; entonces $\mathbf{G}\mathbf{X} = \mathbf{0}$ (pues $\mathbf{A}\mathbf{X} = \mathbf{I}$). Puesto que $\text{Var}(\mathbf{Y} | \mathbf{X}) = \text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2 \mathbf{I}$, se deduce que:

$$\text{Var}(\tilde{\beta} | \mathbf{X}) = \mathbf{F} \text{Var}(\mathbf{Y} | \mathbf{X}) \mathbf{F}^\top = \sigma^2 (\mathbf{A} + \mathbf{G}) \mathbf{I} (\mathbf{A}^\top + \mathbf{G}^\top) = \sigma^2 (\mathbf{A}\mathbf{A}^\top + \mathbf{A}\mathbf{G}^\top + \mathbf{G}\mathbf{A}^\top + \mathbf{G}\mathbf{G}^\top) = \underbrace{\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}}_{\text{Var}(\tilde{\beta} | \mathbf{X})} + \sigma^2 \mathbf{G}\mathbf{G}^\top,$$

pues $\mathbf{A}\mathbf{A}^\top = (\mathbf{X}^\top \mathbf{X})^{-1}$, $\mathbf{G}\mathbf{A}^\top = \underbrace{\mathbf{G}\mathbf{X}}_{\mathbf{0}} (\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{0}$; y donde $\mathbf{G}\mathbf{G}^\top$ es semi-definida positiva. Así, para todo $\mathbf{v} \in \mathbb{R}^k$

$$\text{Var}(\mathbf{v} \cdot \tilde{\beta} | \mathbf{X}) = \mathbf{v} \text{Var}(\tilde{\beta} | \mathbf{X}) \mathbf{v} = \mathbf{v} (\text{Var}(\hat{\beta} | \mathbf{X}) + \sigma^2 \mathbf{G}\mathbf{G}^\top) \mathbf{v} = \text{Var}(\mathbf{v} \cdot \hat{\beta} | \mathbf{X}) + \sigma^2 \mathbf{v} \mathbf{G}\mathbf{G}^\top \mathbf{v}$$

que implica:⁸² $\text{Var}(\mathbf{v} \cdot \tilde{\beta} | \mathbf{X}) \geq \text{Var}(\mathbf{v} \cdot \hat{\beta} | \mathbf{X})$. □

En particular el T^a arriba mencionado implica que

P-4
(87)

$$\text{Var}(\tilde{\beta}_{|j} | \mathbf{X}) \geq \text{Var}(\hat{\beta}_{|j} | \mathbf{X}); \quad \text{para } j = 1 : k.$$

es decir, la relación es cierta para cada uno de los estimadores de cada uno de los parámetros individuales.

⁸²Decimos que $\mathbf{A} \geq \mathbf{B}$ cuando $\mathbf{A} - \mathbf{B}$ es definida positiva. Así que $\text{Var}(\tilde{\beta} | \mathbf{X}) \geq \text{Var}(\hat{\beta} | \mathbf{X})$ significa que la matriz $[\text{Var}(\tilde{\beta} | \mathbf{X}) - \text{Var}(\hat{\beta} | \mathbf{X})]$ es definida positiva.

13.3.1 Consistencia del estimador MCO

Un estimador es consistente si es insesgado y su varianza tiende a cero cuando el tamaño de la muestra tiende a infinito: $N \rightarrow \infty$.

Aunque no vamos a estudiar propiedades asintóticas, es importante que sepa que, bajo los supuestos muestrales que hemos visto, el estimador MCO, $\hat{\beta}$, es *consistente*. Puede experimentarlo simulando con Gretl un mismo modelo, pero cada vez con muestras mayores. Una demostración analítica requiere de teoría asintótica que no vamos a ver. Pero debe saber que incluso con supuestos menos restrictivos, el estimador MCO sigue siendo consistente:

- si en lugar de $E(\mathbf{U} | \mathbf{X}) = \mathbf{0}$, solo pedimos $E(\mathbf{X}\mathbf{U}) = \mathbf{0}$
- y si en lugar de $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2 \mathbf{I}$, solo pedimos $E(\mathbf{U}_n^2(\mathbf{X}^\top \mathbf{X})) = \sigma^2 E(\mathbf{X}^\top \mathbf{X})$, donde $\sigma^2 = E(\mathbf{U}_n^2)$ (es decir, si pedimos que \mathbf{U}^2 esté incorrelado con los regresores \mathbf{X}_j y con sus cuadrados \mathbf{X}_j^2)

el estimador $\hat{\beta}$ sigue siendo consistente.

Fíjese que cuando se verifica el Supuesto 2, es decir, cuando $E(\mathbf{U}^2 | \mathbf{X}) = \sigma^2 \mathbf{I}$ (y si $\mathbf{U}^2 \in \mathcal{E}$ y las variables aleatorias de $\mathbf{X}^\top \mathbf{X}$ también pertenecen a \mathcal{E}) tenemos que $E(\mathbf{U}^2(\mathbf{X}^\top \mathbf{X})) = \sigma^2 E(\mathbf{X}^\top \mathbf{X})$ pues

$$E(\mathbf{U}^2(\mathbf{X}^\top \mathbf{X})) = E(E(\mathbf{U}^2(\mathbf{X}^\top \mathbf{X}) | \mathbf{X})) = E(E(\mathbf{U}^2 | \mathbf{X})(\mathbf{X}^\top \mathbf{X})) = E(\sigma^2 \mathbf{I}(\mathbf{X}^\top \mathbf{X})) = \sigma^2 E(\mathbf{X}^\top \mathbf{X}).$$

Es decir, Supuesto 2 $\implies E(\mathbf{U}^2(\mathbf{X}^\top \mathbf{X})) = \sigma^2 E(\mathbf{X}^\top \mathbf{X})$.

(Lección 6)

T-5

Consistencia del estimador MCO $\hat{\beta}$

Además, $\hat{\beta}$ es **consistente**, es decir,

- es *insesgado*
- la *varianza tiende a cero* cuando la muestra crece

$$\lim_{N \rightarrow \infty} \text{Var}(\hat{\beta}_j | \mathbf{X}) = \mathbf{0}$$

F69

13.4 Caso particular: la constante como único regresor

Sabemos que $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, entonces si $\mathbf{X} = [\mathbf{1};]$,

- $N^{-1}(\mathbf{1}^\top \mathbf{1}) = [m_{\mathbf{1} \odot \mathbf{1}};] = [\mathbf{1};]$ y
- $N^{-1}(\mathbf{1}^\top \mathbf{Y}) = (m_{\mathbf{1} \odot \mathbf{Y}},) = (m_{\mathbf{Y}},)$.

Por tanto, para un *m.a.s.* del modelo $\mathbf{Y} = \beta_1 \mathbf{1} + \mathbf{U}$, donde $\beta = (\beta_1,) \in \mathbb{R}^1$

$$\hat{\beta} = [\mathbf{1};]^{-1} (m_{\mathbf{Y}},) = (m_{\mathbf{Y}},).$$

es decir, $\hat{\beta}_1 = m_{\mathbf{Y}}$. Por tanto $E(\hat{\beta}_1) = E(\mathbf{Y})$. Por otra parte $\text{Var}(\hat{\beta}) = \sigma^2 E(\mathbf{1}^\top \mathbf{1})^{-1} = \sigma^2$.

13.5 Caso particular: modelo lineal simple

Dado que $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, entonces si $\mathbf{X} = [\mathbf{1}; \mathbf{X};]$,

$$\bullet N^{-1}(\mathbf{X}^\top \mathbf{X}) = \begin{bmatrix} \mathbf{1} & m_{\mathbf{X}} \\ m_{\mathbf{X}} & m_{\mathbf{X}^2} \end{bmatrix}$$

$$\bullet N^{-1}(\mathbf{X}^\top \mathbf{Y}) = \begin{pmatrix} m_{\mathbf{X}} \\ m_{\mathbf{X} \odot \mathbf{Y}} \end{pmatrix}.$$

Por tanto, para un *m.a.s.* del modelo $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \mathbf{U}$, el estimador MCO es $\hat{\boldsymbol{\beta}} = \begin{pmatrix} m_{\mathbf{Y}} - (\frac{s_{\mathbf{X}} \mathbf{Y}}{s_{\mathbf{X}}^2}) m_{\mathbf{X}} \\ \frac{s_{\mathbf{X}} \mathbf{Y}}{s_{\mathbf{X}}^2} \end{pmatrix}$. Dicho estimador es insesgado y su varianza es $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{E}(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{E}\left(\frac{\sigma^2}{s_{\mathbf{X}}^2} \begin{bmatrix} m_{\mathbf{X}^2} & -m_{\mathbf{X}} \\ -m_{\mathbf{X}} & 1 \end{bmatrix}\right)$

13.6 Momentos de los valores ajustados $\hat{\mathbf{y}}$ y de los errores $\hat{\mathbf{e}}$

Revise la sección “Expresión matricial de la proyección ortogonal” del [Curso de Álgebra Lineal](#).

Denotemos $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ por \mathbf{P} , y nótese que

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{XA}; \quad \mathbf{M} \in \mathcal{L}(\mathbf{X}),$$

donde $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$; y que $\mathbf{P}^\top = \mathbf{P}$ y $\mathbf{PP} = \mathbf{P}$.

Veamos cómo son la esperanza y varianza de los valores ajustados por MCO:

P-5
(87)

(Lección 6) T-6 Primeros momentos de $\hat{\mathbf{y}}$ (valores ajustados por MCO)

$$\hat{\mathbf{y}} = \underbrace{\mathbf{X}\hat{\boldsymbol{\beta}}}_{\text{(Requiere Sup. IV)}} = \mathbf{X}(\mathbf{I}\boldsymbol{\beta} + \mathbf{AU}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{PU};$$

$$\begin{aligned} \mathbb{E}(\hat{\mathbf{y}} | \mathbf{X}) &= \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{PU} | \mathbf{X}) = \mathbb{E}(\mathbf{X}\boldsymbol{\beta} | \mathbf{X}) + \mathbf{P} \cdot \mathbb{E}(\mathbf{U} | \mathbf{X}) \\ &= \mathbf{X}\boldsymbol{\beta} \quad (\text{por Sup. II}) \end{aligned}$$

$$\text{Así, } \mathbb{E}(\hat{\mathbf{y}}) = \mathbb{E}(\mathbf{X}\boldsymbol{\beta}) \Rightarrow \mathbb{E}(\widehat{\mathbf{Y}_n}) = \mathbb{E}(\mathbf{X}\boldsymbol{\beta})$$

$$\begin{aligned} \text{Var}(\hat{\mathbf{y}} | \mathbf{X}) &= \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{PU} | \mathbf{X}) = \mathbf{P} \text{Var}(\mathbf{U} | \mathbf{X}) \mathbf{P}^\top \\ &= \sigma^2 \mathbf{P} \mathbf{P}^\top = \sigma^2 \mathbf{P} \quad (\text{por Sup. III}) \end{aligned} \tag{32}$$

(matriz “llena”)

F70

Fíjese que, como en este caso la esperanza condicional no es constante, la varianza corresponde a la suma de dos matrices definidas positivas:

$$\begin{aligned} \text{Var}(\hat{\mathbf{y}}) &= \mathbb{E}(\text{Var}(\hat{\mathbf{y}} | \mathbf{X})) + \text{Var}(\mathbb{E}(\hat{\mathbf{y}} | \mathbf{X})) \\ &= \sigma^2 \mathbb{E}(\mathbf{P}) + \text{Var}(\mathbf{X}\boldsymbol{\beta}) \\ &= \sigma^2 \mathbb{E}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) + \text{Var}(\mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

Nótese que la matriz de varianzas y covarianzas generalmente es una matriz “llena” (al contrario que la matriz identidad que fuera de la diagonal está completamente compuesta de ceros) por tanto los valores ajustados son autocorrelados (es decir, en general $\text{Cov}(\hat{Y}_i, \hat{Y}_j) \neq 0$ si $i \neq j$) y heterocedásticos (es decir, en general $\text{Var}(\hat{Y}_i) \neq \text{Var}(\hat{Y}_j) \neq 0$ si $i \neq j$).

Denotemos $\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ por \mathbf{M} . Y nótese que

$$\mathbf{M} \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{XA}; \quad \mathbf{M} \in \mathcal{L}(\mathbf{X}),$$

y que $\mathbf{M}^\top = \mathbf{M}$ y $\mathbf{MM} = \mathbf{M}$.

Nota 3. Respecto al ajuste MCO visto en las primeras lecciones, a partir de ahora tan sólo cambiaremos su interpretación: si suponemos que los datos son realizaciones de las variables aleatorias de un modelo clásico de regresión

lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$ (y si se cumplen los supuestos que ahora veremos) interpretaremos la recta de regresión $\hat{\mathbf{y}}$ como una estimación de la esperanza condicional $\mathbb{E}(\mathbf{Y} | \mathbf{X})$.

Veamos cómo son la esperanza y varianza de los errores MCO:

P-6
(87)

<p style="margin: 0;">(Lección 6)</p>	<p style="margin: 0;">T-7 Primeros momentos de los errores MCO</p>
$\hat{\mathbf{e}} = \underbrace{\mathbf{Y} - \hat{\mathbf{y}}}_{\text{(Requiere Sup. IV)}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{U}) = \mathbf{M}\mathbf{U}$ $\mathbb{E}(\hat{\mathbf{e}} \mathbf{X}) = \mathbb{E}(\mathbf{M}\mathbf{U} \mathbf{X}) = \mathbf{M} \cdot \mathbb{E}(\mathbf{U} \mathbf{X}) = \mathbf{0} \quad (\text{por Sup. II})$ <p style="text-align: center;">Así, $\mathbb{E}(\hat{\mathbf{e}}) = \mathbf{0}$</p> $\text{Var}(\hat{\mathbf{e}} \mathbf{X}) = \mathbf{M} \text{Var}(\mathbf{U} \mathbf{X}) \mathbf{M}^\top = \sigma^2 \mathbf{M} \mathbf{M}^\top = \sigma^2 \mathbf{M} \quad (\text{por Sup. III})$	
<p style="margin: 0;">(matriz “llena”)</p>	

Nótese que de nuevo encontramos una la matriz “llena”, así que también los errores estimados son en general autocorrelados y heterocedásticos. Además

$$\text{Var}(\hat{\mathbf{e}}) = \mathbb{E}(\text{Var}(\hat{\mathbf{e}} | \mathbf{X})) + \text{Var}(\mathbb{E}(\hat{\mathbf{e}} | \mathbf{X})) = \mathbb{E}(\sigma^2 \mathbf{M}) + \text{Var}(\mathbf{0}) = \sigma^2 \mathbb{E}(\mathbf{M}).$$

14 Distribución de los estimadores MCO bajo la hipótesis de Normalidad

(Wooldridge, 2006, Secciones 4.1 y 4.2 y Apéndice E3)

Nota 4. *Distribución conjunta normal implica*

1. *La distribución queda completamente determinada por el par: vector de esperanzas – matriz de varianzas y covarianzas.*
2. *Correlación cero implica independencia*
3. *Cualquier transformación lineal de las variables también tiene distribución normal*

14.1 Quinto supuesto del Modelo Clásico de Regresión Lineal

(Lección 6)

T-8 Supuesto 5: Distribución Normal de las perturbaciones

La inferencia es muy sencilla bajo el siguiente supuesto sobre la distribución conjunta de \mathbf{U} :

$$\mathbf{U} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \Rightarrow \mathbf{Y} \sim N(E(\mathbf{X}\boldsymbol{\beta}), \sigma^2 \mathbf{I})$$

donde \mathbf{I} es la matriz identidad de orden $N \times N$. Puesto que

$$\hat{\boldsymbol{\beta}} = \mathbf{I}\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{U} = \mathbf{I}\boldsymbol{\beta} + \mathbf{A}\mathbf{U}$$

entonces $\hat{\boldsymbol{\beta}}$ tiene distribución normal multivariante.

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 E(\mathbf{X}^\top \mathbf{X})^{-1})$$

F72

(Lección 6)

T-9 Distribución del estimador MCO $\hat{\boldsymbol{\beta}}$

Así pues,

$$\hat{\beta}_j \sim N(\beta_j, \text{Var}(\hat{\beta}_j))$$

donde $\text{Var}(\hat{\beta}_j) = E(\text{Var}(\hat{\beta}_j | \mathbf{X})) = \sigma^2 E\left(\left.\mathbf{I}_{jj}(\mathbf{X}^\top \mathbf{X})^{-1}\right|_{jj}\right)$
(el j -ésimo elemento de la diagonal) y

$$\frac{\hat{\beta}_j - \beta_j}{\text{Dt}(\hat{\beta}_j | \mathbf{X})} \sim N(0, 1)$$

F73

14.2 Estimación de la varianza residual y la matriz de covarianzas

Nota 5. Llamamos “traza” a la suma de los elementos de la diagonal de una matriz. El operador traza es un operador lineal con la siguiente propiedad: Sean \mathbf{A} y \mathbf{B} dos matrices cuadradas, entonces

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$$

Proposición 14.1. $\text{tr}(\mathbf{M}) = N - k$.

P-8
(87)

Proposición 14.2. $E(\hat{\mathbf{e}} \cdot \hat{\mathbf{e}} | \mathbf{X}) = \sigma^2(N - k)\mathbf{1}$.

P-9
(87)

Por tanto, la cuasi-varianza muestral de los errores, $\hat{s}_{\mathbf{e}}^2 \equiv (\hat{\mathbf{e}} \cdot \hat{\mathbf{e}})/(N - k)$, es un estimador insesgado de σ^2 . Consecuentemente emplearemos como estimador de la matriz de varianzas y covarianzas la ecuación (34) de más abajo.

El parámetro σ^2 es desconocido [F55]

Pero la cuasivarianza de $\hat{\mathbf{e}}$

$$\widehat{s}_{\hat{\mathbf{e}}}^2 \equiv (\hat{\mathbf{e}} \cdot \hat{\mathbf{e}})/(N - k)$$

es un estimador *insesgado* de σ^2 puesto que

$$E(\widehat{s}_{\hat{\mathbf{e}}}^2) = E\left(\frac{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}{N - k}\right) = \frac{\sigma^2(N - k)}{N - k} = \sigma^2$$

Así, el estimador insesgado de la matriz de varianzas condicionada de $\hat{\beta}$ es

$$\widehat{\text{Var}}(\hat{\beta} | \mathbf{X}) = \widehat{s}_{\hat{\mathbf{e}}}^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1}. \quad (34)$$

F74

14.3 Más sobre eficiencia de los estimadores

1. La matriz de varianzas y covarianzas de $\hat{\beta}$ alcanza la cota mínima de Cramér-Rao. Es decir, el estimador MCO del Modelo Lineal General es el estimador insesgado de mínima varianza (resultado más fuerte que T^a de Gauss-Markov, pues incluye a los estimadores no lineales).

Un estimador insesgado y de mínima varianza se dice que es un *estimador eficiente*.

2. La varianza del estimador $\widehat{s}_{\hat{\mathbf{e}}}^2$ es $\frac{2\sigma^4}{N}$, y por tanto no alcanza la cota mínima de Cramér-Rao. No obstante, no existe ningún estimador *insesgado* de σ^2 con varianza menor a $\frac{2\sigma^4}{N}$.

15 Más sobre medidas de ajuste

Para complementar la Sección 4.2 en la página 31, citamos aquí dos criterios para la selección de modelos basados en la función de verosimilitud, y por tanto basados en el quinto supuesto.

Los criterios de información de Akaike y de Schwartz permiten seleccionar entre modelos alternativos. (Están calculados bajo el supuesto de normalidad).

Aquí es preferido el modelo que arroja un resultado más bajo
(¡justo al revés que con los coeficientes de determinación!)

Akaike (AIC) Premia la bondad de ajuste, pero penaliza la complejidad del modelo (aunque tiende a sobre-parametrizar)

$$AIC = -2\ell(\hat{\theta}) + k$$

Schwartz (BIC) Basado en el criterio de Akaike, la penalización por el número de parámetros es mayor que en el AIC para evitar una posible sobre-parametrización.

$$SBC = -2\ell(\hat{\theta}) + k \log N$$

Hannan-Quinn (HQC) Basado en el criterio de Akaike, la penalización por el número de parámetros es mayor que en el AIC para evitar una posible sobre-parametrización.

$$HQC = -2\ell(\hat{\theta}) + 2k \log \log N$$

Véase los resultados de estimación para el precio de las viviendas (página 34). F75

16 Apéndice de definiciones y resultados

Un vector \mathbf{c} por una variable aleatoria \mathbf{X} es un vector cuyas componentes son múltiplos de la variable aleatoria:

$$\mathbf{X}\mathbf{c} = \mathbf{X}(c_1, \dots, c_k) = (\mathbf{X}c_1, \dots, \mathbf{X}c_k) = (c_1\mathbf{X}, \dots, c_k\mathbf{X}) = (c_1, \dots, c_k)\mathbf{X} = \mathbf{c}\mathbf{X}$$

o expresado con operadores selectores:

$$(\mathbf{X}\mathbf{c})_{|i} = \mathbf{X}(\mathbf{c}_{|i}) = ({}_{|i}\mathbf{c})\mathbf{X} = {}_{|i}(\mathbf{c}\mathbf{X})$$

Así, $\mathbf{c}\mathbf{1}$ es un vector de variables aleatorias constantes iguales a las componentes del vector \mathbf{c} :

$$\mathbf{c}\mathbf{1} = (c_1(1), \dots, c_k(1),).$$

Del mismo modo, $\mathbf{A}\mathbf{X}$ es una matriz cuyas componentes son múltiplos de la variable aleatoria: $(\mathbf{X}\mathbf{A})_{|j} = \mathbf{X}(\mathbf{A}_{|j})$.

En lo que sigue, nótese que tanto un sistema de variables aleatorias \mathbf{Y} como un vector aleatorio \mathbf{Y} son listas ordenadas de variables aleatorias. Pero téngase en cuenta que hemos seguido el convenio de que los sistemas de variables aleatorias se representan en horizontal y se pueden transponer; aunque esto no es así en el caso de los vectores.

$$\mathbf{Y} = [Y_1; \dots; Y_n]; \quad \mathbf{Y}^\top = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = (Y_1, \dots, Y_n); \quad \text{de manera que } [\mathbf{Y};] = \mathbf{Y}^\top,$$

es decir, que la matriz estocástica cuya única columna es \mathbf{Y} es igual a \mathbf{Y}^\top .

Esperanza condicional de un sistema (vector o matriz) de variables aleatorias

La esperanza de un sistema (vector o matriz) de variables aleatorias condicionada a un sistema (vector o matriz) de variables aleatorias es el sistema (vector o matriz) cuyas componentes son la proyección ortogonal sobre el subespacio probabilístico generado por las variables aleatorias que componen el sistema (vector o matriz) que condiciona. Por ejemplo:

Definición 28. *Esperanza de un sistema de variables aleatorias \mathbf{Y} condicionada a otro sistema \mathbf{X} es el sistema de variables aleatorias cuyas componentes son las proyecciones de las variables aleatorias Y_j sobre un espacio de probabilidad $\mathcal{L}(\mathbf{X})$ generado por las variables aleatorias de \mathbf{X} :*

$$\mathbb{E}(\mathbf{Y} | \mathbf{X}) = [\mathbb{E}(Y_1 | \mathbf{X}); \dots; \mathbb{E}(Y_k | \mathbf{X});]$$

o expresado componente a componente: $\mathbb{E}(\mathbf{Y} | \mathbf{X})_{|j} = \mathbb{E}(Y_j | \mathbf{X})$.

Nótese cómo aplica el Teorema de las Esperanzas Iteradas

$$\mathbb{E}(\mathbb{E}(\mathbf{Y} | \mathbf{X})) = [\mathbb{E}(\mathbb{E}(Y_1 | \mathbf{X})); \dots; \mathbb{E}(\mathbb{E}(Y_k | \mathbf{X}));] = [\mathbb{E}(Y_1); \dots; \mathbb{E}(Y_k);] = \mathbb{E}(\mathbf{Y}).$$

Un caso particular es la *esperanza de \mathbf{Y} condicionada a $\mathbf{1}$* , que contiene las proyecciones de las variables aleatorias X_j sobre el espacio de probabilidad de las variables aleatorias constantes $\mathcal{L}(1)$:

$$\begin{aligned} \mathbb{E}(\mathbf{X} | \mathbf{1}) &= [\mathbb{E}(X_1 | 1); \dots; \mathbb{E}(X_k | 1);] \\ &= [\mathbb{E}(X_1) 1; \dots; \mathbb{E}(X_k) 1;] = \mathbb{E}(\mathbf{X}) 1. \end{aligned}$$

o expresado componente a componente: $\mathbb{E}(\mathbf{X} | \mathbf{1})_{|j} = \mathbb{E}(X_j | 1) = \mathbb{E}(X_j) 1$.

Varianza de un sistema (o vector) de variables aleatorias

Definición 29 (Matriz de varianzas y covarianzas de un sistema de variables aleatorias). *Definimos la matriz de varianzas y covarianzas de un sistema de variables aleatorias $\mathbf{X} = [X_1; \dots; X_k]$ como*

$$\text{Var}(\mathbf{X}) = \mathbb{E}([\mathbf{X} - \mathbb{E}(\mathbf{X}) 1]^\top [\mathbf{X} - \mathbb{E}(\mathbf{X}) 1]);$$

es decir, es aquella matriz cuya componente de la fila i -ésima y la columna j -ésima es

$${}_{ij}\text{Var}(\mathbf{X}) = \mathbb{E} \left((\mathbf{X}_i - \mathbb{E}(\mathbf{X}_i) \mathbf{1}) \cdot (\mathbf{X}_j - \mathbb{E}(\mathbf{X}_j) \mathbf{1}) \right) = \text{Cov}(\mathbf{X}_i, \mathbf{X}_j). \quad (35)$$

Consecuentemente, otra expresión alternativa es

$${}_{ij}\text{Var}(\mathbf{X}) = \mathbb{E}(\mathbf{X}_i \mathbf{X}_j) - \mathbb{E}(\mathbf{X}_i) \mathbb{E}(\mathbf{X}_j);$$

es decir

$$\text{Var}(\mathbf{X}) = \mathbb{E}(\mathbf{X}^\top \mathbf{X}) - \mathbb{E}(\mathbf{X}) \mathbb{E}(\mathbf{X}),$$

y por tanto

$$\begin{aligned} \text{Var}(\mathbf{X}) &= \begin{bmatrix} \mathbb{E}(\mathbf{X}_1^2) & \mathbb{E}(\mathbf{X}_1 \mathbf{X}_2) & \cdots & \mathbb{E}(\mathbf{X}_1 \mathbf{X}_k) \\ & \mathbb{E}(\mathbf{X}_2^2) & \cdots & \mathbb{E}(\mathbf{X}_2 \mathbf{X}_k) \\ & & \ddots & \vdots \\ & & & \mathbb{E}(\mathbf{X}_k^2) \end{bmatrix} - \begin{bmatrix} [\mathbb{E}(\mathbf{X}_1)]^2 & \mathbb{E}(\mathbf{X}_1) \mathbb{E}(\mathbf{X}_2) & \cdots & \mathbb{E}(\mathbf{X}_1) \mathbb{E}(\mathbf{X}_k) \\ & [\mathbb{E}(\mathbf{X}_2)]^2 & \cdots & \mathbb{E}(\mathbf{X}_2) \mathbb{E}(\mathbf{X}_k) \\ & & \ddots & \vdots \\ & & & [\mathbb{E}(\mathbf{X}_k)]^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{\mathbf{X}_1}^2 & \sigma_{\mathbf{X}_1 \mathbf{X}_2} & \cdots & \sigma_{\mathbf{X}_1 \mathbf{X}_k} \\ & \sigma_{\mathbf{X}_2}^2 & \cdots & \sigma_{\mathbf{X}_2 \mathbf{X}_k} \\ & & \ddots & \vdots \\ & & & \sigma_{\mathbf{X}_k}^2 \end{bmatrix}, \quad \text{donde } \sigma_{\mathbf{X}_i}^2 \text{ es } \text{Var}(\mathbf{X}_i), \text{ y } \sigma_{\mathbf{X}_i \mathbf{X}_j} \text{ es } \text{Cov}(\mathbf{X}_i, \mathbf{X}_j). \end{aligned}$$

En el caso de un vector aleatorio la definición es semejante

$${}_{ij}\text{Var}(\mathbf{Y}) = \mathbb{E} \left((\mathbf{Y}_i - \mathbb{E}(\mathbf{Y}_i) \mathbf{1}) \cdot (\mathbf{Y}_j - \mathbb{E}(\mathbf{Y}_j) \mathbf{1}) \right) = \text{Cov}(\mathbf{Y}_i, \mathbf{Y}_j).$$

Pero para escribir las expresiones de forma matricial, debemos hacer lo siguiente

$$\text{Var}(\mathbf{Y}) = \mathbb{E} \left([\mathbf{Y} - \mathbb{E}(\mathbf{Y}) \mathbf{1}] [\mathbf{Y} - \mathbb{E}(\mathbf{Y}) \mathbf{1}]^\top \right);$$

o lo que es lo mismo

$$\text{Var}(\mathbf{Y}) = \mathbb{E}([\mathbf{Y}] [\mathbf{Y}]^\top) - \mathbb{E}([\mathbf{Y}]) \mathbb{E}([\mathbf{Y}]^\top).$$

Denotaremos con \mathbf{I} a una matriz cuadrada cuyas componentes en la diagonal son 1 y cuyas componentes fuera de la diagonal son 0 .

Puesto que

$$(\mathbf{Y} + \mathbf{I}\mathbf{v}) - \mathbf{I}\mathbb{E}(\mathbf{Y} + \mathbf{I}\mathbf{v}) = (\mathbf{Y} + \mathbf{I}\mathbf{v}) - \mathbf{I}\mathbb{E}(\mathbf{Y}) - \mathbf{I}\mathbb{E}(\mathbf{I}\mathbf{v}) = \mathbf{Y} + \mathbf{I}\mathbf{v} - \mathbf{I}\mathbb{E}(\mathbf{Y}) - \mathbf{I}\mathbf{v} = \mathbf{Y} - \mathbf{I}\mathbb{E}(\mathbf{Y}).$$

la varianza no cambia al sumar un vector constante:

$$\text{Var}(\mathbf{Y} + \mathbf{I}\mathbf{v}) = \text{Var}(\mathbf{Y}).$$

Otro caso especial es del sistema resultante de multiplicar un vector $\mathbf{b} \in \mathbb{R}^k$ por $\mathbf{I}_{k \times k}$, puesto que $\mathbf{I}\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}$, tenemos que $\mathbb{E}(\mathbf{I}\mathbf{b}) = \mathbb{E}(\mathbf{I})\mathbf{b} = \mathbf{I}\mathbf{b} = \mathbf{b}$; así

$$\mathbb{E}(\mathbf{I}\mathbf{b} - \mathbf{I}\mathbb{E}(\mathbf{I}\mathbf{b})) = \mathbb{E}(\mathbf{I}\mathbf{b} - \mathbf{I}\mathbf{b}) = 0;$$

y

$$\text{Var}(\mathbf{I}\mathbf{b}) = \mathbb{E}([\mathbf{I}\mathbf{b}]([\mathbf{I}\mathbf{b}]^\top)) - \mathbb{E}([\mathbf{I}\mathbf{b}])\mathbb{E}([\mathbf{I}\mathbf{b}]^\top) = \mathbf{0}. \quad (36)$$

El producto de una matriz \mathbf{Q} de m filas y N columnas por un vector de N variables aleatorias \mathbf{Y} , es otro vector de m

variables, de manera que cada variable i -ésima es el producto de la i -ésima fila ${}_i|\mathbf{Q}$ por \mathbf{Y} .

$$\mathbf{Q}\mathbf{Y} = \begin{pmatrix} {}_1|\mathbf{Q}\mathbf{Y} \\ \vdots \\ {}_m|\mathbf{Q}\mathbf{Y} \end{pmatrix} = ({}_{1|}\mathbf{Q}\mathbf{Y}, \dots, {}_{m|}\mathbf{Q}\mathbf{Y}) = (\mathbf{Y}(\mathbf{Q}^T)_{|1}, \dots, \mathbf{Y}(\mathbf{Q}^T)_{|m},) = \mathbf{Y}\mathbf{Q}^T;$$

ya que $\mathbf{Y}(\mathbf{Q}^T)_{|i} = {}_{i|}\mathbf{Q}\mathbf{Y} = q_{i1}\mathbf{Y}_1 + q_{i2}\mathbf{Y}_2 + \dots + q_{iN}\mathbf{Y}_N$. Así pues,

$$\mathbb{E}(\mathbf{Q}\mathbf{Y}) = \mathbf{Q}\mathbb{E}(\mathbf{Y}) = \mathbb{E}(\mathbf{Y})\mathbf{Q}^T = \mathbb{E}(\mathbf{Y}\mathbf{Q}^T).$$

La demostración del siguiente resultado se deja como ejercicio.

Nota 6. Sea $\mathbf{Q}_{m \times N}$, entonces, $\text{Var}(\mathbf{Q}\mathbf{Y}) = \mathbf{Q}\text{Var}(\mathbf{Y})\mathbf{Q}^T$. (P-2)
(87)

De manera similar se demuestran los siguientes resultados:

Nota 7. Sean $\mathbf{Q}_{n \times N}$ y $\mathbf{R}_{m \times N}$ matrices, y \mathbf{v} y \mathbf{w} vectores de \mathbb{R}^n y \mathbb{R}^m respectivamente. Entonces

$$\mathbb{E}(\mathbf{Q}\mathbf{U} + \mathbf{I}\mathbf{v}) = \mathbb{E}(\mathbf{Q}\mathbf{U}) + \mathbb{E}(\mathbf{I}\mathbf{v}) = \mathbf{Q}\mathbb{E}(\mathbf{U}) + \mathbf{v},$$

y

$$\text{Var}(\mathbf{Q}\mathbf{U} + \mathbf{I}\mathbf{v}) = \text{Var}(\mathbf{Q}\mathbf{U}) = \mathbf{Q}\text{Var}(\mathbf{U})\mathbf{Q}^T,$$

además

$$\text{Cov}(\mathbf{Q}\mathbf{U} + \mathbf{I}\mathbf{v}, \mathbf{R}\mathbf{U} + \mathbf{I}\mathbf{w}) = \text{Cov}(\mathbf{Q}\mathbf{U}, \mathbf{R}\mathbf{U}) = \mathbf{Q}\text{Cov}(\mathbf{U}, \mathbf{U})\mathbf{R}^T = \mathbf{Q}\text{Var}(\mathbf{U})\mathbf{R}^T$$

Definición 30. Llamaremos *varianza condicional* $\text{Var}(\mathbf{U} | \mathbf{X})$ de un vector de variables aleatorias \mathbf{U} a la matriz de variables aleatorias \mathbf{X}

$$\text{Var}(\mathbf{U} | \mathbf{X}) = \mathbb{E}([\mathbf{U} - \mathbb{E}(\mathbf{U} | \mathbf{X})][\mathbf{U} - \mathbb{E}(\mathbf{U} | \mathbf{X})]^T | \mathbf{X});$$

cuya componente de la fila i -ésima y columna j -ésima es

$${}_{ij}\text{Var}(\mathbf{U} | \mathbf{X})_{|j} = \text{Cov}(U_i, U_j | \mathbf{X}) = \mathbb{E}((U_i - \mathbb{E}(U_i | \mathbf{X}))(U_j - \mathbb{E}(U_j | \mathbf{X})) | \mathbf{X})$$

Nótese que hay una segunda expresión equivalente

$$\text{Var}(\mathbf{U} | \mathbf{X}) = \mathbb{E}([\mathbf{U}][\mathbf{U}]^T | \mathbf{X}) - ([\mathbb{E}(\mathbf{U} | \mathbf{X})][\mathbb{E}(\mathbf{U} | \mathbf{X})]^T).$$

Nota 8. Sean $\mathbf{Q}_{n \times N} = f(\mathbf{X})$ y $\mathbf{R}_{m \times N} = g(\mathbf{X})$ matrices de variables aleatorias función de la matriz aleatoria \mathbf{X} (y por tanto, $f(\mathbf{X}) \in \mathcal{L}(\mathbf{X})$ y $g(\mathbf{X}) \in \mathcal{L}(\mathbf{X})$); y sean \mathbf{v} y \mathbf{w} vectores de \mathbb{R}^n y \mathbb{R}^m respectivamente. Entonces

$$\mathbb{E}(\mathbf{Q}\mathbf{U} + \mathbf{I}\mathbf{v} | \mathbf{X}) = \mathbb{E}(\mathbf{Q}\mathbf{U} | \mathbf{X}) + \mathbb{E}(\mathbf{I}\mathbf{v} | \mathbf{X}) = \mathbf{Q}\mathbb{E}(\mathbf{U} | \mathbf{X}) + \mathbf{I}\mathbf{v}.$$

y

$$\text{Var}(\mathbf{Q}\mathbf{U} + \mathbf{I}\mathbf{v} | \mathbf{X}) = \text{Var}(\mathbf{Q}\mathbf{U} | \mathbf{X}) = \mathbf{Q}\text{Var}(\mathbf{U} | \mathbf{X})\mathbf{Q}^T.$$

además

$$\text{Cov}((\mathbf{Q}\mathbf{U} + \mathbf{I}\mathbf{v}), (\mathbf{R}\mathbf{U} + \mathbf{I}\mathbf{w}) | \mathbf{X}) = \text{Cov}(\mathbf{Q}\mathbf{U}, \mathbf{R}\mathbf{U} | \mathbf{X}) = \mathbf{Q}\text{Var}(\mathbf{U} | \mathbf{X})\mathbf{R}^T.$$

Problemas de la Lección 6

(L-6) PROBLEMA 1. Dado un *m.a.s* \mathbf{Y} de \mathbf{Y} , demuestre que la cuasi-varianza muestral $s_{\mathbf{Y}}^2$ es un estimador insesgado de $\text{Var}(\mathbf{Y})$.

Varianza de una matriz por un vector de variables aleatorias

(L-6) PROBLEMA 2. Sea $\mathbf{Q}_{m \times N}$, y \mathbf{Y} un vector con N variables aleatorias. Demuestre que $\text{Var}(\mathbf{Q}\mathbf{Y}) = \mathbf{Q} \text{Var}(\mathbf{Y})\mathbf{Q}^T$.

Varianza de los estimadores

(L-6) PROBLEMA 3. Calcule las varianzas y covarianzas de los estimadores MCO en el Modelo Lineal Simple para una muestra \mathbf{y} y \mathbf{x} .

(L-6) PROBLEMA 4. Si en particular \mathbf{v} es el vector selector $\mathbf{I}_{lj} = (0, \dots, 0, 1, 0, \dots, 0)$; es decir, el vector con un 1 en la posición j -ésima y ceros en el resto (la fila i -ésima de la matriz identidad); ¿qué implica el Teorema de Gauss-Markov para cada uno de los estimadores $\tilde{\beta}_j$?

Momentos de los valores ajustados y los errores

(L-6) PROBLEMA 5. Denotemos $\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ por \mathbf{P} . Nótese que

$$\mathbf{P} \equiv \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{XA}.$$

Verifique que $\mathbf{PX} = \mathbf{X}$. Demuestre además que $\mathbf{P}^T = \mathbf{P}$ y que $\mathbf{PP} = \mathbf{P}$; es decir, que \mathbf{P} es simétrica e idempotente.

(L-6) PROBLEMA 6. Denotemos $\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ por \mathbf{M} . Nótese que

$$\mathbf{M} \equiv \mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{XA}.$$

Verifique que $\mathbf{MX} = \mathbf{0}$, y que $\mathbf{AM} = \mathbf{0}$. Demuestre además que $\mathbf{M} = \mathbf{M}^T$ y que $\mathbf{MM} = \mathbf{M}$; es decir, que \mathbf{M} es simétrica e idempotente.

(L-6) PROBLEMA 7. Demuestre que el estimador de la suma residual es $\widehat{SRC} = \mathbf{UMU}$.

Estimación de la varianza residual

(L-6) PROBLEMA 8. Demuestre la Proposición 14.1 en la página 82, es decir, que $\text{tr}(\mathbf{M}) = N - k$.

(L-6) PROBLEMA 9. Demuestre la Proposición 14.2 en la página 82, es decir, que $\mathbb{E}(\hat{\mathbf{e}} \cdot \hat{\mathbf{e}} | \mathbf{X}) = \sigma^2(N - k)\mathbf{1}$.

Fin de los Problemas de la Lección 6

Enlace a algunas **prácticas de la Lección 6**

Fin de la lección

Part III

Inferencia en el Modelo Clásico de Regresión lineal

LECCIÓN 7: Inferencia. Contrastes de hipótesis lineales

17 Introducción a la contrastación de hipótesis

Bibliografía:

Básica: Wooldridge (2006, Secciones 4.1, 4.2 y Apéndices E3)

Complementaria: Hayashi (2000, Capítulo 1)

(Lección 7)

T-1

Contrastes de hipótesis paramétricas

Hipótesis afirmación sobre uno o varios parámetros

- H_0 : hipótesis *nula* (hipótesis cuya veracidad se cuestiona)
- H_1 : hipótesis complementaria (*alternativa*)

Contraste de hipótesis es una regla que establece

- para que valores muestrales \mathbf{X} se rechaza H_0
(*región crítica, RC*)
- para que valores muestrales \mathbf{X} no se rechaza H_0
(*región de no rechazo (\neq aceptación), RA*)

Toma de decisión sobre el rechazo o no de H_0 (Es la realización del contraste)

F77

(Lección 7)

T-2

Contrastes de hipótesis paramétricas

Caracterizamos RC mediante un estadístico $g(\mathbf{X})$.

Ejemplo

- Tren sale cada hora en punto (tarde 10' en llegar al andén)
- H_0 : **me da tiempo**
 H_1 : NO me da tiempo
- $g(\mathbf{X})$: hora media de los relojes de los presentes
- $RC = \{\mathbf{X} \text{ tales que: } g(\mathbf{X}) = m_x \geq hh : 40'\}$ (*nivel significación α*)
- Pregunto la hora, y decido si voy al andén

Pero el estadístico podría ser

- $g^*(\mathbf{X})$: hora media de los relojes de más de 60 euros.
- $RC^* = \{\mathbf{X} \text{ tales que: } g^*(\mathbf{X}) \geq hh : 45'\}$ (*nivel de significación α*)

F78

El *nivel de significación* acota el riesgo que estoy dispuesto a asumir de rechazar *erróneamente* la hipótesis nula. En el ejemplo de los relojes, podría ser que todos mis amigos llevaran el reloj adelantado. Si en tal caso decido como regla que “a menos 10” decido que no me da tiempo a tomar el tren, es probable que haya rechazado H_0 erróneamente, (muy probablemente faltan más de 10 minutos para la hora en punto si todos llevan el reloj adelantado). Para cubrir ese riesgo, uno debe graduar qué hora límite define su región crítica. El *nivel de significación* limita la probabilidad de cometer el error Tipo I (i.e., rechazar H_0 erróneamente).

1. Establecimiento de la hipótesis nula H_0 sobre θ

$$H_0 : \mathbf{X} \sim f_{\mathbf{X}}(x; \boldsymbol{\theta}); \quad \boldsymbol{\theta} \in \Theta_0$$

y la hipótesis complementaria (*alternativa*)

$$H_1 : \mathbf{X} \sim f_{\mathbf{X}}(x; \boldsymbol{\theta}); \quad \boldsymbol{\theta} \in \Theta_1$$

donde $\Theta = \Theta_0 \cup \Theta_1$, y $\Theta_0 \cap \Theta_1 = \emptyset$

2. Elección del estadístico $g(\mathbf{X})$

3. División del espacio muestral en dos regiones: *RC* (*región crítica*) y *RA* (*región no crítica*)
(dado un nivel de significación α)

$RC \cap RA = \emptyset$; $RC \cup RA =$ espacio muestral

- ¿Donde está mi muestra \mathbf{X} ?
- Cálculo del estadístico: $g(\mathbf{X})$ para decidir si $\mathbf{X} \in RC$.
- En consecuencia, Rechazo o no rechazo H_0 (toma de decisión)

F79

18 Estadístico t de Student

Proposición 18.1. Si una matriz \mathbf{Q} es idempotente entonces $\text{rg}(\mathbf{Q}) = \text{tr}(\mathbf{Q})$.

Demostración. (Demostración en Bujosa, 2022b, ultimo ejercicio de la Lección 12). \square

Proposición 18.2. Sea el vector $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$, y sea \mathbf{Q} simétrica e idempotente, entonces $\mathbf{Z}^T \mathbf{Q} \mathbf{Z} \sim \chi^2_{(\text{rg}(\mathbf{Q}))}$.

Demostración. (Demostración en Mittelhammer, 1996, pp. 329) \square

Denominamos *cuasi-varianza de $\hat{\mathbf{e}}$* al estadístico $\hat{s}_{\hat{\mathbf{e}}}^2 = \frac{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}{N-k}$, que verifica la siguiente

Proposición 18.3. $\frac{N-k}{\sigma^2} \hat{s}_{\hat{\mathbf{e}}}^2 = \frac{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}{\sigma^2} \sim \chi^2_{(N-k)}$

P-1
(94)

Además,

Proposición 18.4. Los vectores de variables aleatorias $(\hat{\beta} - \beta)$ y $\hat{\mathbf{e}}$ son probabilísticamente independientes. \square

P-3
(94)

Nota 9. Si dos variables aleatorias \mathbf{X} e \mathbf{Y} son probabilísticamente independientes, entonces transformaciones de ellas, $h(\mathbf{X})$ y $g(\mathbf{Y})$, también son probabilísticamente independientes.

Proposición 18.5. El estadístico T_j de distribuye como una t con $N - k$ grados de libertad, es decir, $T_j \sim t_{N-k}$. \square

P-4
(94)

(Lección 7) T-4 Estadístico t de Student (T) para los parámetros β_j

Bajo los supuestos muestrales:

$$\frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\sqrt{\sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})} \sim N(0, 1)$$

y sustituyendo σ^2 por su estimador, $\hat{s}_{\hat{\mathbf{e}}}^2 = \frac{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}{N-k}$, obtenemos el estadístico T

$$\frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\sqrt{\hat{s}_{\hat{\mathbf{e}}}^2 [(\mathbf{X}^T \mathbf{X})^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})} \stackrel{\text{E}(\hat{\beta}_j) = \beta_j}{\sim} t_{N-k}, \quad (37)$$

Nótese que β_j es desconocido.

F80

donde $\sqrt{\widehat{s}_{\hat{e}}^2[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}} = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j | \mathbf{X})} \equiv \widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})$

Ejemplo 11. Continuación de “precio de las viviendas”:

Dada una muestra concreta \mathbf{X} , sustituimos $(\mathbf{X}^\top \mathbf{X})^{-1}$ por la inversa de la matriz $\mathbf{X}^\top \mathbf{X}$:

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 9.1293e-01 & -4.4036e-04 \\ -4.4036e-04 & 2.3044e-07 \end{bmatrix};$$

como no conocemos σ^2 ; sustituimos su valor por la la cuasi-varianza de los errores de ajuste

$$\widehat{s}_{\hat{e}}^2 = \frac{\hat{e} \cdot \hat{e}}{N-n} = \frac{18273.6}{14-2} = 1522.8.$$

Así las desviaciones típicas estimadas de \hat{a} y \hat{b} son (véase 68 en la página 153)

$$\begin{aligned} \widehat{\text{Dt}}(\hat{a}) &= \sqrt{(1522.8) \cdot (9.1293e-01)} = 37.285 & = \sqrt{\widehat{s}_{\hat{e}}^2 \cdot \frac{\sum X_n^2}{N \sum (X_n - m_x)^2}}; \\ \widehat{\text{Dt}}(\hat{b}) &= \sqrt{(1522.8) \cdot (2.3044e-07)} = 0.01873 & = \sqrt{\widehat{s}_{\hat{e}}^2 \cdot \frac{1}{\sum (X_n - m_x)^2}} \end{aligned}$$

Véase los resultados de estimación en el ejemplo del precio de las viviendas (página 34).

Por otra parte,

$$\widehat{\text{Cov}}(\hat{a}, \hat{b}) = (1522.8) * (-4.4036e-04) = -0.671 = \widehat{s}_{\hat{e}}^2 \cdot \frac{-\sum X_n}{N \sum (X_n - m_x)^2}$$

(véase 69 en la página 153).

19 Contraste de hipótesis sobre coeficientes individuales de la regresión

19.1 Contrastos de dos colas

(Lección 7)
T-5
Contraste de la t : de dos colas

1. $H_0: \beta_j = b$; $H_1: \beta_j \neq b$
 $\frac{\hat{\beta}_j - b}{\widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})} \equiv \mathcal{T}_{H_0} \sim t_{N-k}$
3. Se rechaza H_0 cuando $|\mathcal{T}| > t_{N-k}^{(1-\alpha)}$

(α determina RC)

Distribución t con $(N - k)$ grados de libertad

$t_{N-k}^{(\alpha/2)}$ $t_{N-k}^{(1-\alpha/2)} = -t_{N-k}^{(\alpha/2)}$

$t_{N-k}^{(\alpha/2)}$ y $t_{N-k}^{(1-\alpha/2)}$ son los valores críticos

F81

Donde \widehat{T} es la realización del estadístico (i.e., la variable aleatoria) \mathcal{T} para una muestra concreta \mathbf{X} .

Ejemplo 12. Continuación de “precio de las viviendas”: Contraste de significación individual de a :

$$H_0 : a = 0; \quad H_1 : a \neq 0$$

En este caso la región crítica debe ser

$$RC = \left\{ \mathbf{X} \text{ tales que } \left| \frac{\widehat{a}-0}{\widehat{D}_t(\widehat{a}|\mathbf{X})} \right| > k_2 \right\}, \text{ donde } \frac{\widehat{a}}{\widehat{D}_t(\widehat{a}|\mathbf{X})} \equiv \mathcal{T}_a \stackrel{H_0}{\sim} t_{12}.$$

Recordando los resultados para el Modelo Lineal Simple (Ec. (68))

$$RC = \left\{ \mathbf{X} \text{ tales que } \left| \frac{\widehat{a}-0}{\sqrt{\frac{s_e^2}{N} \frac{\sum x_t^2}{\sum (x_t - m_x)^2}}} \right| < k_2 \right\}$$

donde $N = 14$.

Si $\alpha = 0.05$, el valor crítico es $k_2 = 2.18 = -k_1 = t_{12}^{(0.025)}$:

$$\widehat{T}_a = \frac{52.351}{37.285} = 1.4041 < k_2 \quad \text{no rechazamos } H_0 \text{ para } \alpha \text{ del 5%}.$$

Véase los resultados de estimación del ejemplo del precio de las viviendas (página 34).

Para $\alpha = 0.1$, el valor critico es $k_2 = 1.78 = -k_1 = t_{12}^{(0.05)}$.

?

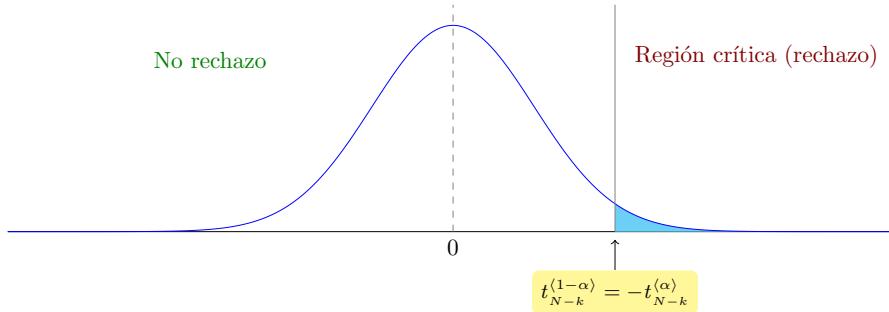
19.2 Contrastes de una sola cola

(Lección 7)

T-6 Contraste de la t : de una sola cola (derecha)

1. $H_0 : \beta_j = b; \quad H_1 : \beta_j > b$
2. (De Ec. 37) $\frac{\widehat{\beta}_j - b}{\widehat{D}_t(\widehat{\beta}_j|\mathbf{X})} \equiv \mathcal{T} \stackrel{H_0}{\sim} t_{N-k}$
3. Se rechaza H_0 cuando $\widehat{T} > t_{N-k}^{(1-\alpha)}$ (α determina RC)

Distribución t con $(T - k)$ grados de libertad



$t_{N-k}^{(1-\alpha)}$ es el valor crítico

F83

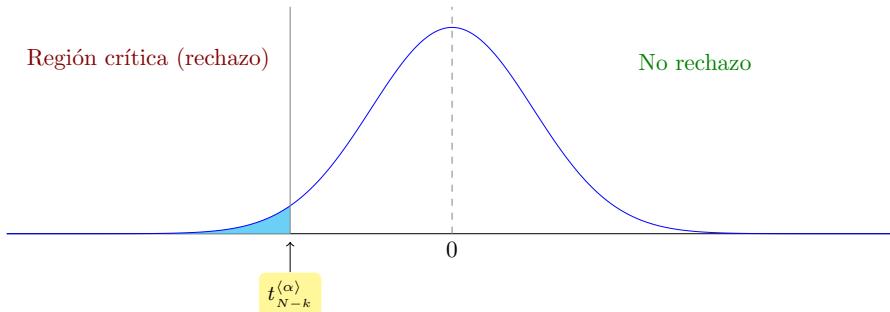
$$1. H_0 : \beta_j = b; \quad H_1 : \beta_j < b$$

$$2. (\text{De Ec. 37}) \quad \frac{\widehat{\beta}_j - b}{\widehat{D}t(\widehat{\beta}_j | \mathbf{X})} \equiv \mathcal{T}_{H_0} \sim t_{N-k}$$

3. Se rechaza H_0 cuando $\widehat{\mathcal{T}} < t_{N-k}^{(\alpha)}$

(α determina RC)

Distribución t con $(T - k)$ grados de libertad



$t_{N-k}^{(\alpha)}$ es el valor crítico

F84

Ejemplo 13. Continuación de “precio de las viviendas”: Un experto del mercado de la vivienda afirma que un pie cuadrado adicional en la superficie supone un incremento de (*como poco*) 150 dólares, pero *nunca menos*. ¿Podemos creer al experto con una significación del 2.5%?

$$H_0 : b = 0.15; \quad H_1 : b < 0.15$$

La región critica de cola izquierda

$$RC = \left\{ \mathbf{x} \mid \frac{\widehat{b} - 0.15}{\widehat{D}t(\widehat{b} | \mathbf{x})} < k \right\}$$

(o bien, para este caso particular (68))

$$RC = \left\{ \mathbf{x} \mid \frac{\widehat{b} - 0.15}{\sqrt{\frac{\widehat{s}_e^2}{\sum(x_t - m_x)^2}}} < k \right\}$$

sustituyendo valores estimados, tenemos que

$$\widehat{\mathcal{T}}_b = \frac{0.139 - 0.15}{0.01873} = -0.58729 > t_{12}^{(0.025)} = -2.18$$

?

19.3 Reglas de decisión para contrastes de una cola y de dos colas usando el p -valor

El p -valor es la probabilidad, bajo la hipótesis nula, de obtener un test estadístico al menos tan extremo como el que ha sido observado.

El investigador “rechazará la hipótesis nula” cuando considere que el p -valor es muy pequeño. Por ejemplo, si está dispuesto a realizar contrastes al 5% de nivel de significación, entonces rechazará H_0 cuando el p -valor sea menor a 0,05.

El *p*-valor es la *probabilidad* (*bajo* H_0) de obtener un *resultado* (igual o) “más extremo” que el observado.

El significado de “más extremo” depende de H_1

- p -valor = $\mathbb{P}_{H_0}(\mathcal{T}_j > \widehat{\mathcal{T}}_j)$ (cola derecha)
- p -valor = $\mathbb{P}_{H_0}(\mathcal{T}_j < \widehat{\mathcal{T}}_j)$ (cola izquierda)
- p -valor = $2 \times \min \left\{ \mathbb{P}_{H_0}(\mathcal{T}_j > \widehat{\mathcal{T}}_j) H_0, \mathbb{P}_{H_0}(\mathcal{T}_j < \widehat{\mathcal{T}}_j) \right\}$ (bilateral)

Cuando el *p*-valor es “*pequeño*” se rechaza H_0

Véase los resultados de estimación del ejemplo del precio de las viviendas

F86

Problemas de la Lección 7

Estadístico t de Student

(L-7) PROBLEMA 1. Demuestre la Proposición 18.3 en la página 90, es decir, que $\frac{N-k}{\sigma^2} \widehat{s}_{\widehat{e}}^2 = \frac{\widehat{e} \cdot \widehat{e}}{\sigma^2} \sim \chi_{(N-k)}^2$.

(L-7) PROBLEMA 2. Dado que toda variable aleatoria con distribución χ_{N-k}^2 tiene esperanza $N - k$ y varianza $2(N - k)$; y que $\widehat{s}_{\widehat{e}}^2$ es una variable aleatoria χ_{N-k}^2 multiplicada por $\frac{\sigma^2}{N-k}$; calcule la esperanza y la varianza de $\widehat{s}_{\widehat{e}}^2$.

(L-7) PROBLEMA 3. Demuestre la Proposición 18.4 en la página 90 es decir, demuestre que los vectores aleatorios $(\widehat{\beta} - \mathbf{I}\beta)$ y \widehat{e} son probabilísticamente independientes.

(L-7) PROBLEMA 4. Demuestre que \mathcal{T}_j de distribuye como una *t* con $N - k$ grados de libertad, es decir, $\mathcal{T}_j \sim t_{N-n}$.

(L-7) PROBLEMA 5. Discuta la veracidad o falsedad de la siguiente afirmación:

Si la hipótesis nula no es rechazada, entonces es cierta.

(L-7) PROBLEMA 6. Discuta la veracidad o falsedad de la siguiente afirmación:

Si se rechaza H_0 con un nivel de significación $\alpha = 0.05$ también se rechaza H_0 con un nivel de significación $\alpha = 0.10$

(L-7) PROBLEMA 7. En el siguiente modelo de regresión simple, $\mathbf{Y} = \alpha \mathbf{1} + \beta \mathbf{X} + \mathbf{U}$, \mathbf{Y} mide la tasa de crecimiento del PIB real y \mathbf{X} la tasa de crecimiento de la masa monetaria. Tenemos una muestra de tamaño 35. Queremos contrastar la hipótesis de neutralidad monetaria en esta muestra (es decir, la masa monetaria no tiene efecto real sobre el nivel del PIB). Proponga la hipótesis nula y alternativa que, según la t^a económica, crea más conveniente para realizar el contraste, así como la región crítica. ¿Cuál es la distribución del estadístico de contraste?

Fin de los Problemas de la Lección 7

Enlace a algunas **prácticas de la Lección 7**

Fin de la lección

LECCIÓN 8: Inferencia. Contrastes de hipótesis lineales (combinaciones lineales de parámetros). Intervalos y regiones de confianza

20 Contraste de hipótesis sobre combinaciones lineales de coeficientes de la regresión

Bibliografía:

Básica: Wooldridge (2006, Secciones 4.1, 4.2 y Apéndices E3)

Complementaria: Hayashi (2000, Capítulo 1)

(Lección 8) T-1 Hipótesis lineales

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

\mathbf{R} es matriz con $\text{rg}(\mathbf{R}) = r$, ($r \leq k$); y $\mathbf{r} \in \mathbb{R}^r$ es vector.

Las r ecuaciones son hipótesis sobre valores de los coeficientes.

Condición $\text{rg}(\mathbf{R}) = r$, garantiza:

- no hipótesis redundantes
- no hipótesis incompatibles

F88

Ejemplo 14. Ecuación de salarios (continuación Ejemplo 2 en la página 11):

$$\ln(SALAR) = \beta_1 \mathbf{1} + \beta_2 EDUC + \beta_3 ANTIG + \beta_4 EXPER + U$$

Supongamos que queremos contrastar si educación y antigüedad tienen el mismo efecto en el incremento del salario, y que además, la experiencia no tiene ningún efecto (por tanto $r = 2$)

$$\beta_2 = \beta_3 \quad \text{y} \quad \beta_4 = 0.$$

En forma matricial, $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, donde

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

donde \mathbf{R} cumple la condición de rango completo.

Añadiendo restricciones que no cumplen la condición de rango:

- Supongamos que adicionalmente imponemos que

$$\beta_2 - \beta_3 = \beta_4.$$

Esta es una restricción redundante, pues ya se cumple con las dos primeras restricciones; en forma matricial

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & -1 \end{bmatrix}, \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix};$$

- Supongamos que imponemos una condición incompatible con las dos primeras:

$$\beta_4 = 0.5,$$

que evidentemente es incompatible con $\beta_4 = 0$. Matricialmente

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ 0.5 \end{pmatrix}.$$

De nuevo la condición de rango se incumple.

20.1 El test F

(Lección 8)

T-2 Estadístico F

Bajo supuestos 1 a 5; y si $H_0: \mathbf{R}\beta = \mathbf{r}$ cierta,
donde $\text{rg}(\mathbf{R})_{r \times k} = r$, definimos el **Estadístico F** :

$$\mathcal{F} = (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) \left[\widehat{\text{Var}}(\mathbf{R}\hat{\beta} | \mathbf{X}) \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) / r \stackrel{H_0}{\sim} F_{r, N-k} \quad (38)$$

(39)

$$= (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) \left[\mathbf{R}\widehat{\text{Var}}(\hat{\beta} | \mathbf{X}) \mathbf{R}^\top \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) / r$$

(de la Ecuación 34) sustituyendo $\widehat{\text{Var}}(\hat{\beta} | \mathbf{X}) = \hat{s}^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1}$

$$= \frac{1}{\hat{s}^2} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) / r \quad (40)$$

F91

Nota 10. Sean $\mathbf{W} \sim \chi_r^2$ y $(\hat{\mathbf{e}} \cdot \hat{\mathbf{e}} / \sigma^2) \sim \chi_{N-k}^2$ dos variables aleatorias independientes, entonces

$$\frac{\mathbf{W}/r}{\frac{1}{\sigma^2} \hat{\mathbf{e}} \cdot \hat{\mathbf{e}} / (N - k)} \stackrel{H_0}{\sim} F_{r, N-k}$$

(Demostración en (Mittelhammer, 1996, Teorema 6.21, página 345))

Proposición 20.1 (Distribución del Estadístico F): *El Estadístico F se distribuye como una F con r y $N - k$ grados de libertad, es decir $\mathcal{F} \stackrel{H_0}{\sim} F_{r, N-k}$.* (P-1)
(102)

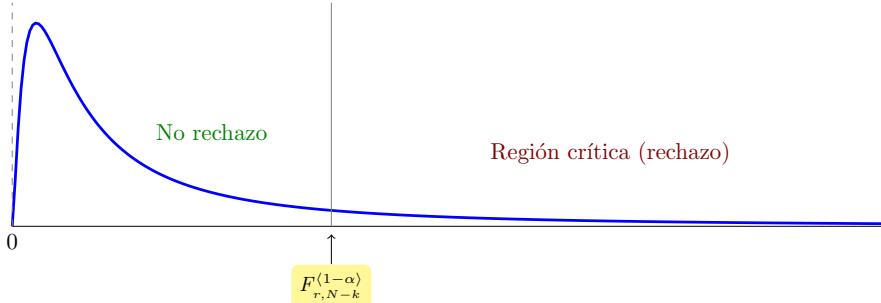
$$1. H_0 : \mathbf{R}\beta = \mathbf{r}; \quad H_1 : \mathbf{R}\beta \neq \mathbf{r}$$

$$2. (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) \left[\widehat{\text{Var}}(\mathbf{R}\hat{\beta} | \mathbf{X}) \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) / r \underset{H_0}{\sim} F_{r, N-k}$$

3. Cuando $\hat{\mathcal{F}} \in RC$ se rechaza H_0

(α determina RC)

Distribución F con $(r, N-k)$ grados de libertad



... o bien: cuando p -valor se considera pequeño, se rechaza H_0

F92

20.2 t versus F

Contrastación de hipótesis individual es caso particular, donde $r = 1$ y

$$\mathbf{R} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ & & & (j) \\ 1 \times k & & & 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{r} = b_j$$

(38) se reduce a

$$\begin{aligned} \mathcal{F} &= (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) \left[\widehat{\text{Var}}(\mathbf{R}\hat{\beta} | \mathbf{X}) \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) / 1 \\ &= (\hat{\beta}_j - b_j \mathbf{1},) \left[\widehat{\text{Var}}(\hat{\beta}_j | \mathbf{X}) \right]^{-1} (\hat{\beta}_j - b_j \mathbf{1},) \underset{H_0: \beta_j = b_j}{\sim} F_{1, N-k} \end{aligned} \quad (41)$$

que es cuadrado^a del estadístico \mathcal{T} de (37), página 90.

F93

^a $F_{1, N-k}$ es el cuadrado de una t_{N-k} .

Nota 11. Nótese que si $\mathbf{R} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ & & & (j) \\ 1 \times k & & & 0 & \cdots & 0 \end{bmatrix}$ y $\mathbf{r} = (b_j,)$, entonces:

$$\begin{aligned} \mathcal{F} &= (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) \left[\widehat{\text{Var}}(\mathbf{R}\hat{\beta} | \mathbf{X}) \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{I}\mathbf{r}) / 1 \\ &= (\hat{\beta}_j - b_j \mathbf{1},) \left[\widehat{\text{Var}}(\hat{\beta}_j | \mathbf{X}) \right]^{-1} (\hat{\beta}_j - b_j \mathbf{1},) \underset{H_0: \beta_j = b_j}{\sim} F_{1, N-k} \quad \text{pues se selecciona el elemento } j \\ &= \frac{(\hat{\beta}_j - b_j \mathbf{1},)^2}{\widehat{\text{Var}}(\hat{\beta}_j | \mathbf{X})} = \left(\frac{\hat{\beta}_j - b_j \mathbf{1}}{\widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})} \right)^2 = (\mathcal{T})^2 \quad \text{por ser un escalar} \end{aligned}$$

donde

$$\mathcal{T} \underset{H_0}{\sim} t_{N-k}$$

Si el contraste individual es de una sola cola, entonces no se puede realizar mediante el estadístico \mathcal{F} . El estadístico \mathcal{F}

sólo tiene en cuenta las desviaciones al cuadrado, por lo tanto, penaliza de igual manera una diferencia positiva entre H_0 y la estimación obtenida, que si la diferencia es negativa, es decir, que no distingue la cola derecha de la izquierda. Si el contraste individual es bilateral, entonces se puede realizar empleando el estadístico \mathcal{F} , que en el caso bilateral es idéntico al cuadrado del estadístico \mathcal{T} .

Nota 12. No solo el contraste de significación individual tiene una distribución $(\mathcal{T})^2$. Si $\mathbf{R} = [r_1, \ r_2, \ \dots \ r_k]_{1 \times k}$ y, consecuentemente, \mathbf{r} tiene una única componente (es decir, si hay una única restricción lineal), el estadístico resultante siempre es $\mathcal{F} = (\mathcal{T})^2$; veámoslo:

$$\mathcal{F} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) \left[\widehat{\text{Var}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) / 1$$

operando tenemos:

$$= (r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k - b\mathbb{1},) \left[\widehat{\text{Var}}(r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k | \mathbf{X}) \right]^{-1} (r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k - b\mathbb{1},)$$

y por ser una expresión escalar:

$$= \frac{(r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k - b\mathbb{1})^2}{\widehat{\text{Var}}(r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k | \mathbf{X})} = \left(\frac{r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k - b\mathbb{1}}{\widehat{\text{Dt}}(r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k | \mathbf{X})} \right)^2 = (\mathcal{T})^2,$$

ya que $r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k$ es una combinación lineal de Normales, es una variable aleatoria escalar con distribución Normal⁸³; así pues, la expresión dentro del paréntesis cumple las mismas propiedades descritas en la Proposición 18.5 de la página 90.

(Lección 8) T-5 Contraste t para una combinación lineal de betas

Si $\mathbf{R} = [r_1, \ r_2, \ \dots \ r_k]_{1 \times k}$ y $b = \mathbf{R}\boldsymbol{\beta}$, entonces

$$\frac{(r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k - b\mathbb{1})}{\widehat{\text{Dt}}(r_1\widehat{\beta}_1 + \dots + r_k\widehat{\beta}_k | \mathbf{X})} = \frac{\mathbf{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbb{1})}{\widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X})} = \mathcal{T}_{H_0} \sim t_{N-k}.$$

F95

Significación conjunta del modelo Con el contraste \mathcal{F} se puede contrastar la significación conjunta de varios (o todos) los coeficientes de la regresión.

Por ejemplo, para el contraste de significación conjunta basta con:

$$\mathbf{R} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ (k-1) \times k & (k-1) \times (k-1) \end{bmatrix}; \quad \mathbf{r} = \mathbf{0} \in \mathbb{R}^{k-1}.$$

En este caso:

$$\begin{aligned} H_0 : & \text{ todos los coeficientes (excepto el de la constante) son nulos;} \\ H_1 : & \text{ al menos uno es distinto de cero.} \end{aligned}$$

Este contraste no es equivalente a realizar $k-1$ contrastes individuales *por separado*.

Este estadístico F para la contrastación de significación conjunta es el que aparece en las regresiones de los programas informáticos. Véase por ejemplo el “recuadro de resultados de estimación del precio de las viviendas” (pag. 34).

⁸³Nótese que si los estimadores de las covarianzas de los estimadores $\widehat{\text{Cov}}(\widehat{\beta}_i, \widehat{\beta}_j | \mathbf{X})$ fuesen cero, entonces

$$\left[\widehat{\text{Var}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right] = \left[r_1^2 \widehat{\text{Var}}(\widehat{\beta}_1 | \mathbf{X}) + \dots + r_k^2 \widehat{\text{Var}}(\widehat{\beta}_k | \mathbf{X}) \right]$$

es decir, la suma de las varianzas multiplicadas por los factores r_j^2 ; pero como en general dichas covarianzas nunca son cero, la expresión es más complicada pues se deben añadir los términos $2r_i r_j \widehat{\text{Cov}}(\widehat{\beta}_i, \widehat{\beta}_j | \mathbf{X})$, para todos los pares $i \neq j$.

- En este contraste las hipótesis son

$$H_0 : \text{todos los coeficientes (excepto el de la constante) son nulos};$$

$$H_1 : \text{al menos uno es distinto de cero}.$$

- Este contraste no es equivalente a realizar $k - 1$ contrastes individuales *por separado*.
- Es un contraste F y su valor y p -valor se muestran en las regresiones por MCO.

Véase los resultados de estimación del ejemplo del precio de las viviendas (con esto ya sabe que significan casi todos los números del cuadro de resultados) (página 34).

F96

21 Regiones e intervalos de confianza

Sea $\mathbf{R}_{1 \times k}$, de la Nota 12 de la Página 98 (o bien de (38) y (41)) se deduce la primera expresión de la siguiente transparencia:

El test *t*-Student *bilateral* rechaza $H_0 : \mathbf{R}_{1 \times k} \boldsymbol{\beta} = r$ si

$$|\mathcal{T}| = \frac{|\mathbf{R}\hat{\boldsymbol{\beta}} - r\mathbf{1}|}{\widehat{\text{Dt}}(\mathbf{R}\hat{\boldsymbol{\beta}} | \mathbf{X})} > t^{(1-\alpha/2)}, \quad \boxed{\text{F81}}$$

donde α es el nivel de significación; por tanto

$$\begin{aligned} |\mathcal{T}| > t^{(1-\alpha/2)} &\Leftrightarrow |\mathbf{R}\hat{\boldsymbol{\beta}} - r\mathbf{1}| > t^{(1-\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\hat{\boldsymbol{\beta}} | \mathbf{X}) \\ &\Leftrightarrow |r\mathbf{1} - \mathbf{R}\hat{\boldsymbol{\beta}}| > t^{(1-\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\hat{\boldsymbol{\beta}} | \mathbf{X}) \\ &\Leftrightarrow (r\mathbf{1} - \mathbf{R}\hat{\boldsymbol{\beta}}) \notin [\pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\hat{\boldsymbol{\beta}} | \mathbf{X})] \\ &\Leftrightarrow r\mathbf{1} \notin [\mathbf{R}\hat{\boldsymbol{\beta}} \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\hat{\boldsymbol{\beta}} | \mathbf{X})] \end{aligned} \quad (42)$$

No se rechaza H_0 si y solo si: $r\mathbf{1} \in [\mathbf{R}\hat{\boldsymbol{\beta}} \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\hat{\boldsymbol{\beta}} | \mathbf{X})] = \widehat{\text{IC}}_{1-\alpha}^{\mathbf{R}\hat{\boldsymbol{\beta}}}.$

Si $\mathbf{R} = [0 \quad \cdots \quad 0 \quad \underset{(j)}{1} \quad 0 \quad \cdots \quad 0]_{1 \times k}$: el test *t*-student *bilateral* rechaza $H_0 : \mathbf{R}\boldsymbol{\beta} = \beta_j = b$ si

$$|\mathcal{T}_j| = \frac{|\hat{\beta}_j - b\mathbf{1}|}{\widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})} > t^{(1-\alpha/2)}$$

$$\begin{aligned} |\mathcal{T}_j| > t^{(1-\alpha/2)} &\Leftrightarrow |\hat{\beta}_j - b\mathbf{1}| > t^{(1-\alpha/2)} \cdot \widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X}) \\ &\Leftrightarrow |b\mathbf{1} - \hat{\beta}_j| > t^{(1-\alpha/2)} \cdot \widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X}) \\ &\Leftrightarrow (b\mathbf{1} - \hat{\beta}_j) \notin [\pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})] \\ &\Leftrightarrow b\mathbf{1} \notin [\hat{\beta}_j \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})] \end{aligned} \quad (43)$$

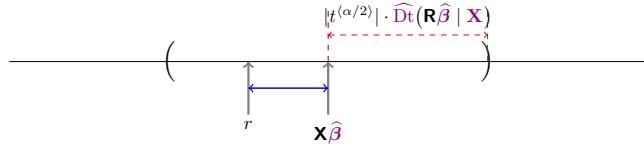
No se rechaza H_0 si y solo si: $b\mathbf{1} \in [\hat{\beta}_j \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\hat{\beta}_j | \mathbf{X})] = \widehat{\text{IC}}_{1-\alpha}^{\hat{\beta}_j}.$

F98

Denominamos *intervalo de confianza* a:

$$\widehat{IC}_{1-\alpha}^{\widehat{R}\widehat{\beta}} \equiv \left[\widehat{R}\widehat{\beta} \pm t^{\langle\alpha/2\rangle} \cdot \widehat{Dt}(\widehat{R}\widehat{\beta} | \mathbf{X}) \right].$$

$$\widehat{IC}_{1-\alpha}^{\widehat{R}\widehat{\beta}} = \{ \text{Hipótesis no rechazables para } \mathbf{R}\beta \text{ con significación } \alpha \}$$



$$H_0 : \mathbf{R}\beta = r \text{ no se rechaza si: } r \mathbb{1} \in \widehat{IC}_{1-\alpha}^{\widehat{R}\widehat{\beta}}.$$

F99

Intervalos de confianza frente a contrastes de hipótesis

Contrastes bilaterales al $\alpha\%$

$(\mathbf{R})_{1 \times k}$

- Rechazar $H_0 : \mathbf{R}\beta = c$; frente $H_1 : \mathbf{R}\beta \neq c$;

$$\text{si y sólo si } c \mathbb{1} \notin \widehat{IC}_{1-\alpha}^{\widehat{R}\widehat{\beta}}$$

- Rechazar $H_0 : \beta_j = b_j$; frente $H_1 : \beta_j \neq b_j$;

$$\text{si y sólo si } b_j \mathbb{1} \notin \widehat{IC}_{1-\alpha}^{\widehat{\beta}_j}$$

En el caso del intervalo de un parámetro individual (41) $\mathbf{R} = \begin{bmatrix} 0 & \cdots & 0 & 1_{(j)} & 0 & \cdots & 0 \end{bmatrix}_{1 \times k}$,

$$\widehat{IC}_{1-\alpha}^{\widehat{\beta}_j} = \left[\widehat{\beta}_j \pm t_{N-k}^{\langle\alpha/2\rangle} \cdot \widehat{Dt}(\widehat{\beta}_j | \mathbf{X}) \right]$$

La amplitud de los intervalos de confianza es una medición de la precisión de los estimadores. Fijado un nivel de confianza $0 < 1 - \alpha < 1$, mayor imprecisión, i.e., a mayor desviación típica del estimador, $\widehat{Dt}(\widehat{\beta}_j | \mathbf{X})$, mayor amplitud del intervalo.

Ejemplo 15. Continuación de “precio de las viviendas”:

Los intervalos de confianza de los parámetros a y b son de la forma

$$\widehat{IC}_{1-\alpha}^{\widehat{\beta}_j} = \left[\widehat{\beta}_j \pm t_{N-k}^{\langle\alpha/2\rangle} \cdot \widehat{Dt}(\widehat{\beta}_j | \mathbf{X}) \right]$$

por tanto, en el caso del efecto marginal de la superficie sobre el precio y de la constante sus estimaciones son respectivamente

$$\widehat{IC}_{1-\alpha}^{\widehat{b}} = \left[0.139 \pm (t_{12}^{\langle\alpha/2\rangle}) \cdot 0.01873 \right];$$

$$\widehat{IC}_{1-\alpha}^{\widehat{a}} = \left[52.3509 \pm (t_{12}^{\langle\alpha/2\rangle}) \cdot 37.285 \right];$$

☞ Código: EjPvivienda2.inp

(Lección 8) T-10 Estimación por intervalos de confianza (de una combinación lineal de betas)

Si se cumplen los supuestos: $\frac{\mathbf{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}1)}{\widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X})} \sim t_{N-k}$, donde $\mathbf{R} : r \times k$:

$$\begin{aligned} \mathbb{P}_{H_0} \left(t^{(\alpha/2)} < \frac{\mathbf{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}1)}{\widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X})} < t^{(1-\alpha/2)} \right) &= 1 - \alpha \\ \mathbb{P}_{H_0} \left(|\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}1| < t^{(1-\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right) &= 1 - \alpha \\ \mathbb{P}_{H_0} \left(|\mathbf{R}\boldsymbol{\beta}1 - \mathbf{R}\widehat{\boldsymbol{\beta}}| < t^{(1-\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right) &= 1 - \alpha \\ \mathbb{P}_{H_0} \left(\mathbf{R}\boldsymbol{\beta}1 \in \left[\mathbf{R}\widehat{\boldsymbol{\beta}} \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right] \right) &= 1 - \alpha \\ \mathbb{P}_{H_0} \left(\mathbf{R}\boldsymbol{\beta}1 \in \widehat{\text{IC}}_{1-\alpha}^{\mathbf{R}\widehat{\boldsymbol{\beta}}} \right) &= 1 - \alpha. \end{aligned} \quad (44)$$

$\widehat{\text{IC}}_{1-\alpha}^{\mathbf{R}\widehat{\boldsymbol{\beta}}}$ se denomina *estimador por intervalo* de $\mathbf{R}\boldsymbol{\beta}$ y $1 - \alpha$ es el *nivel de confianza* del intervalo.

F101

(Lección 8) T-11 Regiones de confianza

Si \mathbf{R} es de rango r , la condición

$$\mathcal{F} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}1) \left[\widehat{\text{Var}}(\mathbf{R}\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}1)/r \leq c$$

define un elipsoide en \mathbb{R}^k . De esta manera, de 38 en la página 96 se deduce que

$$\begin{aligned} \mathbb{P}_{H_0}(\mathcal{F} < F^{(1-\alpha)}) &= 1 - \alpha && \text{(operando como para el test-}t\text{)} \\ \mathbb{P}_{H_0}(\mathbf{R}\widehat{\boldsymbol{\beta}} \in \widehat{\text{IC}}_{1-\alpha}^{\mathbf{R}\widehat{\boldsymbol{\beta}}}) &= 1 - \alpha, \end{aligned}$$

donde $\widehat{\text{IC}}_{1-\alpha}^{\mathbf{R}\widehat{\boldsymbol{\beta}}} \subset \mathbb{R}^r$ se denomina *elipse (o elipsoide) de confianza*.

$\widehat{\text{IC}}_{1-\alpha}^{\mathbf{R}\widehat{\boldsymbol{\beta}}}$ contiene los vectores $\mathbf{r} \in \mathbb{R}^r$ tales que $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ no se rechaza con un nivel de significación α .

F102

Ejemplo 16. Región de confianza de dos parámetros: $H_0 : \beta_1 = a$, y $\beta_2 = b$; $k = 2$; $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$; $\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$; $\mathbf{r} = \begin{pmatrix} a \\ b \end{pmatrix}$.

solución tentativa pero incorrecta

No rechazar si

$$\begin{pmatrix} a \\ b \end{pmatrix} \in \text{región tal que } \begin{cases} a \in [\widehat{\beta}_1 \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\widehat{\beta}_1 | \mathbf{X})] \\ b \in [\widehat{\beta}_2 \pm t^{(\alpha/2)} \cdot \widehat{\text{Dt}}(\widehat{\beta}_2 | \mathbf{X})] \end{cases}$$

que es un rectángulo (formado por el producto cartesiano de los intervalos de confianza individuales).

solución correcta

No rechazar si

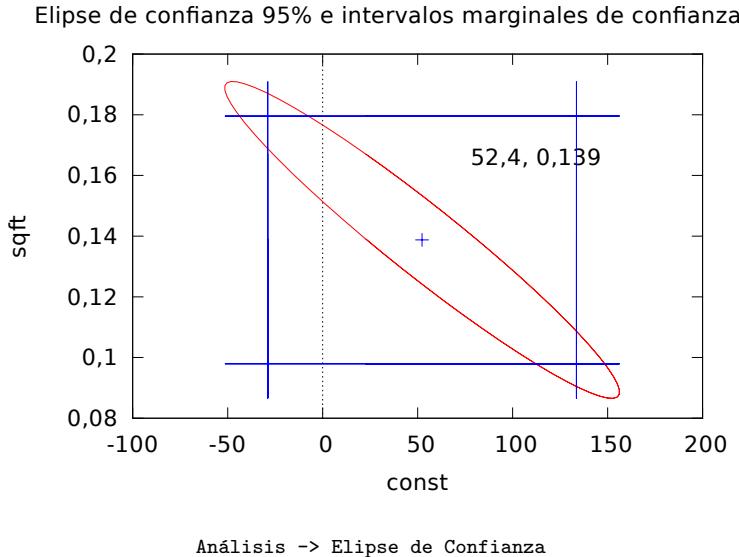
$$\begin{pmatrix} a \\ b \end{pmatrix} \in \left\{ \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \mid (\widehat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[\widehat{\text{Var}}(\widehat{\boldsymbol{\beta}} | \mathbf{X}) \right]^{-1} (\widehat{\boldsymbol{\beta}} - \mathbf{r}) < 2 \cdot F_{2, N-2}^{(1-\alpha)} \right\}$$

que es una elipse.

mostrar regiones con Gretl. Diferencia entre test de la F “una elipsoide” y tests de la t “un rectángulo” para dos parámetros. Explicar que la correlación positiva (negativa) entre los estimadores “estira” la elipse hacia los cuadrantes 1 y 3 (2 y 4).

Ejemplo 17. Continuación de “precio de las viviendas”

■ Código: [EjPvivienda2.inp](#)



Problemas de la Lección 8

Contraste de hipótesis sobre combinaciones lineales de coeficientes de la regresión

(L-8) PROBLEMA 1. Demuestre la Proposición 20.1 en la página 96, es decir, que el ratio- F se distribuye como una F con r y $N - k$ grados de libertad, $\mathcal{F} \underset{H_0}{\sim} F_{r, N-k}$.

(L-8) PROBLEMA 2. [Novales (1993, Ejercicio 4.7; pag 153)]

(a) Contrastar $H_0 : 3\beta_2 - \beta_3 = 7$ frente a $H_1 : 3\beta_2 - \beta_3 \neq 7$ en el modelo $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \mathbf{U}$; si se han obtenido los siguientes resultados: $\widehat{\beta}_2 = 4.5$; $\widehat{\beta}_3 = 5.0$; SRC=40; $N = 25$;

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 50 & 13 & 21 & -120 \\ & 4 & 2 & -20 \\ & & 6 & -10 \\ & & & 10 \end{bmatrix}$$

(b) ¿Cómo se contrastaría la hipótesis anterior junto con la hipótesis $\beta_4 = 0$, si se ha estimado $\widehat{\beta}_4 = -1.5$?

Fin de los Problemas de la Lección 8

Enlace a algunas **prácticas de la Lección 8**

Fin de la lección

LECCIÓN 9: Mínimos cuadrados restringidos y contrastes de hipótesis lineales

22 Estimación bajo restricciones lineales generales

Bibliografía:

Básica: Wooldridge (2006, Secciones 4.1, 4.2 y Apéndices E3)

Complementaria: Hayashi (2000, Capítulo 1)

(Lección 9)

T-1

Estimación restringida

Motivos:

- análisis previo → restricciones plausibles
(restricciones correctas → estimación más precisa)
- comparación entre estimación restringida y no restringida permite contrastar la validez de las restricciones

Ejecución:

- por sustitución
- método de mínimos cuadrados restringidos linealmente (MCR)

F106

Ejemplo 18. **Estimación restringida vía sustitución** Suponga el modelo en logaritmos (de una función de Cobb-Douglas):

$$\ln Y = \beta_1 \ln K + \beta_2 \ln L + \beta_3 \ln U$$

Considere la restricción: $\beta_2 + \beta_3 = 1$. La estimación imponiendo *rendimientos constantes a escala* se logra reescribiendo el modelo:

$$\begin{aligned}\ln Y &= \beta_1 \ln K + \beta_2 \ln L + (1 - \beta_2) \ln U \\ \ln Y - \ln L &= \beta_1 \ln K + \beta_2 (\ln L - \ln U) + U \\ \ln \frac{Y}{L} &= \beta_1 \ln K + \beta_2 \ln \frac{L}{U} + U\end{aligned}$$

y estimando por MCO el modelo con los nuevos regresores.

...pero hay otra forma de lograrlo...

(Lección 9)

T-2

Mínimos cuadrados restringidos (MCR)

Bajo los supuestos habituales, buscamos un estimador $\widehat{\beta}^*$ que cumpla el conjunto de restricciones lineales:

$$\mathbf{R}\widehat{\beta}^* = \mathbf{I}\mathbf{r}; \quad \text{rg}(\mathbf{R})_{r \times k} = r.$$

El estimador de **Mínimos Cuadrados con Restricciones Lineales**

$$\widehat{\beta}^* = \widehat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R} \widehat{\beta} - \mathbf{I}\mathbf{r}) \quad (45)$$

La estimación correspondiente a la muestra \mathbf{X} es

$$\widehat{\beta}^* = \widehat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R} \widehat{\beta} - \mathbf{r}) \quad (46)$$

F108

En las expresiones de más arriba, nótese que $\text{Var}(\mathbf{R}\widehat{\beta}^* | \mathbf{X}) = \sigma^2 \mathbf{R} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top$, y $\text{Cov}(\widehat{\beta}, \mathbf{R}\widehat{\beta}^* | \mathbf{X}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top$.

Así

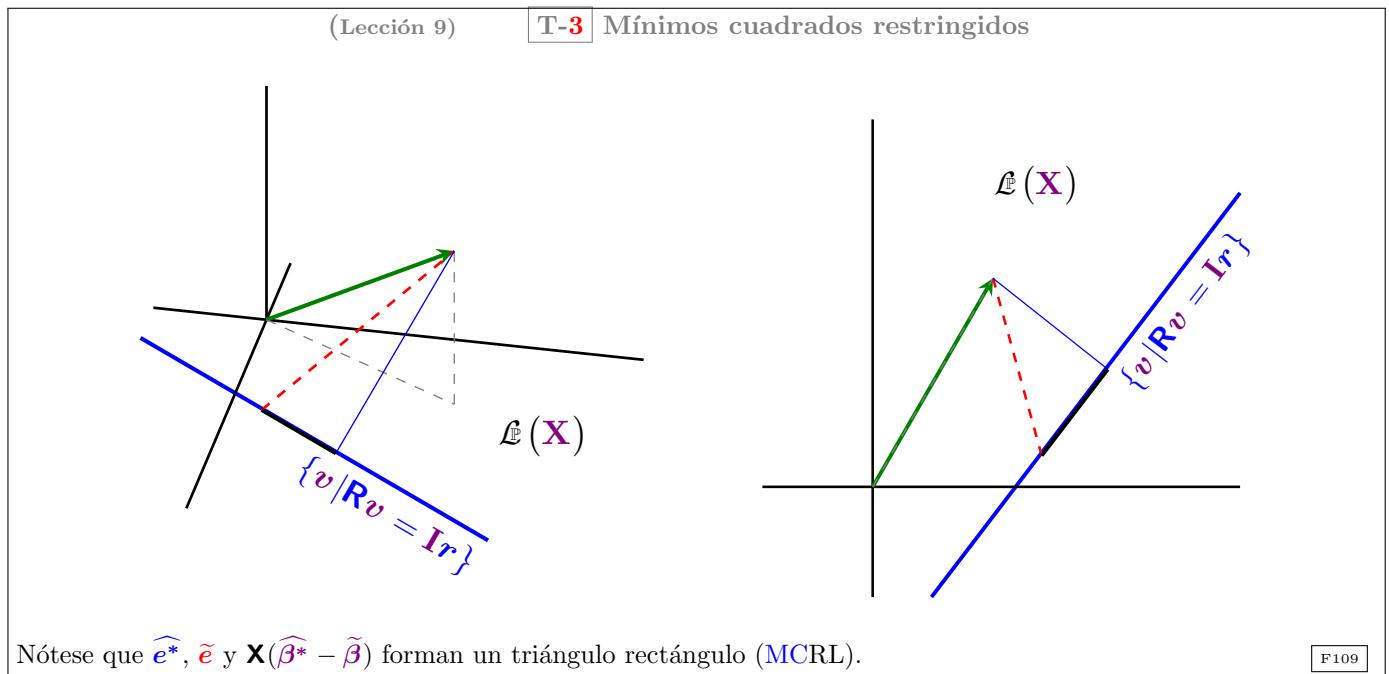
$$\widehat{\beta}^* = \widehat{\beta} - \text{Cov}(\widehat{\beta}, \mathbf{R}\widehat{\beta} \mid \mathbf{X}) [\text{Var}(\mathbf{R}\widehat{\beta} \mid \mathbf{X})]^{-1} \underbrace{\left(\mathbf{R}\widehat{\beta} - \mathbf{I}\mathbf{r} \right)}_{\text{error de predicción MCO}}$$

corrección en función del error cometido

Lo más relevante del estimador $\widehat{\beta}^*$ de *Mínimos Cuadrados con Restricciones Lineales* (MCR) se enuncia en la siguiente proposición; a saber, que de todos los estimadores que cumplen la restricción $\mathbf{R}\tilde{\beta} = \mathbf{I}\mathbf{r}$, éste es el que tiene asociado el vector de errores más pequeño.

Proposición 22.1. *De todos los estimadores $\tilde{\beta}$ que verifican la restricción lineal $\mathbf{R}\tilde{\beta} = \mathbf{I}\mathbf{r}$, el estimador de Mínimos Cuadrados con Restricciones Lineales, $\widehat{\beta}^*$, alcanza la mínima suma de residuos al cuadrado $\text{SRC}(\tilde{\beta})$.*

Demostración. Véase la Sección 24.1 \square



En la sección de ejercicios de esta sección se le pide que encuentre la expresión del estimador restringido en el Modelo Lineal Simple cuando uno de los parámetros se restringe a cero.

22.1 Propiedades estadísticas del estimador MCR

(Lección 9) T-4 Estimador MCRL

El estimador **siempre verifica** la condición: $\mathbf{R}\widehat{\boldsymbol{\beta}}^* = \mathbf{I}r$

Si $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ se cumple (restricción es cierta), de (45)

$$E(\widehat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta} \quad \text{¡sólo cuando se cumple restricción!... } (\boldsymbol{\beta} \text{ es desconocido})$$

y además, tanto si la restricción es cierta como si no

$$\text{Var}(\widehat{\boldsymbol{\beta}} | \mathbf{X}) \geq \text{Var}(\widehat{\boldsymbol{\beta}}^* | \mathbf{X})$$

ya que

$$\text{Var}(\widehat{\boldsymbol{\beta}}^* | \mathbf{X}) = \text{Var}(\widehat{\boldsymbol{\beta}} | \mathbf{X}) - \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1},$$

donde las tres matrices son **definidas positivas**.

F110

Proposición 22.2. Si la restricción es cierta, es decir, si $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, entonces el estimador $\widehat{\boldsymbol{\beta}}^*$ es insesgado.

Demostración. Si la restricción es cierta, es decir, si $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, entonces de (45)

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}^* | \mathbf{X}) &= E\left(\widehat{\boldsymbol{\beta}} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}r) \mid \mathbf{X}\right) \\ &= \mathbf{I}\boldsymbol{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} \mathbf{I}(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \\ &= \mathbf{I}\boldsymbol{\beta} \end{aligned} \quad \text{pues } \mathbf{R}\boldsymbol{\beta} = \mathbf{r}.$$

Así pues, $E(\widehat{\boldsymbol{\beta}}^*) = E(E(\widehat{\boldsymbol{\beta}}^* | \mathbf{X})) = E(\mathbf{I}\boldsymbol{\beta}) = \boldsymbol{\beta}$. □

Proposición 22.3. El estimador $\widehat{\boldsymbol{\beta}}^*$ tiene una matriz de varianzas y covarianzas menor o igual que el estimador $\widehat{\boldsymbol{\beta}}$.

Demostración. Véase la Sección 24.2 □

Aunque MCO y MCR son estimadores lineales, esto no contradice el T^a de Gauss-Markov (en la página 78):

- Cuando los parámetros $\boldsymbol{\beta}$ verifican la restricción $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, los estimadores $\widehat{\boldsymbol{\beta}}^*$ y $\widehat{\boldsymbol{\beta}}$ son idénticos, y por tanto ambos son insesgados y ambos tienen la misma varianza.
- Cuando los parámetros $\boldsymbol{\beta}$ NO verifican la restricción $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, el estimador $\widehat{\boldsymbol{\beta}}^*$ es sesgado (el teorema solo habla de los estimadores insesgados).

22.2 Contraste de la F mediante sumas residuales

Hay una expresión alternativa del estadístico \mathcal{F} empleado para la contrastación de la hipótesis nula $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. De (51) deducimos que

$$\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^* - \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} = (\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X} (\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}),$$

y por (45) sabemos que $(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}r)$; así pues, tenemos que

$$\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^* - \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}r) [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}r)$$

Sustituyendo $(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}r) [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}r)$ por $\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^* - \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}}$ en el numerador del estadístico \mathcal{F} de la Ecuación 40 en la página 96; y teniendo en cuenta que $s_e^2 = \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} / (N - k)$, podemos re-escribir el estadístico tal como

aparece en la Ecuación (47), más abajo:

(Lección 9)

T-5 Contraste de la F mediante sumas residuales

Como

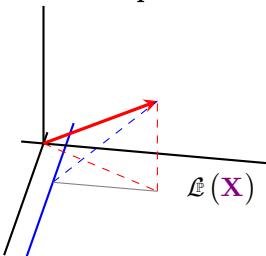
$$\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^* - \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r})$$

de (40)

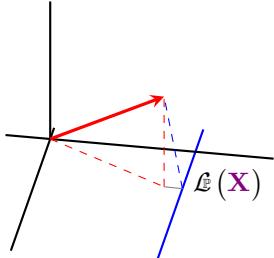
$$F = \frac{(\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^* - \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}})/r}{\widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}}/(N-k)} = \frac{N-k}{r} \cdot \frac{SRC^* - SRC}{SRC} \underset{H_0}{\sim} F_{r, N-k} \quad (47)$$

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

Restricción poco creíble



Restricción creíble



F111

Nótese que bajo la hipótesis nula, el numerador debe ser cero, y que cuanto mayor es el numerador (cuanto más difiere de cero), mayor evidencia hay en contra de que la H_0 sea cierta; así pues, este contraste es de una sola cola, la cola de la derecha.

P-1
(111)

Contraste de la F calculado con coeficientes de determinación. En ambos modelos, restringido y sin restringir, se verican relaciones similares pues $SRC = STC - SEC$, y $SRC^* = STC - SEC^*$; y por tanto, $R^2 = \frac{SEC}{STC}$ y $R^{2*} = \frac{SEC^*}{STC}$. Así podemos re-escribir (47) como:

$$F = \frac{N-k}{r} \cdot \frac{(STC - SEC^*) - (STC - SEC)}{STC - SEC} \quad \text{por ser modelo con término cte.}$$

$$= \frac{N-k}{r} \cdot \frac{SEC - SEC^*}{STC - SEC}$$

$$= \frac{N-k}{r} \cdot \frac{R^2 - R^{2*}}{1 - R^2} \quad \text{dividiendo y multiplicando por } STC$$

Contraste de significación global. Si el modelo restringido tiene como único regresor a la constante, entonces el coeficiente de determinación R^{2*} es siempre cero, por lo que el estadístico se reduce a

$$\frac{N-k}{r} \cdot \frac{\textcolor{violet}{R}^2}{1-\textcolor{violet}{R}^2} \underset{H_0}{\sim} F_{k-1, N-k}$$

(Lección 9)

T-6 Contraste de la F en modelos con constante

$$\mathcal{F} = \frac{N-k}{r} \cdot \frac{\textcolor{violet}{R}^2 - R^{2*}}{1-\textcolor{violet}{R}^2} \underset{H_0}{\sim} F_{r, N-k}$$

Contraste de significación global

$$\mathcal{F} = \frac{N-k}{k-1} \cdot \frac{\textcolor{violet}{R}^2}{1-\textcolor{violet}{R}^2} \underset{H_0}{\sim} F_{k-1, N-k} \quad (\text{caso especial})$$

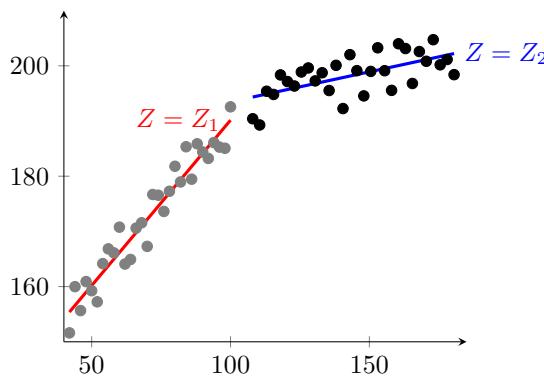
F112

22.2.1 Test de Chow

- Novales (1993, Sección 4.10, pps. 139–140)
- Novales (1997, Sección 15.3.4, pps. 572)
- Wooldridge (2006, Sección 7.4 y Sección 13.1)

(Lección 9)

T-7 Cambio estructural del modelo



F113

(Lección 9)

T-8 Contrastes de cambio estructural: *Test de Chow*

H_0 : parámetros no varían en la muestra (No cambio estructural)

H_1 : σ^2 cte., pero betas toman dos conjuntos de valores.

Modelo sin restringir

$$Y_n = {}_{n|} \mathbf{X} \boldsymbol{\beta}_A + U_n \quad n \in \{\text{índices correspondientes al caso } A\}$$

$$Y_n = {}_{n|} \mathbf{X} \boldsymbol{\beta}_B + U_n \quad n \in \{\text{índices correspondientes al caso } B\},$$

Modelo restringido $H_0 : \boldsymbol{\beta}_A = \boldsymbol{\beta}_B$, es decir,

$$Y_n = {}_{n|} \mathbf{X} \boldsymbol{\beta} + U_n \quad n = 1 : N ,$$

$U_n \sim N(0, \sigma^2)$ para $n = 1, \dots, N$ en ambos modelos.

F114

Modelo sin restringir $2k$ parámetros estimados (β_A, β_B);

y además $SR\mathcal{C} = SR\mathcal{C}_A + SR\mathcal{C}_B$.

Modelo restringido k restricciones lineales: $(\beta_A)_{|j} = (\beta_B)_{|j}; j = 1 : k$.

Por lo tanto,

$$\begin{aligned}\mathcal{F} &= \frac{N-2k}{k} \frac{SR\mathcal{C}^* - SR\mathcal{C}}{SR\mathcal{C}} \\ &= \frac{N-2k}{k} \frac{SR\mathcal{C}^* - (SR\mathcal{C}_A + SR\mathcal{C}_B)}{(SR\mathcal{C}_A + SR\mathcal{C}_B)}\end{aligned}$$

F115

Añadir práctica. Por ejemplo el Example 8.8 o 9.5 del Gujarati (table 8.9).

23 Contraste de normalidad Jarque-Bera

El contraste de **Jarque-Bera (JB)** es un contraste sobre la bondad de ajuste de la asimetría y el apuntamiento (exceso de curtosis) de la muestra de datos con respecto a la distribución normal. El nombre del contraste proviene de sus proponentes Carlos Jarque y Anil K. Bera.

$$JB = \frac{N-k}{6} \left(S^2 + \frac{1}{4}(K-3)^2 \right)$$

donde S es el coeficiente de asimetría muestral, y K el coeficiente de curtosis

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n} \sum_{i=1}^N (x_i - m_x)^3}{\left(\frac{1}{n} \sum_{i=1}^N (x_i - m_x)^2 \right)^{3/2}} \quad (48)$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - m_x)^4}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - m_x)^2 \right)^2}, \quad (49)$$

donde $\hat{\mu}_3$ y $\hat{\mu}_4$ son las estimaciones del tercer y cuarto momentos centrados respectivamente, m_x es la media muestral y $\hat{\sigma}^2$ es la estimación del segundo momento centrado (la varianza muestral).

Si la muestra proviene de una distribución normal, el contraste JB se distribuye asintóticamente como una χ^2_2

(Gretl dispone de varios contrastes de normalidad, entre ellos el JB)

F116

Así que este estadístico se puede usar para contrastar si la muestra proviene de una distribución normal. La hipótesis nula es la hipótesis conjunta: asimetría cero y exceso de curtosis cero; pues las muestras de una distribución normal tiene una asimetría esperada y un exceso de curtosis esperado nulos (que es lo mismo que decir la curtosis esperada es 3).

Dada la definición de JB , cualquier desviación de estos valores esperados aumenta el valor del estadístico JB .

24 Apéndice: Demostraciones

24.1 Demostración de la Proposición 22.1 (*minimización de residuos sujeto a restricción*)

Primero demostramos que $\widehat{\beta}^*$ cumple la restricción:

$$\begin{aligned}\widehat{\beta}^* &= \widehat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r) \\ \mathbf{R}\widehat{\beta}^* &= \mathbf{R}\widehat{\beta} - \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r) && \text{premultipliando por } \mathbf{R} \\ \mathbf{R}\widehat{\beta}^* &= \mathbf{R}\widehat{\beta} - (\mathbf{R}\widehat{\beta} - \mathbf{I}r) && \text{pues } \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} = \mathbf{I} \\ \mathbf{R}\widehat{\beta}^* &= \mathbf{I}r\end{aligned}$$

La demostración de la segunda parte también es sencilla... pero un poco más larga. Puesto que $\tilde{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\tilde{\beta}$, sumando y restando $\mathbf{X}\widehat{\beta}^*$ tenemos

$$\tilde{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\widehat{\beta}^* - (\mathbf{X}\tilde{\beta} - \mathbf{X}\widehat{\beta}^*) = (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) + \mathbf{X}(\widehat{\beta}^* - \tilde{\beta})$$

así pues $SR(\tilde{\beta}) \equiv \tilde{\mathbf{e}} \cdot \tilde{\mathbf{e}}$, es

$$\begin{aligned}\tilde{\mathbf{e}} \cdot \tilde{\mathbf{e}} &= (\mathbf{Y} - \mathbf{X}\tilde{\beta}) \cdot (\mathbf{Y} - \mathbf{X}\tilde{\beta}) \\ &= ((\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) + \mathbf{X}(\widehat{\beta}^* - \tilde{\beta})) \cdot ((\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) + \mathbf{X}(\widehat{\beta}^* - \tilde{\beta})) \\ &= (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) \cdot (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) + (\widehat{\beta}^* - \tilde{\beta}) \mathbf{X}^\top \mathbf{X} (\widehat{\beta}^* - \tilde{\beta}) + \underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) \mathbf{X} (\widehat{\beta}^* - \tilde{\beta})}_a + \underbrace{(\widehat{\beta}^* - \tilde{\beta}) \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*)}_{a'}. \quad (50)\end{aligned}$$

Veamos que las expresiones “a” y “a’” (que son iguales) son $\mathbf{0}$...

Multiplicando la expresión (46) del estimador restringido $\widehat{\beta}^*$ por $\mathbf{X}^\top \mathbf{X}$ tenemos

$$\begin{aligned}\mathbf{X}^\top \mathbf{X} \widehat{\beta}^* &= \mathbf{X}^\top \mathbf{X} \widehat{\beta} - \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r) \\ &= \mathbf{X}^\top \mathbf{Y} - \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r)\end{aligned}$$

donde $\mathbf{X}^\top \mathbf{X} \widehat{\beta} = \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{Y}$ por tanto, despejando la última parte tenemos

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) = \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r).$$

Ahora multiplicando ambos lados expresión de arriba por $(\widehat{\beta}^* - \tilde{\beta})$ llegamos a

$$(\widehat{\beta}^* - \tilde{\beta}) \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) = (\widehat{\beta}^* - \tilde{\beta}) \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r)$$

pero como las estimaciones $\widehat{\beta}^*$ y $\tilde{\beta}$ verifican la restricción, entonces $\mathbf{R}(\widehat{\beta}^* - \tilde{\beta}) = \mathbf{R}\widehat{\beta}^* - \mathbf{R}\tilde{\beta} = \mathbf{I}r - \mathbf{I}r = \mathbf{0}$, y por tanto, también $(\widehat{\beta}^* - \tilde{\beta}) \mathbf{R}^\top = \mathbf{0}$, sustituyendo tenemos

$$(\widehat{\beta}^* - \tilde{\beta}) \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) = \mathbf{0} \mathbf{R}^\top [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\beta} - \mathbf{I}r) = \mathbf{0}.$$

Así que finalmente la suma residual $SR(\tilde{\beta}) = \tilde{\mathbf{e}} \cdot \tilde{\mathbf{e}}$ de la ecuación (50) se reduce a

$$\begin{aligned}\tilde{\mathbf{e}} \cdot \tilde{\mathbf{e}} &= (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*) + (\widehat{\beta}^* - \tilde{\beta}) \mathbf{X}^\top \mathbf{X} (\widehat{\beta}^* - \tilde{\beta}) \\ &= \widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^* + (\widehat{\beta}^* - \tilde{\beta}) \mathbf{X}^\top \mathbf{X} (\widehat{\beta}^* - \tilde{\beta}) && \text{donde } \widehat{\mathbf{e}}^* = (\mathbf{Y} - \mathbf{X}\widehat{\beta}^*); \quad (51)\end{aligned}$$

expresión que es mínima cuando $(\widehat{\beta}^* - \tilde{\beta}) = \mathbf{0}$, es decir cuando el estimador $\tilde{\beta}$ (un estimador que cumple la restricción $\mathbf{R}\tilde{\beta} = \mathbf{I}r$) es igual a $\widehat{\beta}^*$ (el estimador de Mínimos Cuadrados Restringido). **(Fin de la demostración).**

En la mayoría de libros de texto encontrará una demostración alternativa derivando el Lagrangiano de un problema de minimización con restricciones lineales (cf., Novales (1993, pag. 132)).

24.2 Demostración de la Proposición 22.3 (varianza del estimador restringido)

Resultado preliminar Para la demostración usaremos que la matriz

$$\mathbf{P}_R = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}. \quad (52)$$

verifica que

$$(\mathbf{I} - \mathbf{P}_R)(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{P}_R^\top = \mathbf{0} \quad (53)$$

ya que

$$\begin{aligned} (\mathbf{I} - \mathbf{P}_R)(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{P}_R^\top &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{P}_R^\top - \mathbf{P}_R(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{P}_R^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \mathbf{P}_R(\mathbf{X}^\top \mathbf{X})^{-1} - \mathbf{P}_R(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{0}. \end{aligned}$$

Demostración de la proposición.

Sea $\mathbf{Id} = (\mathbf{R}\mathbf{I}\beta - \mathbf{I}\mathbf{r})$ y denotemos el vector aleatorio $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{Id}$ con \mathbf{F} ; de (45):

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^* &= \widehat{\boldsymbol{\beta}} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{R}\mathbf{I}\beta - \mathbf{Id}) && \text{sustituyendo } \mathbf{I}\mathbf{r} \text{ por } \mathbf{R}\mathbf{I}\beta - \mathbf{Id} \\ \widehat{\boldsymbol{\beta}}^* - \mathbf{I}\beta &= \widehat{\boldsymbol{\beta}} - \mathbf{I}\beta - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{R}\mathbf{I}\beta) - \mathbf{F} && \text{restando a ambos lados } \mathbf{I}\beta \\ &= \left[\mathbf{I} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R} \right] (\widehat{\boldsymbol{\beta}} - \mathbf{I}\beta) - \mathbf{F} && \text{sacando el factor común } (\widehat{\boldsymbol{\beta}} - \mathbf{I}\beta) \end{aligned}$$

ahora sustituyendo \mathbf{P}_R (Ecuación 52) tenemos $\widehat{\boldsymbol{\beta}}^* - \mathbf{I}\beta = (\mathbf{I} - \mathbf{P}_R)(\widehat{\boldsymbol{\beta}} - \mathbf{I}\beta) - \mathbf{F}$. Así,

$$\text{Var}(\widehat{\boldsymbol{\beta}}^* | \mathbf{X}) = \text{Var}(\widehat{\boldsymbol{\beta}}^* - \mathbf{I}\beta | \mathbf{X}) = \text{Var}((\mathbf{I} - \mathbf{P}_R)(\widehat{\boldsymbol{\beta}} - \mathbf{I}\beta) - \mathbf{F} | \mathbf{X}) = \text{Var}((\mathbf{I} - \mathbf{P}_R)(\widehat{\boldsymbol{\beta}} - \mathbf{I}\beta) | \mathbf{X});$$

pues $\mathbf{F} \in \mathcal{L}(\mathbf{X})$. Por tanto,

$$\begin{aligned} \text{Var}(\widehat{\boldsymbol{\beta}}^* | \mathbf{X}) &= (\mathbf{I} - \mathbf{P}_R) \text{Var}(\widehat{\boldsymbol{\beta}} - \mathbf{I}\beta | \mathbf{X}) (\mathbf{I} - \mathbf{P}_R)^\top \\ &= \sigma^2 (\mathbf{I} - \mathbf{P}_R)(\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{I} - \mathbf{P}_R)^\top \\ &= \sigma^2 (\mathbf{I} - \mathbf{P}_R)(\mathbf{X}^\top \mathbf{X})^{-1} - \underbrace{\sigma^2 (\mathbf{I} - \mathbf{P}_R)(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{P}_R^\top}_{= \mathbf{0} \text{ (Ecuación (53))}} \\ &= \sigma^2 \left[(\mathbf{X}^\top \mathbf{X})^{-1} - \mathbf{P}_R(\mathbf{X}^\top \mathbf{X})^{-1} \right] \\ &= \sigma^2 \left[(\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \right] \\ &= \text{Var}(\widehat{\boldsymbol{\beta}} | \mathbf{X}) - \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \end{aligned}$$

donde tanto $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1}$ como las matrices $\text{Var}(\widehat{\boldsymbol{\beta}}^* | \mathbf{X})$ y $\text{Var}(\widehat{\boldsymbol{\beta}} | \mathbf{X})$ son definidas positivas, y por tanto el estimador restringido tiene menor varianza que el estimador MCO sin restringir:

$$\text{Var}(\widehat{\boldsymbol{\beta}} | \mathbf{X}) \geq \text{Var}(\widehat{\boldsymbol{\beta}}^* | \mathbf{X}).$$

(Fin de la demostración).

Problemas de la Lección 9

Estimación restringida

(L-9) PROBLEMA 1. Sabiendo que $R^2 = 1 - \frac{SRC}{STC}$ (i.e., el caso general) demuestre que el estadístico (47) se puede expresar como

$$F = \frac{N - k}{r} \cdot \frac{R^2 - R^{2*}}{1 - R^2}$$

Pista. En la estimación restringida $SRC^* = (1 - R^{2*}) \cdot STC$ y en la estimación sin restringir $SRC = (1 - R^2) \cdot STC$, donde STC es común a ambas (¿por qué?).

(L-9) PROBLEMA 2. Para el caso de Modelo Lineal Simple, encuentre la expresión de la estimación MCRL para los siguientes casos

- (a) cuando la restricción es $\beta_1 = 0$
- (b) cuando la restricción es $\beta_2 = 0$

(L-9) PROBLEMA 3. [(Ejemplo 4.6 de Novales, 1993, pp. 137)] Consideremos el modelo

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U.$$

donde las variables se hayan en diferencias respecto a la media, y las sumas de los productos cruzados calculados a partir de 103 observaciones muestrales son:

	y	x_1	x_2	x_3
y	10	3	-3	-2
x_1	3	10	4	-2
x_2	-3	4	8	0
x_3	-2	-2	0	6

Contrastar la hipótesis lineal $H_0: \beta_1 + \beta_2 = 0$ frente a $H_1: \beta_1 + \beta_2 \neq 0$.

Fin de los Problemas de la Lección 9

Enlace a algunas **prácticas de la Lección 9**

Fin de la lección

Part IV

Interpretación

LECCIÓN 10: Interpretación de coeficientes en modelos con logaritmos

Bibliografía:

Básica: Ramanathan (2002), Wooldridge (2006)

Complementaria: Novales (1993), Novales (1997)

25 Interpretación de los parámetros en un modelo de regresión

25.1 Interpretación “*Ceteris páribus*”

Uno de los objetivos del análisis empírico es determinar si la variación de algunas magnitudes está relacionado con la variación de otras, y de qué modo lo están... ¿Afectan los tipos de interés a la tasa de variación de los precios? ¿Está relacionado el PIB per cápita con el nivel de emisiones de CO₂ a la atmósfera? ¿Y el número de horas de estudio con las calificaciones finales del estudiante?

En este sentido, la interpretación del coeficiente o parámetro que acompaña uno de los regresores del modelo siempre se realiza “*Ceteris páribus*”, es decir, manteniendo el resto de factores fijos. Por ejemplo, si queremos estudiar qué relación existe entre los años de estudio y el salario percibido por un trabajador, lo querremos hacer controlando el efecto de otras magnitudes tales como la experiencia o la habilidad del trabajador. Es decir, de algún modo querremos aislar el efecto de la formación del trabajador respecto de otras características que también pudieran afectar la determinación del salario, tales como su experiencia o de su habilidad innata (medida, por ejemplo, por su coeficiente intelectual “IQ”). El modo de hacerlo es estimar la esperanza condicional $E(Y|X, Z)$ donde Y es el salario, X los años de formación, y Z es un sistema que incluye las variables “años de experiencia” e “IQ”. Bajo los supuestos clásicos que ya hemos visto, la esperanza condicional será una combinación lineal de los regresores. Así, interpretaremos el parámetro que acompaña a cada regresor como el efecto, “*Ceteris páribus*”, del correspondiente regresor sobre el regresando. En particular, interpretaremos el parámetro que acompaña a X como el efecto que tiene sobre el salario esperado el estudio de un año adicional, una vez descontando los efectos que pudieran tener los años de experiencia o el coeficiente intelectual del trabajador.

Otro ejemplo: si estimamos un modelo que relaciona el precio de una vivienda con su tamaño, su número de dormitorios y su número de cuartos de baño; el parámetro que acompaña al regresor “número de cuartos de baño” se interpretará como el efecto, “*Ceteris páribus*”, que dicho número tiene sobre el precio de la vivienda; es decir, si no cambiara ni el tamaño ni el número de dormitorios de la vivienda, qué variación en el precio esperado de la vivienda supondría la existencia de un cuarto de baño adicional.

Recuerde que la interpretación de cada parámetro siempre es “*Ceteris páribus*”, es decir, “si no cambia el valor de las otras variables explicativas”. Fíjese que he dicho otras variables explicativas en lugar de otros regresores, el motivo es que...

25.2 Regresor y variable explicativa no son siempre lo mismo

... y tampoco regresando y “variable explicada”. Veámoslo.

En el modelo $Y = X\beta + U$, la variable de la izquierda se denomina regresando y las variables de la lista X se denominan regresores. Ya sabemos que bajo los supuestos clásicos $E(Y|X) = X\beta$. Consecuentemente β_j es el efecto marginal de X_j sobre la esperanza condicional (cuando X_j es continua) o el incremento de la esperanza condicional cuando X_j es discreta y aumenta en 1 unidad.

Así, si Y es el precio de una vivienda en euros y X es una sistema con las variables aleatorias $X_1 = 1$ con X_2 (superficie), X_3 (n^o de dormitorios) y X_4 (n^o de cuartos de baño), entonces β_2 es la pendiente (el efecto marginal de la superficie sobre el precio esperado), pero β_3 es la variación (en euros) del precio esperado cuando hay un dormitorio adicional. En un ejemplo como este, la variable explicada (precio) es el regresando Y y los regresores X_2 , X_3 y X_4 son las variables explicativas.

Sin embargo hemos visto modelos con otras especificaciones. En la Lección 3 (final del Ejemplo 5 en la página 35) ajustamos un modelo que expresado con variables aleatorias sería $Y = \beta_1 1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3 + U$ donde Y es el peso de un niño y X es la edad. En este modelo la variable explicativa es la edad, pero los regresores son, además de la edad, la edad al cuadrado y la edad al cubo. Es decir, en ocasiones los regresores pueden ser transformaciones de las variables explicativas. En este modelo el efecto marginal de la edad sobre el peso esperado es la derivada de $\beta_1 1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$; y por tanto dicho efecto marginal depende de la edad. Consecuentemente la interpretación de los parámetros es diferente a la del modelo anterior.

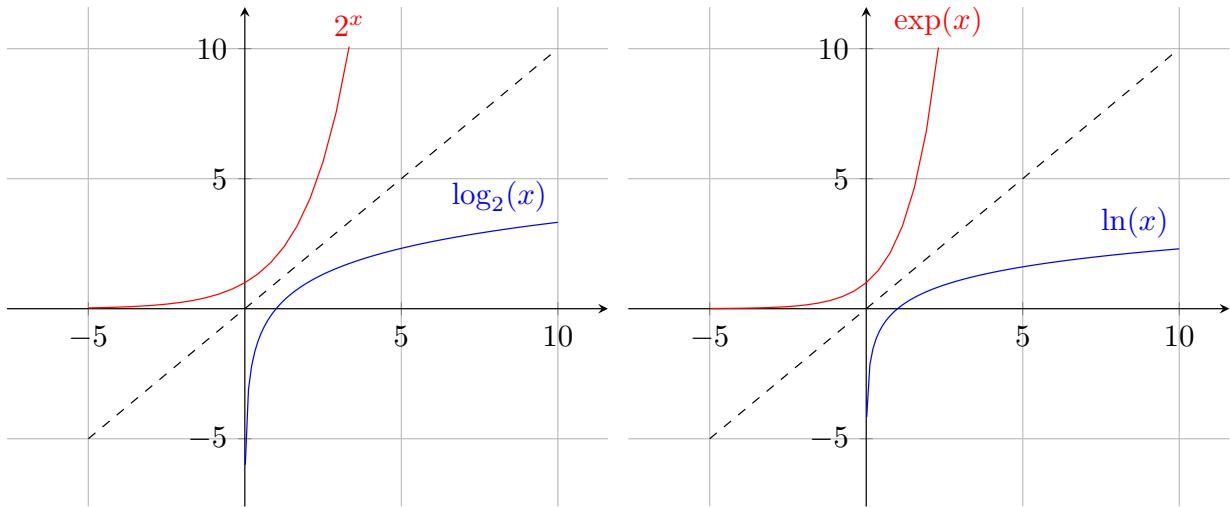
Por otra parte, también el regresando puede ser una transformación de la variable explicada. En el Ejemplo 2 en la página 11 (Lección 1), el modelo original $\mathbf{Y} = \exp(\beta_1 \mathbf{1} + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4) + \mathbf{U}$, (donde \mathbf{Y} es el salario, \mathbf{X}_2 son los años de educación, \mathbf{X}_3 los años de antigüedad en la empresa y \mathbf{X}_4 los años de experiencia) no cumple el primer supuesto. Pero transformando logarítmicamente el modelo tenemos

$$\ln \mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \mathbf{U}.$$

Aunque la esperanza condicional en este modelo transformado ya es lineal en los parámetros, ahora resulta que el regresando es el logaritmo de la variable explicada. Así, aunque los parámetros sean los efectos marginales sobre el regresando, dicho regresando ya no es la variable de interés; y como analistas estaremos interesados en saber cómo afectan los años de formación al salario del trabajador (y no a su logaritmo).

Esta lección trata sobre la interpretación de los parámetros en algunos modelos donde el regresando, el regresor, o ambos son transformaciones de las variables de interés.

26 Función exponencial, función logaritmo y elasticidad



Función exponencial. La función $f(x) = a^x$ (con $a > 0$) se denomina “función exponencial con base a ”. La base de uso más común es el número irracional e :

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = 2.71828182845905\dots$$

Cuando la base es el número e , la función exponencial $f(x) = e^x$ también se denota con $f(x) = \exp(x)$, y se suele denominar sencillamente “exponencial de x ” (sin indicar explícitamente que la base es el número e). Esta función es la única que es igual a su propia función derivada ($\frac{\partial e^x}{\partial x} = e^x$) y que satisface que $f(0) = 1$.

Función logaritmo. La inversa de la función exponencial en base a es la función logaritmo en base a .

$$f(x) = a^x \quad \longrightarrow \quad f^{-1}(x) = \log_a(x).$$

Es decir, el logaritmo de x con base a , $\log_a(x)$, es la potencia a la que se debe elevar a para obtener x :

$$a^{(\log_a(x))} = x.$$

Por ejemplo $\log_2(16) = 4$, pues $2^4 = 16$.

Nótese que para cualquier base $a > 0$, el logaritmo de 1 es cero, pues $a^0 = 1$; y además, $\log_b(b) = 1$ pues $b^1 = b$.

Cuando la base es el número e , esta función se llama *logaritmo natural* o *logaritmo neperiano*, y se escribe $\ln x$. Así, la inversa de $\exp(x)$ es $\ln(x)$.

$$f(x) = e^x = \exp(x) \quad \longrightarrow \quad f^{-1}(x) = \ln(x) \quad \longrightarrow \quad \exp(\ln(x)) = x = \ln(\exp(x)).$$

Recordatorio de algunas propiedades

1. La función exponencial y la función logaritmo son monótonas crecientes, es decir, si $a < b$ entonces $f(a) < f(b)$.
2. El logaritmo del producto es la suma de los logaritmos

$$\log_c(a \cdot b) = \log_c(a) + \log_c(b), \quad \text{para } a, b > 0 \text{ y } c \neq 1.$$

3. El logaritmo del cociente es la diferencia de los logaritmos

$$\log_c\left(\frac{a}{b}\right) = \log_c(a) - \log_c(b), \quad \text{para } a, b > 0 \text{ y } c \neq 1.$$

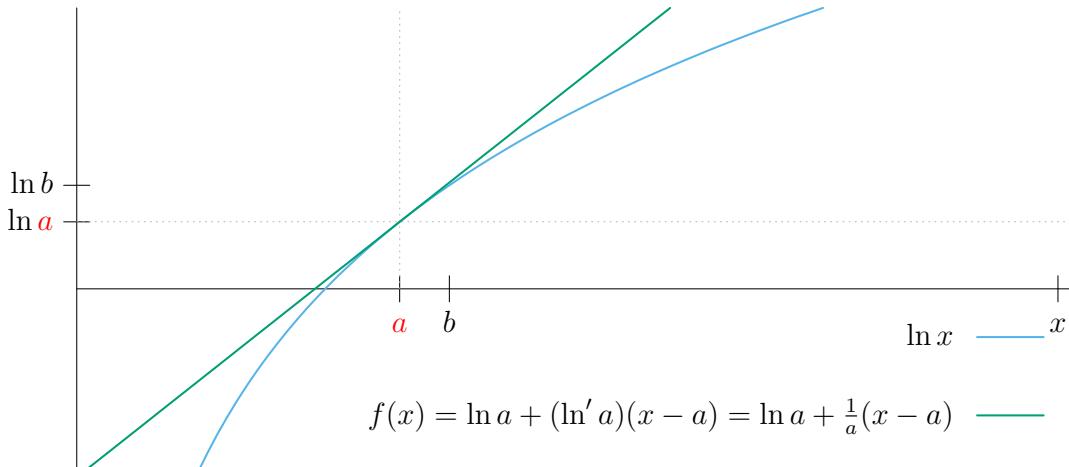
De donde se deduce que $\log_c\left(\frac{1}{b}\right) = \log_c(1) - \log_c(b) = 0 - \log_c(b) = -\log_c(b)$.

4. $\log_c(a^x) = x \log_c(a)$, y consecuentemente $\log_c(\sqrt[x]{a}) = \frac{\log_c(a)}{x}$
5. Como caso particular de lo anterior, $\ln(a^x) = x \ln(a)$; y tomando la trasformación exponencial se deduce que $a^x = \exp(x \ln a)$.
6. $a^x \cdot a^y = a^{(x+y)}$ y $(a^x)^y = a^{x \cdot y}$.
7. ¡La función exponencial de x es igual a su función derivada! Así, si $y = \exp(x)$, entonces $\frac{\partial y}{\partial x} = \exp(x)$.
8. La derivada de $\exp(a^x)$ es $a \exp(a^x)$.
9. La derivada de $\ln(x)$ es $\frac{1}{x}$.
10. De 5, 8 y 7 se deduce que la derivada de a^x es $\ln(a) \cdot a^x$.

Logaritmo neperiano (ln) y cambios relativos De la derivada de la función logaritmo neperiano tenemos:

$$\frac{\partial \ln x}{\partial x} = \frac{1}{x} \Rightarrow \partial \ln x = \frac{\partial x}{x} = \text{cambio relativo (infinitesimal) de } x$$

Recta $f(x)$ tangente en $x = a$ a la función logaritmo neperiano $\ln x$



Puesto que la recta tangente a la función logaritmo neperiano evaluada en a es $f(x) = \ln(a) + \frac{1}{a}(x - a)$ tenemos que

$$f(b) = \ln(a) + \frac{1}{a}(b - a).$$

La recta tangente en el punto a es una aproximación a la función evaluada en dicho punto. Por tanto, para incrementos muy pequeños, el incremento de la recta tangente es casi igual que el incremento de la función en dicho punto; por lo

que logramos la siguiente *aproximación* cuando el incremento $\Delta x = b - a$ es pequeño:

$$\Delta \ln x = \ln(b) - \ln a \approx \Delta f(x) = \ln(a) + \frac{1}{a}(b-a) - \ln(a) = \frac{1}{a}\Delta x.$$

(Lección 10)

[T-1]

Elasticidad

De la derivada de la función logaritmo neperiano tenemos:

$$\frac{\partial \ln z}{\partial z} = \frac{1}{z} \Rightarrow \partial \ln z = \frac{\partial z}{z} = \text{cambio relativo (infinitesimal) de } z$$

La elasticidad η de y respecto a x se define cómo:

$$\eta = \frac{\text{cambio relativo infinitesimal de } y}{\text{cambio relativo infinitesimal de } x} = \frac{\partial \ln y}{\partial \ln x} = \frac{\partial y/y}{\partial x/x} = \frac{x}{y} \cdot \frac{\partial y}{\partial x}.$$

Relacionemos esto con distintas formas funcionales de los modelos

(¡todos lineales en los parámetros!).

F118

Como $\partial \ln y = \frac{\partial x}{x}$ = cambio relativo de x ; tenemos que la elasticidad de y respecto a x también es

$$\eta = \frac{\partial \ln y}{\partial \ln x}.$$

Recuerde que $\partial \ln f(z) = \frac{1}{f(z)} \frac{\partial f(z)}{\partial z}$; así, si derivamos un modelo de la forma $\ln(y) = \beta \ln(x)$ deducimos que

$$\frac{\partial \ln(y)}{\partial x} = \beta \frac{\partial \ln(x)}{\partial x}$$

y sustituyendo $\partial \ln(y)$ por $\frac{\partial y}{y}$ y $\partial \ln(x)$ por $\frac{\partial x}{x}$ tenemos

$$\frac{1}{\partial x} \frac{\partial y}{y} = \beta \frac{1}{\partial x} \frac{\partial x}{x} = \beta \frac{1}{x}.$$

Despejando β concluimos que β es igual a la elasticidad:

$$\beta = \frac{x}{y} \frac{\partial y}{\partial x}$$

y despejando $\frac{\partial y}{y}$ deducimos que

$$\frac{\partial y}{y} = \beta \frac{\partial x}{x};$$

es decir, que el incremento relativo (infinitesimal) de y es β por el incremento relativo (infinitesimal) de x .

Operando de manera similar se logra completar la siguiente tabla.

Efectos marginales y elasticidades para distintas funciones lineales en los parámetros

Nombre	Forma Funcional	Efecto Marginal: $\frac{dy}{dx}$	Elasticidad: $\frac{x}{y} \frac{dy}{dx}$
Lineal	$y = \alpha + \beta x$	β	$\beta x/y$
Lin-Log	$y = \alpha + \beta \ln x$	β/x	β/y
Reciproco	$y = \alpha + \beta 1/x$	$-\beta/x^2$	$-\beta/(xy)$
Cuadrático	$y = \alpha + \beta x + \gamma x^2$	$\beta + 2\gamma x$	$(\beta + 2\gamma x)x/y$
Interacción	$y = \alpha + \beta x + \gamma xy$	$\beta + \gamma z$	$(\beta + \gamma z)x/y$
Log-Lin	$\ln y = \alpha + \beta x$	βy	βx
Log-Reciproco	$\ln y = \alpha + \beta(1/x)$	$-\beta y/x^2$	$-\beta/x$
Log-Cuadrático	$\ln y = \alpha + \beta x + \gamma x^2$	$y(\beta + 2\gamma x)$	$x(\beta + 2\gamma x)$
Log-Log	$\ln y = \alpha + \beta \ln x$	$\beta y/x$	β
Logístico	$\ln \left[\frac{y}{1-y} \right] y = \alpha + \beta x$	$\beta y(1-y)$	$\beta(1-y)x$

Table 8: Efectos marginales y elasticidades para distintas formas funcionales

27 Interpretación de los coeficientes de una regresión lineal cuando el modelo original no era lineal y se transformó logarítmicamente para linealizarlo

(Lección 10) T-2 Interpretación de coeficientes en modelos con logs	
Modelo	Interpretación
$y = \alpha + \beta x$	$\beta = \frac{\partial y}{\partial x}$ Cambio esperado en nivel de y si x aumenta una unidad
$\ln(y) = \alpha + \beta \ln(x)$	$\beta = \frac{x}{y} \frac{\partial y}{\partial x}$ (Aprox.) Cambio <u>porcentual</u> esperado de y si x aumenta un uno por ciento (en tanto por uno, i.e., 0.01)
$\ln(y) = \alpha + \beta x$	$\beta = \frac{1}{y} \frac{\partial y}{\partial x}$ (Aprox.) Cambio relativo esperado de y (en tanto por uno) si x aumenta una unidad
$y = \alpha + \beta \ln(x)$	$\beta = x \frac{\partial y}{\partial x}$ (Aprox.) Cambio esperado en el nivel de y si x aumenta un uno por ciento (en tanto por uno)

(derivando respecto a x , sustituyendo $\partial \ln z$ por $\frac{\partial z}{z}$ y despejando)

F120

Ejemplo 19. Función de consumo (lin-lin):

$$CON = \beta_1 \mathbf{1} + \beta_2 RD + \mathbf{U}$$

donde CON y RD son el consumo y la renta disponible respectivamente, y \mathbf{U} son otros factores que afectan al consumo distintos a la renta disponible (activos financieros, estado de ánimo, etc.).

Ejemplo 20. Ecuación de salarios (log-lin):

$$SALAR = e^{(\beta_1 \mathbf{1} + \beta_2 EDUC + \beta_3 ANTIG + \beta_4 EXPER + \mathbf{U})};$$

donde $SALAR$ es el salario, $EDUC$ son los años de educación, $ANTIG$ los años de antigüedad en la empresa, y $EXPER$ los años de experiencia en el sector.

Al tomar logaritmos tenemos un nuevo modelo para $\ln(SALAR)$ que es lineal en los parámetros:

$$\ln(SALAR) = \beta_1 \mathbb{1} + \beta_2 EDUC + \beta_3 ANTIG + \beta_4 EXPER + U$$

Ejemplo 21. Precio de la vivienda (lin-log):

$$PRICE = \beta_1 \mathbb{1} + \beta_2 \ln SQFT + U.$$

Ejemplo 22. Función de producción Cobb-Douglas (log-log):

$$Q = c K^{\beta_2} L^{\beta_3} \nu;$$

donde Q es la producción, K es el capital empleado; L el trabajo empleado. Supongamos, además, que hay un efecto aleatorio adicional ν debido a otras causas o factores.

Tomando logaritmos tenemos

$$\ln Q = \beta_1 \mathbb{1} + \beta_2 \ln K + \beta_3 \ln L + U,$$

donde $\beta_1 = \ln c$, y $U = \ln \nu$. (es decir, $\nu = e^U$.)

27.1 Relaciones lineales en las variables

(Lección 10) Ejercicio en clase. N-1.

Código: POE2-4.inp

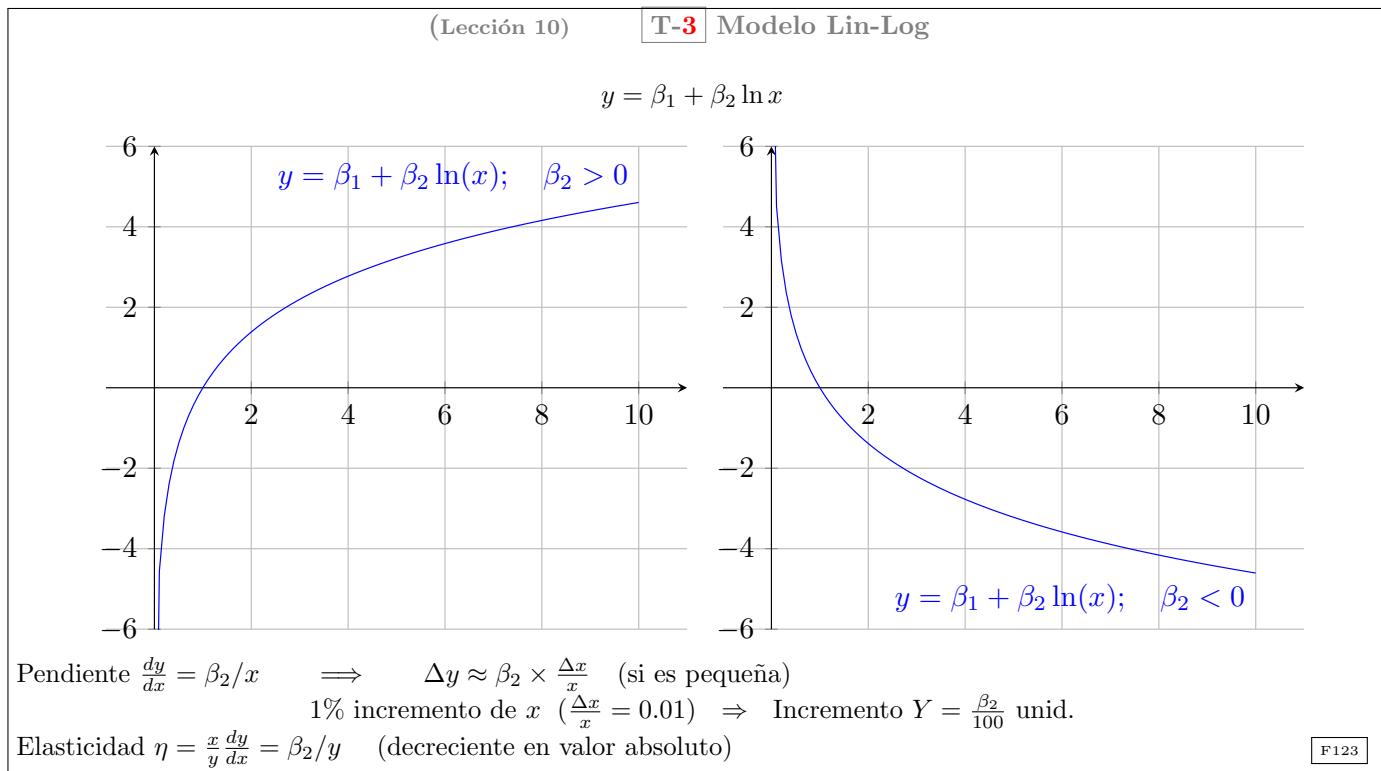
Cargue los datos `food.gdt` del libro POE, sobre los gastos en alimentación `food_exp` de las familias y la renta disponible `income`.

- (a) Ajuste por MCO el gasto en comida en función de la renta disponible
- (b) Observe los estadísticos principales de ambas variables
- (c) Grafique un diagrama de dispersión del gasto sobre la renta
- (d) Muestre los valores de ambas variables
- (e) Calcule la elasticidad de la demanda de alimentos respecto de la renta en el valor medio muestral de la renta, donde

$$\left(\frac{\text{variación \% de } x}{\text{variación \% de } y} \right) \approx \text{elasticidad} = \frac{\partial y / y}{\partial x / x} = \frac{\partial y}{\partial x} \frac{x}{y} \approx \widehat{\beta}_2 \frac{m_x}{m_y}$$

- (f) ¿Qué gasto prevé este modelo para una familia cuya renta asciende a 20?
- (g) Realice un contraste de normalidad para los residuos ¿Puede rechazar que la distribución es normal?
- (h) Grafique los residuos de la regresión ¿Le parece que la varianza de los residuos es independiente de la renta? ¿Es creíble que se cumple el supuesto de homocedasticidad en este modelo?

27.2 Relaciones Lin-Log



(Lección 10) Ejercicio en clase. N-2.

»Código: RamanathanEX6-1.inp

Precio de casas unifamiliares Use data4-1.gdt.

- (a) Estime por MCO: $PRICE = \beta_1 \mathbb{1} + \beta_2 SQFT + U$; y añádalo a la tabla de modelos.
- (b) Estime después el siguiente modelo Lin-Log

$$PRICE = \beta_1 \mathbb{1} + \beta_2 \ln SQFT + \beta_3 \ln BEDRMS + \beta_4 \ln BATH + U;$$

- (c) Decida si es necesario quitar alguna variable del modelo. Opere secuencialmente (añadiendo a la tabla de modelos aquellos que le parezcan interesantes) hasta quedarse con un modelo definitivo.
- (d) Compare los resultados de los distintos modelos ajustados. ¿Hay grandes diferencias? ¿Son comparables los ajustes?
- (e) Calcule las elasticidades del modelo lineal y del siguiente modelo Lin-Log:

$$PRICE = \beta_1 \mathbb{1} + \beta_2 \ln SQFT + U;$$

para casas con superficies de 1500, 2000 y 2500 pies al cuadrado respectivamente. ¿Qué diferencias encuentra? También lo puede hacer con el modelo que incluye información sobre los dormitorios (pero es ligeramente más trabajoso).

- (f) ¿Cuanto aumenta el precio de las casa por un aumento del 1% de su superficie (nótese que este aumento es independiente del tamaño de la casa (lin-log)).

27.3 Relaciones semi-logarítmicas (Log-lineal)

(Lección 10)

T-4 Modelo en semi-logaritmos (Log-Lin)

Ejemplo 23. Modelo de crecimiento constante: Suponga que la variable P crece a una tasa constante g :

$$P_t = P_{t-1} \cdot (1 + g).$$

Mediante sustituciones sucesivas, llegamos a

$$P_t = P_0(1 + g)^t.$$

Este modelo se puede linealizar tomando logaritmos:

$$\underbrace{\ln P_t}_Y = \underbrace{\beta_1}_{\beta_1} + \underbrace{\ln(1 + g)}_{\beta_2} \cdot \underbrace{t}_X \Rightarrow g = \exp(\beta_2) - 1 \quad (54)$$

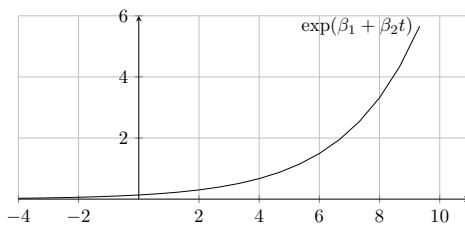
F124

Consideré el modelo

$$\ln P_t = \beta_1 \mathbb{1} + \beta_2 \cdot t \mathbb{1} + U_t;$$

tomando la exponencial obtenemos otro modelo para P_t :

$$P_t = \exp(\beta_1 \mathbb{1} + \beta_2 \cdot t \mathbb{1} + U_t).$$



De (54) sabemos que $\widehat{\beta}_1 = \widehat{\ln P_0}$; y $\widehat{\beta}_2 = \ln(1 + \widehat{g})$; es decir:

$$\widehat{P}_0 = \exp(\widehat{\beta}_1); \quad \text{y} \quad \widehat{g} = \exp(\widehat{\beta}_2) - 1$$

Tomando esperanzas: $E(P_t) = e^{\beta_1 + \beta_2 t} E(e^{U_t})$, pero $E(e^{U_t}) \neq 1$

Esperanza de una distribución lognormal: Si $Z \sim N(\mu, \sigma^2)$, entonces $\exp Z$ tiene distribución lognormal y

$$E(\exp Z) = \exp\left(\mu + \frac{\sigma^2}{2}\right);$$

así, puesto que $U \sim N(0, \sigma^2)$

$$E(\exp U) = \exp\left(\frac{\sigma^2}{2}\right);$$

resultado que si se quiere ser preciso ha de tenerse en cuenta del siguiente modo

Si $\ln Y = X\beta + U$, tomando la esperanza de $Y = \exp(X\beta + U)$,

$$E(Y) = \exp(X\beta) \cdot E(\exp(U)),$$

donde $E(\exp(U)) = \exp\left(\frac{\sigma^2}{2}\right)$.

Por tanto como predictor de Y se debe usar,

$$\widehat{Y} = \exp\left(\widehat{X\beta} + \frac{\widehat{s^2}}{2}\right) = \exp\left(\widehat{\ln Y} + \frac{\widehat{s^2}}{2}\right)$$

que es un estimador consistente de $E(\hat{Y})$.

Por otra parte, un estimador consistente de la tasa de crecimiento g en el modelo

$$\ln P_t = \beta_1 \mathbf{1} + \beta_2 \cdot t \mathbf{1} + U_t;$$

es

$$\tilde{g} = \exp\left(\widehat{\beta}_2 + \frac{1}{2}\widehat{\text{Var}}(\widehat{\beta}_2)\right) - 1$$

(Lección 10)

T-5 Ejemplo de modelo en semi-logaritmos (Log-Lin)

Si el retorno de un año adicional de estudios es g , entonces, $w_1 = (1+g)w_0$, y $w_2 = (1+g)^2 w_0$, En general

$$w_t = (1+g)^t w_0.$$

Tomando logs: $\ln w_t = \ln w_0 + \ln(1+g) \cdot t = \beta_1 + \beta_2 \cdot t$.

Ejemplo 24. Supongamos el siguiente modelo no-lineal en los parámetros

$$SALAR = e^{(\beta_1 \mathbf{1} + \beta_2 EDUC + \beta_3 ANTIG + \beta_4 EXPER + U)};$$

donde $SALAR$ es el salario del trabajador, $EDUC$ son sus años de educación, $ANTIG$ sus años de antigüedad en la empresa, y $EXPER$ sus años de experiencia en el sector de la empresa.

Tomando logaritmos → modelo para $\ln(SALAR)$

$$\ln(SALAR) = \beta_1 \mathbf{1} + \beta_2 EDUC + \beta_3 ANTIG + \beta_4 EXPER + U$$

Si $\beta_2 = .03$; cada año educ → incremen. esperado (aprox.) salario 3%

(mejor $g = \exp(\beta_2) - 1$) → $g = \exp(0.03) - 1 = 0.030455$.

F125

De manera similar, en

$$P_t = \exp(\beta_1 \mathbf{1} + \beta_2 \cdot t \mathbf{1} + U_t),$$

y si se cumplen los cinco supuestos de modelo clásico de regresión lineal, resulta que $\widehat{\beta}_2 \sim N(\beta_2, \text{Var}(\widehat{\beta}_2))$ y por tanto

$$E(\exp(\widehat{\beta}_2)) = \exp\left(\beta_2 \mathbf{1} + \text{Var}(\widehat{\beta}_2)/2\right);$$

entonces $\exp\left[\widehat{\beta}_2 - \frac{\widehat{\text{Var}}(\widehat{\beta}_2)}{2}\right] - 1$ es un estimador insesgado del valor esperado de la tasa de crecimiento g .

(Lección 10) Ejercicio en clase. N-3.

■■■ Código: RamanathanEX6-5.inp

Modelo para los salarios. Abra el conjunto de datos **data6-4.gdt**, del libro de Ramanathan, con datos del salarios mensuales (*wage*), años de educación (*educ*) y de experiencia (*exper*), y la edad (*age*) de 49 trabajadores.

(a) Estime el modelo

$$\begin{aligned} \ln W &= \beta_1 \mathbf{1} + \beta_2 \cdot EDUC + \beta_3 \cdot EDUC^2 + \beta_4 \cdot AGE + \beta_5 \cdot AGE^2 \\ &\quad + \beta_6 \cdot EXPER + \beta_7 \cdot EXPER^2 + U \end{aligned}$$

(b) Vaya eliminando variables no significativas hasta obtener un modelo final.

(c) ¿Qué efecto estimado tiene un año adicional de experiencia?

(d) Recordando que

$$\hat{W} = \exp\left(X\hat{\beta} + \hat{s}^2/2\right),$$

calcule los salarios estimados por el modelo y compárelos con los salarios de la muestra. Con el diagrama de dispersión de salarios observados y ajustados podrá comprobar que este modelo no funciona muy bien.

(e) Pese a ello calcule el efecto estimado que tiene un año adicional de educación en el salario de trabajadores con 1

y 7 años de formación respectivamente.

(Lección 10) **T-6** Comparación de coeficientes de determinación entre modelos

R^2 de modelos Lin-Lin y Log-Lin no son comparables
(distinto regresando)

- Una forma de intentar comparar ajustes es calcular el cuadrado de la correlación entre \mathbf{y} y $\hat{\mathbf{y}}$; donde

$$\hat{Y} = \exp\left(\ln \bar{Y} + \sigma^2/2\right)$$

- O calcular los estadísticos de selección empleando la suma de errores al cuadrado y la varianza estimada:

$$SRC = \sum(Y - \hat{Y})^2; \quad \widetilde{\sigma^2} = \frac{SRC}{n - k}$$

F126

(Lección 10) Ejercicio en clase. N-4.

☞ Código: RamanathanEX6-6.inp

Modelo para los salarios. Abra el conjunto de datos `data6-4.gdt`, del libro de Ramanathan, con datos del salarios mensuales (`wage`), así como años de educación (`educ`), años de experiencia (`exper`) y edad (`age`) de 49 trabajadores.

(a) Estime los modelos

$$W = \beta_1 \mathbb{1} + \beta_2 \cdot EDUC^2 + \beta_3 \cdot EXPER + U$$

$$\ln W = \beta_1 \mathbb{1} + \beta_2 \cdot EDUC^2 + \beta_3 \cdot EXPER + U$$

Aunque los R^2 parecen semejantes, no son comparables.

- (b) Guarde los salarios estimados por el segundo modelo, así como los errores y la varianza estimada de los errores.
- (c) Calcule el cuadrado de la correlación entre los salarios observados y los estimados (o predichos). ¿Qué modelo presenta un mejor ajuste? ¿El primero o el segundo?
- (d) Cargando la función `criteria.gfn`, calcule los criterios de selección de modelo (mire el guión adjunto). A la luz de los resultados, ¿qué modelo parece preferible?

27.4 Modelos Log-Log

(Lección 10) **T-7** Ejemplo de modelo Log-Log

Ejemplo 25. Función de producción Cobb-Douglas Pensemos en la clásica función de producción

$$Q = cK^{\beta_2} L^{\beta_3}$$

donde Q es el volumen de la producción, K es el capital empleado y L el trabajo empleado. Supongamos, además, que hay un efecto aleatorio adicional ν debido a otras causas o factores:

$$Q = cK^{\beta_2} L^{\beta_3} \nu;$$

Tomando logaritmos en $Q = cK^{\beta_2} L^{\beta_3} \nu$, tenemos

$$\ln Q = \beta_1 \mathbb{1} + \beta_2 \ln K + \beta_3 \ln L + U,$$

donde $\beta_1 = \ln c$, y $U = \ln \nu$ (es decir, $\nu = e^U$.)

En los modelos Log-Log los parámetros β_j son elasticidades constantes...

La interpretación de un valor como $\beta_2 = 5$ es que un incremento de capital del 1% aumenta la producción en un 5%.

F127

(Lección 10) Ejercicio en clase. N-5.

■ Código: RamanathanAp6-11.inp

Elasticidades en la demanda del transporte en autobús. Abra el conjunto de datos `data4-4.gdt`, del libro de Ramanathan.

- (a) Estime un modelo de regresión entre el logaritmo de *bustravl* y el resto de variables, también en logaritmos.
- (b) Elimine secuencialmente del modelo las variables no significativas al 10% ni individual ni conjuntamente.
- (c) Decimos que la demanda es inelástica cuando el valor absoluto de la elasticidad es menor que 1 (elástica en caso contrario). Contraste si la elasticidad de la demanda de viajes de autobús con respecto a las distintas variables explicativas es 1.

Enlace a algunas [prácticas de la Lección 10](#)

[Fin de la lección](#)

LECCIÓN 11: Interpretación de coeficientes en modelos con regresores cualitativos

28 Variables ficticias

- Novales (1993, Secciones 4.10 y 4.11, pps. 139–145)
- Wooldridge (2006, Capítulo 7)
- Johnston and Dinardo (2001, Secciones 4.5 y 4.6, pps. 145–160)
- Gujarati (2003, Capítulo 9)

(Lección 11)

T-1

Variables ficticias (*dummies*)

Variable discreta que clasifica “categorías”

(Indicador que toma valores 0 ó 1)

Usos:

- inclusión de información cualitativa (empresa, sexo, etc.)
- división de la muestra en dos períodos (contraste cambio estructural)

En este caso los coeficientes β_j tienen otra interpretación (no son pendientes).

F129

28.1 Interpretación de los coeficientes de las variables ficticias

Ejemplo 26. Relación entre salario por hora trabajada percibido por el trabajador n -ésimo (W_n) y su nivel de estudios (variable cualitativa representada por 3 dummies:)

W = Salario del trabajador n -ésimo

$$\mathbb{1}_P = \begin{cases} 1, & \text{sin estudios o sólo estudios primarios (P)} \\ 0, & \text{en caso contrario} \end{cases}$$

$$\mathbb{1}_M = \begin{cases} 1, & \text{con estudios medios (no superiores) (M)} \\ 0, & \text{en caso contrario} \end{cases}$$

$$\mathbb{1}_S = \begin{cases} 1, & \text{con estudios superiores (S)} \\ 0, & \text{en caso contrario} \end{cases}$$

$$W = \alpha_1 \mathbb{1}_P + \alpha_2 \mathbb{1}_M + \alpha_3 \mathbb{1}_S + U \quad (55)$$

donde $\mathbb{1}_P + \mathbb{1}_M + \mathbb{1}_S = 1$.

La matriz de regresores es X es

$$X = \begin{bmatrix} \mathbf{1}_{N_1 \times 1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{N_2 \times 1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{N_3 \times 1} \end{bmatrix},$$

donde $\mathbf{1}_{N_j \times 1}$ es una columna de unos, con tantos unos como el número de trabajadores con educación de nivel j (N_j).

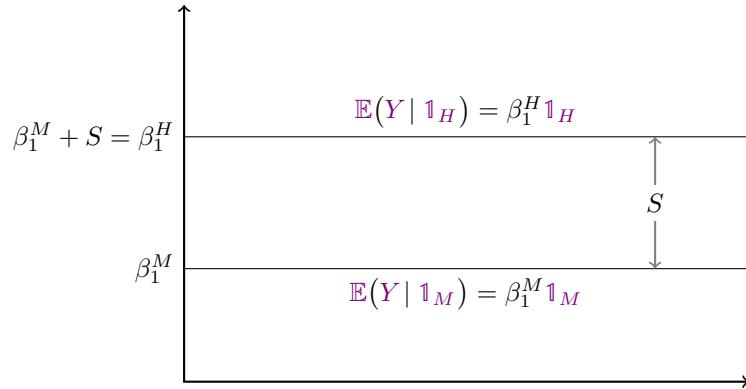
Es un ejemplo de regresión ortogonal particionada, donde las ecuaciones normales son: $\mathbf{X}^\top \mathbf{X} \hat{\alpha} = \mathbf{X}^\top \mathbf{w}$, es decir,

$$\begin{bmatrix} N_1 & 0 & 0 \\ 0 & N_2 & 0 \\ 0 & 0 & N_3 \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} \sum_{n \in P} w_n \\ \sum_{n \in M} w_n \\ \sum_{n \in S} w_n \end{bmatrix},$$

por lo tanto $\hat{\alpha}_j = N_j^{-1} \sum_{n=1}^{N_j} w_n = m_{\mathbf{w}_j}$; es decir, es el salario medio en cada nivel j de educación.

Diferentes términos constantes

$$\mathbb{1}_H(\omega) = \begin{cases} 1 & \omega \in H \\ 0 & \omega \notin H \end{cases} \quad \text{y donde } \mathbb{1}_H + \mathbb{1}_M = 1$$



(Lección 11) Ejercicio en clase. N-1.

■ CÓDIGO: RamanathanPp7-1.inp

Diferencias salariales entre hombres y mujeres.

Abra el conjunto de datos data7-1.gdt, del libro de Ramanathan, con datos sobre 49 trabajadores.

(a) Estime el modelo

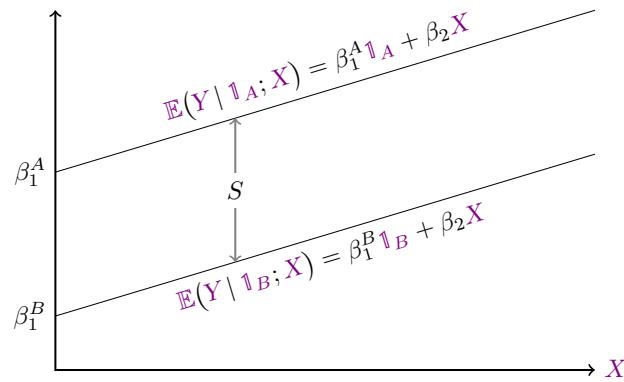
$$WAGE = \beta_1 \mathbb{1} + \beta_2 D + U$$

donde D es una variable que toma el valor 1 si el trabajador es varón.

(b) Interprete los coeficientes.

Calcule los salarios medios de hombres y mujeres, así como la diferencia de dichas medias. ¿Confirmán su interpretación de los coeficientes?

La misma idea se puede generalizar



$$\beta_1^A = \beta_1^B + S$$

Suponga el modelo

$$\ln(Y) = a\mathbb{1} + bX + cD + U$$

donde D solo toma los valores cero o uno.

Calculando la exponencial de esta expresión:

- el crecimiento porcentual $\frac{\Delta Y}{Y}$ al pasar de $D = 0$ a $D = 1$ es

$$100[\exp(c) - 1]$$

- el crecimiento porcentual $\frac{\Delta Y}{Y}$ al pasar de $D = 1$ a $D = 0$ es

$$100[\exp(-c) - 1]$$

F134

(Lección 11) Ejercicio en clase. N-2.

■ Código: RamanathanEX7-1.inp

Diferencias salariales entre hombres y mujeres. Abra el conjunto de datos data7-2.gdt, del libro de Ramanathan, con datos sobre 49 trabajadores.

(a) Estime el modelo

$$WAGE = \beta_1\mathbb{1} + \beta_2D + \beta_3EXPER + U$$

donde D es una variable ficticia que toma el valor 1 si el trabajador es varón.

Interprete los coeficientes.

(b) Estime el modelo

$$\ln WAGE = \beta_1\mathbb{1} + \beta_2D + \beta_3EXPER + U$$

Interprete los coeficientes.

(c) Estime el modelo

$$\ln WAGE = \beta_1\mathbb{1} + \beta_2D + \beta_3EXPER + \beta_4EDUC + U.$$

Interprete los coeficientes y compare los resultados de los modelos.

Nota 13. Se debe tener cuidado con los problemas de multicolinealidad exacta que pueden aparecer, y cómo interpretar los coeficientes asociados a las “dummies”. Veámoslo en el siguiente ejemplo:

Interpretación de los coeficientes: Ejemplo con multicolinealidad

Ejemplo 27. Un modelo de salarios más completo:

Relación entre salario por hora trabajada percibido por un trabajador (W), su antigüedad en la empresa (A), los años de experiencia en el sector (X), y su nivel de estudios (variable cualitativa representada por las 3 dummies anteriores)

$$W = \beta_1\mathbb{1} + \beta_2A + \beta_3X + \alpha_1\mathbb{1}_P + \alpha_2\mathbb{1}_M + \alpha_3\mathbb{1}_S + U \quad (56)$$

Aquí

- β_1 salario “autónomo” común a todos los trabajadores
- β_2 efecto antigüedad
- β_3 efecto experiencia
- α_j efecto del nivel de estudios j

Pero puesto que $\mathbb{1} = \mathbb{1}_P + \mathbb{1}_M + \mathbb{1}_S$, hay *multicolinealidad exacta* y no es posible la estimación de los parámetros.

Hay varias soluciones posibles, y todas pasan por eliminar uno de los regresores linealmente dependientes:

- Reemplazar la constante $\mathbb{1}$ por $\mathbb{1}_P + \mathbb{1}_M + \mathbb{1}_S$.

$$\begin{aligned} W &= \beta_1(\mathbb{1}_P + \mathbb{1}_M + \mathbb{1}_S) + \beta_2 A + \beta_3 X + \alpha_1 \mathbb{1}_P + \alpha_2 \mathbb{1}_M + \alpha_3 \mathbb{1}_S + U \\ &= \beta_2 A + \beta_3 X + (\beta_1 + \alpha_1) \mathbb{1}_P + (\beta_1 + \alpha_2) \mathbb{1}_M + (\beta_1 + \alpha_3) \mathbb{1}_S + U \\ &= \beta_2 A + \beta_3 X + \delta_1 \mathbb{1}_P + \delta_2 \mathbb{1}_M + \delta_3 \mathbb{1}_S + U \end{aligned}$$

Es decir

$$W = \beta_2 A + \beta_3 X + \delta_1 \mathbb{1}_P + \delta_2 \mathbb{1}_M + \delta_3 \mathbb{1}_S + U \quad (57)$$

que es modelo sin término cte. En (56) β_1 es el salario “autónomo” común a todos (indep. de a, x, E), y α_j $\delta_j = (\beta_1 + \alpha_j)$ es una combinación del salario “autónomo” (e inobservable) y del nivel de estudios j .

- Reemplazar $\mathbb{1}_M$ por $(\mathbb{1} - \mathbb{1}_P - \mathbb{1}_S)$. Operando:

$$W = \theta_0 \mathbb{1} + \beta_2 A + \beta_3 X + \theta_1 \mathbb{1}_P + \theta_3 \mathbb{1}_S + U \quad (58)$$

- $\theta_0 = (\beta_1 + \alpha_2)$ es como δ_2 de (57) (autónomo + Est. M)
- $\theta_1 = (\alpha_1 - \alpha_2)$ pérdida por tener estudios P en lugar de M
- $\theta_3 = (\alpha_3 - \alpha_2)$ ganancia por tener estudios S en lugar de M

(el referente es la categoría eliminada: Estudios M)

Piense en la interpretación con otras soluciones alternativas.

28.2 Contrastes de homogeneidad

Los “*contrastes de homogeneidad de los parámetros*” (entre distintas sub-muestras **excluyentes y exhaustivas**, i.e., entre distintas particiones de una muestra dada) se pueden realizar mediante *dummies*. Podemos asociar un subconjunto de índices $\mathbb{1}_C$ “*a una característica determinada*” (sexo, región geográfica, nivel de educación, sector económico, empresa, etc.) que define a una sub-muestra de interés particular.

Ejemplo 28. Contrastes de homogeneidad del salario para distintos niveles educativos: Supongamos que queremos realizar un contraste de homogeneidad de los efectos derivados de los distintos niveles de educación (que en el modelo original (56) se expresaría como $H_0 : \alpha_1 = \alpha_2 = \alpha_3$). Puesto que no podemos trabajar con el modelo original debido al problema de multicolinealidad, podemos reescribir la hipótesis de homogeneidad como

$$H_0 : \alpha_1 - \alpha_2 = 0 \text{ y } \alpha_3 - \alpha_2 = 0$$

es decir, empleando (58) podemos realizar el contraste de significatividad conjunta de los parámetros θ_1 y θ_3 .

$$H_0 : \theta_1 = 0 \text{ y } \theta_3 = 0$$

¿Difiere el salario de trabajadores con distinto nivel de educación?

- Modelo original (56)

$$W = \beta_1 \mathbb{1} + \beta_2 A + \beta_3 X + \alpha_1 \mathbb{1}_P + \alpha_2 \mathbb{1}_M + \alpha_3 \mathbb{1}_S + U$$

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3$$

No se puede contrastar debido a la multicolinealidad

- Modelo transformado (58) (*quitando $\mathbb{1}_M$*)

$$W = \theta_0 \mathbb{1} + \beta_2 A + \beta_3 X + \theta_1 \mathbb{1}_P + \theta_3 \mathbb{1}_S + U$$

$$H_0 : \theta_1 = 0 \text{ y } \theta_3 = 0$$

F137

Observación. Supongamos que deseamos estimar el efecto de los tres niveles de educación sobre el salario: (56) no puede estimarse por el problema de multicolinealidad; y al estimar (57) ó (58) no disponemos de estimaciones individuales de α_1, α_2 y α_3 .

Una forma de estimar estos parámetros es imponer una restricción lineal sobre α_1, α_2 y α_3 (siempre y cuando dicha restricción tenga sentido, claro!). Por ejemplo, si impusieramos la restricción $\alpha_1 + \alpha_2 + \alpha_3 = 0$ (algo que implica un tipo de renormalización tal que la suma de los efectos es cero —algo que no está claro que sea cierto...) y sustituyendo α_1 por $(-\alpha_2 - \alpha_3)$ en el modelo original (56) tenemos:

$$\begin{aligned} W &= \beta_1 \mathbb{1} + \beta_2 A + \beta_3 X + \alpha_1 \mathbb{1}_P + \alpha_2 \mathbb{1}_M + \alpha_3 \mathbb{1}_S + U \\ &= \beta_1 \mathbb{1} + \beta_2 A + \beta_3 X + \alpha_2 (\mathbb{1}_M - \mathbb{1}_P) + \alpha_3 (\mathbb{1}_S - \mathbb{1}_P) + U \\ &= \beta_1 \mathbb{1} + \beta_2 A + \beta_3 X + \alpha_2 D^M + \alpha_3 D^S + U \end{aligned} \quad (59)$$

donde $D^M = (\mathbb{1}_M - \mathbb{1}_P)$, y $D^S = (\mathbb{1}_S - \mathbb{1}_P)$. Al estimar (59) se obtiene $\widehat{\alpha}_2$ y $\widehat{\alpha}_3$; Finalmente podemos calcular $\widehat{\alpha}_1 = -\widehat{\alpha}_2 - \widehat{\alpha}_3$.

Pero nótese que el efecto estimado para los distintos niveles de educación está “forzado” para que cumpla la restricción $\alpha_1 + \alpha_2 + \alpha_3 = 0$, algo que producirá sesgos cuando dicha restricción sea falsa.

28.2.1 Más contrastes de homogeneidad: uso dummies para contrastar cambios estructurales

$$Y = \beta_1 \mathbb{1} + \beta_2 X + U^*. \quad (60)$$

Partición en sub-muestras A y B .

Si sospechamos que β_1 cambia \rightarrow Modelo no restringido:

$$Y = \beta_1^A \mathbb{1}_A + \beta_1^B \mathbb{1}_B + \beta_2 X + U, \quad (61)$$

donde

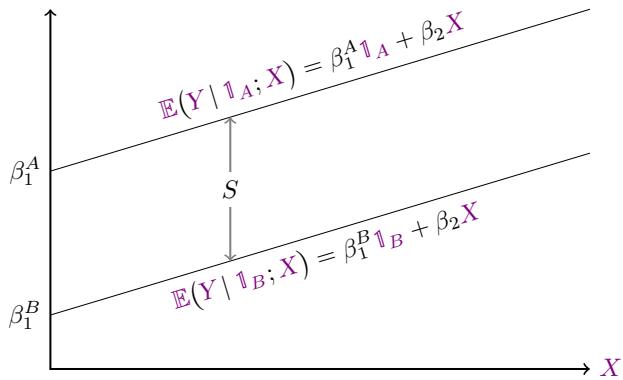
$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}, \quad \text{y donde } \mathbb{1}_A + \mathbb{1}_B = \mathbb{1}.$$

F138

Por claridad de exposición, supongamos que tenemos los datos ordenados; primero los que pertenecen a la categoría

A, y luego los de la categoría B. Entonces la matriz de regresores tiene la forma

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & X_{j-1} \\ 1 & 0 & X_j \\ 0 & 1 & X_{j+1} \\ 0 & 1 & X_{j+2} \\ \vdots & \vdots & \vdots \\ 0 & 1 & X_N \end{bmatrix}$$



(Lección 11) **T-5** Variables ficticias: contrastes de homogeneidad (constante)

Contraste $H_0 : \beta_1^A = \beta_1^B$. Dos opciones:

1. Contraste F de sumas residuales (página 106):

Estimando (60) y (61) ($H_1 : \beta_1^A \neq \beta_1^B$)

2. Por sustitución: $1_B = 1 - 1_A$ en (61);

$$Y = \beta_1^B 1 + \alpha 1_A + \beta_2 X + U, \quad (62)$$

donde $\alpha \equiv \beta_1^A - \beta_1^B$ (60 y 62 idénticas bajo H_0).

Ahora $H_0 : \alpha = 0$;

(basta contraste de signif. individual; uni o bilateral).

F140

Sustituyendo $1_B = 1 - 1_A$ en (61) tenemos:

$$\begin{aligned} Y &= \beta_1^A 1_A + \beta_1^B 1_B + \beta_2 X + U \\ &= \beta_1^A 1_A + \beta_1^B (1 - 1_A) + \beta_2 X + U \\ &= \beta_1^B 1 + (\beta_1^A - \beta_1^B) 1_A + \beta_2 X + U \\ &= \beta_1^B 1 + \alpha 1_A + \beta_2 X + U, \end{aligned}$$

Ahora la matriz de regresores es

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & X_{j-1} \\ 1 & 1 & X_j \\ 1 & 0 & X_{j+1} \\ 1 & 0 & X_{j+2} \\ \vdots & \vdots & \vdots \\ 1 & 0 & X_N \end{bmatrix}$$

(Lección 11)

T-6 Variables ficticias: contrastes de homogeneidad (pendiente)

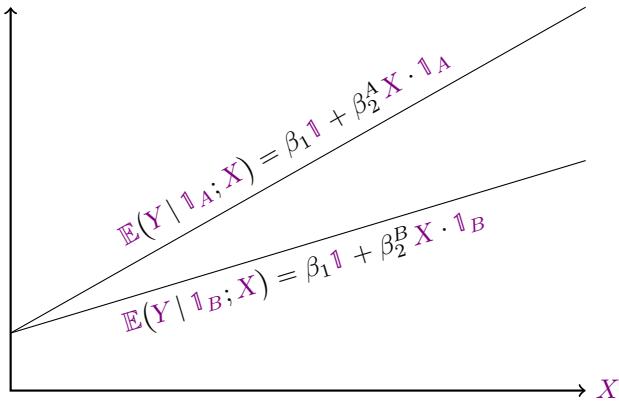
$$Y = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \mathbf{U}^*.$$

Partición en sub-muestras A y B .

Si sospechamos β_2 (pendiente) cambia \rightarrow Modelo no restringido:

$$Y = \beta_1 \mathbf{1} + \beta_2^A b \mathbf{X} \cdot \mathbf{1}_A + \beta_2^B \mathbf{X} \cdot \mathbf{1}_B + \mathbf{U}, \quad (63)$$

F141



(Lección 11)

T-7 Variables ficticias: contrastes de homogeneidad (pendiente)

Contraste $H_0 : \beta_2^A = \beta_2^B$. Dos opciones:

1. Por sumas residuales:

Estimando (60) y (63) ($H_1 : \beta_2^A \neq \beta_2^B$)

2. Por sustitución: $\mathbf{1}_B = \mathbf{1} - \mathbf{1}_A$ en (63);

$$Y = \beta_1 \mathbf{1} + \beta_2^B \mathbf{X} + \delta \mathbf{X} \cdot \mathbf{1}_A + \mathbf{U}, \quad (64)$$

donde $\delta \equiv \beta_2^A - \beta_2^B$,

(60 y 64 idénticas bajo H_0).

Ahora $H_0 : \delta = 0$;

(basta contraste de signif. individual; uni o bilateral).

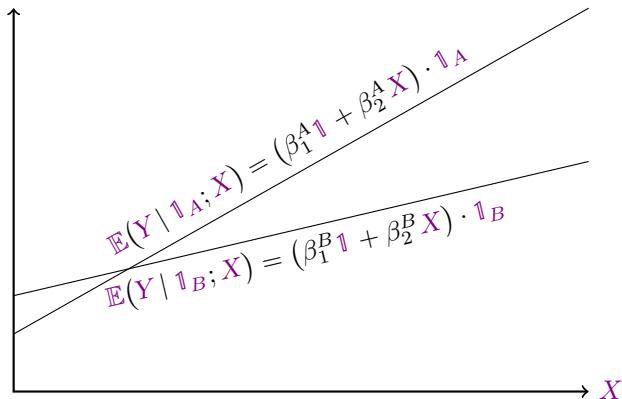
F143

Sustituyendo $\mathbb{1}_B = \mathbb{1} - \mathbb{1}_A$ en (63) tenemos,

$$\begin{aligned}\mathbf{Y} &= \beta_1 \mathbb{1} + \beta_2^A \mathbf{X} \mathbb{1}_A + \beta_2^B \mathbf{X} \mathbb{1}_B + \mathbf{U} \\ &= \beta_1 \mathbb{1} + \beta_2^A \mathbf{X} \mathbb{1}_A + \beta_2^B (\mathbb{1} - \mathbb{1}_A) \mathbf{X} + \mathbf{U} \\ &= \beta_1 \mathbb{1} + \beta_2^B \mathbf{X} + (\beta_2^A - \beta_2^B) \mathbf{X} \mathbb{1}_A + \mathbf{U} \\ &= \beta_1 \mathbb{1} + \beta_2^B \mathbf{X} + \delta \mathbf{X} \mathbb{1}_A + \mathbf{U},\end{aligned}$$

Ahora la matriz de regresores es

$$\mathbf{x} = \begin{bmatrix} 1 & X_1 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & X_{j-1} & X_{j-1} \\ 1 & X_j & X_j \\ 1 & X_{j+1} & 0 \\ 1 & X_{j+2} & 0 \\ \vdots & \vdots & \vdots \\ 1 & X_N & 0 \end{bmatrix}$$



EJERCICIO 1. Generalice el procedimiento para el caso en el que tanto la constante como la pendiente cambian de una sub-muestra a otra.

Este procedimiento se puede generalizar a más de dos sub-muestras y más variables explicativas simultáneamente.

28.3 Términos de interacción

(Lección 11)

T-8 Términos de interacción

Considere el modelo de consumo

$$C = \alpha \mathbb{1} + \beta Y + \mathbf{U}$$

Considere la hipótesis de que la propensión marginal al consumo (β) depende de la posesión de activos A . Entonces

$$C = \alpha \mathbb{1} + (\beta_1 + \beta_2 \mathbb{1}_A) Y + \mathbf{U},$$

o

$$C = \alpha \mathbb{1} + \beta_1 Y + \beta_2 (\mathbb{1}_A \cdot Y) + \mathbf{U}.$$

El término $(\mathbb{1}_A \cdot Y)$ se llama término de interacción.

F144

Prácticas de la Lección 11

- A continuación tiene algunos ejercicios adicionales propuestos.

La siguiente práctica reproduce la aplicación 7.6 del libro de Ramanathan.

(Lección 11) Ejercicio en clase. N-3.

■ CÓDIGO: [RamanathanPS7-6.inp](#)

Possible cambio estructural en la participación de las mujeres en el mercado laboral Abra el conjunto de datos `data7-4.gdt`, del libro de Ramanathan, con datos de 50 estados de EEUU sobre la participación de las mujeres en el mercado laboral. Los 50 primeros son del año 1980 y los 50 últimos de 1990. La variable a explicar es `WLFP`, que es el **porcentaje de participación** de mujeres mayores de 16 años en el mercado laboral. `YF` es el salario mediano de las mujeres (en miles de dólares); `YM` es el salario mediano de los hombres (en miles de dólares); `EDUC` es el porcentaje, de entre las mujeres con 24 o más años, con el título de bachillerato; `UE` es la tasa de desempleo; `MR` es el porcentaje de mujeres mayores de 16 años que están casadas; `DR` es el porcentaje de mujeres divorciadas; `URB` es el porcentaje de población urbana; `WH` es el porcentaje de mujeres mayores de 16 años que son de raza blanca.

Por último, la variable ficticia `D90` vale 1 si el dato corresponde al año 1990 y 0 en caso contrario.

- Estime un modelo para `WLFP` empleando todas las variables explicativas (excepto `D90`).
- Realice un contraste de cambio estructural (Contraste de Chow), para estudiar si ha habido un cambio en la disposición de las mujeres a entrar en el mercado laboral entre los años 1980 y 1990.
- Si rechaza H_0 de ausencia de cambio estructural, genere todas las variables de interacción necesarias para captar el cambio (genere todas las variables que han sido empleadas en el test de cambio estructural). Re-estime el modelo con ellas.
- Este último modelo tiene muchos regresores. Si hay variables estadísticamente no significativas, reduzca el modelo como de costumbre.
- Interprete los resultados. En particular,
 - ¿Son distintos los efectos del porcentaje de mujeres casadas (`MF`)? ¿Cuales son sus efectos? ¿Es significativo el efecto en los años 90?
 - ¿Son distintos los efectos “desaliento” debidos a la tasa de paro (`UE`)? ¿Cuales son sus efectos? ¿Es significativo el efecto en los años 90?
 - ¿Son distintos los efectos debidos al salario mediano de las mujeres (`YF`)? ¿Cuales son sus efectos? ¿Es significativo el efecto en los años 90? Ramanathan hace notar que este efecto no está justificado y lo atribuye a una difícil identificación de los efectos de ésta variable. ¿Cuál puede ser el problema?

(Lección 11) Ejercicio en clase. N-4.

■ CÓDIGO: [wage1dummiesB.inp](#)

Log-lin con variables ficticias. Estimaremos las diferencias salariales entre cuatro grupos: hombres casados (`marrmale`), mujeres casadas (`marrfem`), hombres solteros y mujeres solteras (`singfem`)

- Cargue los datos `wage1.gdt` del libro de texto de Wooldridge (2006, Ejemplo 7.6)
- Genere las variables ficticias necesarias para indicar los cuatro grupos.
- Estime por MCO el siguiente modelo

$$\begin{aligned}\log(wage) = & \beta_1 + \beta_2 \cdot marrmale + \beta_3 marrfem + \beta_4 singfem \\ & + \beta_5 educ + \beta_6 exper + \beta_7 exper^2 + \beta_8 tenure + \beta_9 tenure^2 + \text{OtrosFactores}\end{aligned}$$

- ¿Quién es el grupo de referencia? Interprete los coeficientes correspondientes a las variables ficticias que ha generado; en particular, ¿en qué porcentaje varía el salario con cada una de estas variables ficticias? (recuerde que el cálculo es $100 * (\exp(\beta) - 1)$)
- ¿Qué pasaría si también incluimos en el modelo la variable ficticia correspondiente a los hombres solteros?
- ¿Es significativa la diferencia de salarios entre mujeres solteras y casadas al 5%? Calcule un intervalo de confianza para $\beta_4 - \beta_3$ al 95% para comprobarlo.
- A partir del modelo estimado no es fácil ver si esta última diferencia salarial es estadísticamente significativa. Hay

una alternativa. Cambiar el grupo de referencia. Estime el siguiente modelo

$$\begin{aligned}\log(wage) = & \beta_1 + \beta_2 \cdot marrmale + \beta_3 singmale + \beta_4 singfem \\ & + \beta_5 educ + \beta_6 exper + \beta_7 exper^2 + \beta_8 tenure + \beta_9 tenure^2 + OtrosFactores\end{aligned}$$

y verifique que la estimación e intervalo de confianza para β_4 (diferencia entre mujer soltera y el grupo de referencia, que ahora es mujer casada) coincide con lo calculado en el apartado anterior.

- (h) Calcule la diferencia estimada en el salario (no en el logaritmo del salario) entre mujeres solteras y casadas. Calcule también el intervalo de confianza al 95%.

(Lección 11) Ejercicio en clase. N-5.

☞ Código: [RamanathanEX7-2.inp](#)

Precio de viviendas unifamiliares Abra el conjunto de datos `data7-3.gdt` del libro de Ramanathan.

- (a) Estime un modelo para el precio en función del tamaño.
- (b) Estime un modelo para el precio en función de todas las variables explicativas disponibles.
- (c) Elimine del último modelo aquellas variables no significativas.
- (d) Compare los resultados e interprete los coeficientes de este último modelo.

- (e) Repita los pasos anteriores pero usando el logaritmo del `sqft` en lugar de `sqft`
- (f) Elimine del último modelo el regresor `ln sqft`. ¿Empeoran los resultados?

Fin de la lección

LECCIÓN 12: Multicolinealidad

29 Multicolinealidad

- Novales (1993, Capítulo 10, pp. 344)
- Wooldridge (2006, Sección 3.4, pps. 103–108)

La existencia de multicolinealidad significa que existen enormes dificultades para aislar e identificar el efecto individual de algunas (o todas) las variables explicativas del modelo.

(Lección 12)

T-1

Dos tipos de multicolinealidad

Estricta o perfecta: $|\mathbf{X}^T \mathbf{X}| = 0$ $\text{rg}(\mathbf{X}^T \mathbf{X}) < k$.

- incumplimiento del Sup-IV de independencia lineal
- Infinitas soluciones (β no identificado).

No estricta o de grado: $|\mathbf{X}^T \mathbf{X}| \simeq 0$

- Alta correlación entre regresores.
- A mayor correlación menor determinante $|\mathbf{X}^T \mathbf{X}|$ y mayor gravedad del problema.

Multicolinealidad → incertidumbre sobre el valor los β_j .

F146

Multicolinealidad exacta. Se denomina multicolinealidad exacta al caso en que las ecuaciones normales

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

tienen infinitas soluciones debido a que $\mathbf{X}^T \mathbf{X}$ no es una matriz invertible (es el caso en el que las columnas de \mathbf{X} son linealmente dependientes).

Ejemplo de multicolinealidad exacta: Sea

$$\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \mathbf{U},$$

donde $\mathbf{X}_2 = \lambda \mathbf{X}_3$ entonces, sustituyendo \mathbf{X}_2 por $\lambda \mathbf{X}_3$ tenemos:

$$\mathbf{Y} = \beta_1 \mathbf{1} + (\lambda \beta_2 + \beta_3) \mathbf{X}_3 + \mathbf{U},$$

y consecuentemente $(\lambda \beta_2 + \beta_3) = \frac{\text{Cov}(\mathbf{Y}, \mathbf{X}_3)}{\text{Var}(\mathbf{X}_3)}$, por tanto son posibles infinitos valores para β_2 y β_3 .

Volvamos sobre el Ejemplo 2 en la página 11. ¿Cómo distinguir entre el efecto de la experiencia y de la antigüedad cuando en la muestra de trabajadores nadie ha cambiado de trabajo? ¡No se puede! sólo se puede captar el efecto conjunto.

El Supuesto 4 (independencia lineal de los regresores \mathbf{X}) excluye la multicolinealidad exacta.

Multicolinealidad “no exacta” o “de grado”. Modifiquemos el ejemplo de los salarios... ¿Y si en la muestra disponible **casi** nadie ha cambiado de trabajo? Entonces el determinante $|\mathbf{X}^T \mathbf{X}|$ no es cero, pero estará **próximo** a cero. En tal caso las varianzas de los estimadores se hacen muy grandes.

Variables explicativas con una fuerte correlación entre ellas

- Varios regresores son series temporales con tendencia
- Inclusión de varios retardos de variables explicativas.
- Regresores cuya información ya está en otras vbles. expl.
- Inclusión de muchos regresores aumenta las posibilidades de multicolinealidad.

Caso opuesto: variables explicativas ortogonales. ... (Práctica Simulación 1 y 2).

F147

Puesto que muchas de las magnitudes económicas se mueven al unísono, o de manera muy parecida, la aparición de la *multicolinealidad “de grado”* es relativamente frecuente al emplear datos económicos. En tales casos la correlación entre los datos de distintas variables suele ser muy elevada. Si algunas de estas variables son regresores de un mismo modelo, es fácil que aparezcan problemas de multicolinealidad; es decir, que *existen dificultades para aislar e identificar el efecto individual de cada una de ellas*, ya que no será posible observar cambios de una sola de las variables junto a un comportamiento más o menos estable de las otras (recuerde que la interpretación de los coeficientes es *ceteris paribus*). Como todas se mueven a la vez de manera similar, es imposible aislar sus efectos individuales y, por lo tanto, no es posible interpretar correctamente el valor de los coeficientes estimados.

¡No confunda esta situación con que haya una fuerte correlación entre la variable dependiente y los regresores. Esto último no es un problema en absoluto! El problema solo surge cuando existe una fuerte correlación *de los regresores entre si*.

También es frecuente la aparición de multicolinealidad no estricta en modelos con regresores que son distintos retardos de una misma serie temporal, pues la inercia de algunas magnitudes hace que los retardos de la variable se parezcan mucho entre si. Consecuentemente tampoco es posible saber qué efecto tiene cada uno de los retardos individualmente (esto es frecuente en variables que muestran un crecimiento tendencial).

Pese a que MCO es ELIO:

- Difícil interpretación coeficientes: varianza de $\hat{\beta}$ enorme:

$$\text{Var}(\hat{\beta}) = \sigma^2 \mathbf{E}(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{\sigma^2}{|\mathbf{E}(\mathbf{X}^\top \mathbf{X})|} \text{Adj}(\mathbf{E}(\mathbf{X}^\top \mathbf{X}))$$

Poca precisión (pequeñas variaciones muestra → grandes variaciones en estimación).

Alta correlación entre estimadores.

- Propensión a *no rechazar (casi) cualquier hipótesis* F99

$$(H_0 : \beta_i = a \rightarrow T = \frac{\hat{\beta}_j - a}{\widehat{\text{Dt}}(\hat{\beta}_j)} \sim t_{N-k}.)$$

- Pero no afecta al contraste de significación conjunta

$$F = \frac{N - k}{k - 1} \cdot \frac{\widehat{R}^2}{1 - \widehat{R}^2} \sim F_{k-1, N-k,}$$

Ningún problema si sólo nos preocupa la predicción.

F148

- Elevados R^2 con parámetros no significativos individualmente.
- Examen de la correlación de las vbles. expl.
 - Cálculo de correlaciones simples (sólo entre pares)
 - Regresiones entre vbles. expl.
- Añadir o quitar regresores al modelo ocasiona grandes cambios en los coeficientes estimados.
- Tests de colinealidad y análisis del “tamaño” de $\mathbf{X}^\top \mathbf{X}$

F149

¡Difícil solución!

No es un problema de especificación de modelo; es un problema de datos (Columnas de **X**) (insuficiente información para estimar TODOS los parámetros con precisión)

- Ignorar el problema si sólo nos interesa la predicción.
- Obtener más datos.
- Reformular el modelo
 - Imponer restricciones (información extra-muestral)
 - Transformar variables (primeras diferencias, datos per-cápita, etc.)
- Suprimir variables (si NO hay razones teóricas para mantenerlas)
 - ¡Ojo! quitar variables puede suponer incumplimiento del Supuesto 2: $E(\mathbf{U} | \mathbf{X}) = \mathbf{0}$
- Ortogonalizar regresores con regresiones auxiliares.

F150

Enlace a algunas **prácticas de la Lección 12**

- Simulación: Ejemplo de datos simulados
- Simulación 2: Ejemplo de datos simulados (Incorporando correlación entre regresores)
- Houses 3: Precio de casas unifamiliares (tasación de inmobiliarias y características de las viviendas)
- A continuación tiene algunos ejercicios adicionales propuestos.

Fin de la lección

(Lección 12) Ejercicio en clase. N-1.

☞ Código: [RamanathanEX5-1.inp](#)

Construcción de vivienda nueva. Usando `data4-3.gdt` del libro de Ramanathan, con el nº de viviendas iniciadas (en miles) en los EEUU (*housing*), población (*pop*), PIB en miles de millones de dólares (*gnp*) y tipo de interés del crédito hipotecario (*inrate*):

(a) Estime el siguiente modelo “ModeloA” y añádalo a la tabla de modelos

$$housing_n = \alpha_1 + \alpha_2 \cdot inrate_n + \alpha_3 \cdot pop_n + \text{others factores}_n$$

(b) Estime el siguiente modelo “ModeloB” y añádalo a la tabla de modelos

$$housing_n = \beta_1 + \beta_2 \cdot inrate_n + \beta_3 \cdot gnp_n + \text{others factores}_n$$

(c) Estime el siguiente modelo “ModeloC” y añádalo a la tabla de modelos

$$housing_n = \gamma_1 + \gamma_2 \cdot inrate_n + \gamma_3 \cdot pop_n + \gamma_4 \cdot gnp_n + \text{others factores}_n$$

(d) Compare los resultados.

(e) Realice un contraste de colinealidad entre regresores (**vif**).

(f) Contraste la significatividad conjunta de *gnp* y *pop* en ModeloC.

(g) Mire las correlaciones entre los tres regresores.

(Lección 12) Ejercicio en clase. N-2.

☞ Código: [RamanathanEX5-1v2.inp](#)

Construcción de vivienda nueva (transformación a datos per-cápita). Abra el conjunto de datos `data4-3.gdt`, del libro de Ramanathan, con datos de número de viviendas iniciadas (en miles) en los EEUU (*housing*), además de datos de población (*pop*), PIB en miles de millones de dólares con base en el año 1982 (*gnp*), y tipo de interés del crédito hipotecario (*inrate*).

- (a) Estime de nuevo el siguiente modelo (que tiene multicolinealidad)

$$housing_n = \gamma_1 + \gamma_2 \cdot intrate_n + \gamma_3 \cdot pop_n + \gamma_4 \cdot gnp_n + \text{otros factores}_n$$

- (b) Defina la construcción de viviendas per-cápita ($housingPC=housing/pop$) y la correspondiente producción per-cápita ($gnpPC$); y estime el siguiente modelo

$$housingPC_n = \beta_1 + \beta_2 \cdot intrate_n + \beta_3 \cdot gnpPC_n + \text{otros factores}_n$$

- (c) Compare los resultados.

- (d) Mire las correlaciones entre los nuevos regresores y compárelas con las del antiguo.

- (e) Realice un contraste de colinealidad entre regresores en ambos modelos.

Gastos de mantenimiento de los automóviles

La siguiente práctica reproduce el ejemplo 5.2 del libro de Ramanathan.

(Lección 12) Ejercicio en clase. N-3.

 Código: [RamanathanEX5-2.inp](#)

Abra el conjunto de datos **data3-7.gdt**, del libro de Ramanathan, sobre costes de mantenimiento de coches Toyota (*cost*), así como la edad de cada coche en semanas (*age*) y en número de millas recorridas (*miles*).

- (a) Mire las correlaciones entre los dos regresores.

- (b) Estime el siguiente modelo “ModeloA” y añádalo a la tabla de modelos

$$cost_n = \alpha_1 + \alpha_2 \cdot age_n + \text{otros factores}_n$$

- (c) Estime el siguiente modelo “ModeloB” y añádalo a la tabla de modelos

$$cost_n = \beta_1 + \beta_2 \cdot miles_n + \text{otros factores}_n$$

- (d) Estime el siguiente modelo “ModeloC” y añádalo a la tabla de modelos

$$cost_n = \gamma_1 + \gamma_2 \cdot age_n + \gamma_4 \cdot miles_n + \text{otros factores}_n$$

- (e) Compare los resultados.

- (f) Realice un contraste de colinealidad entre regresores.

- (g) Contraste la significatividad conjunta de *miles* y *age* en ModeloC.

- (h) La serie de errores absolutos de predicción se calcula como

$$ape_n = 100 \left| \frac{\hat{e}_n}{y_n} \right|$$

Calcule la serie de errores absolutos de predicción de cada modelo.

- (i) Calcule los MAPE de cada modelo. ¿Qué modelo predice mejor?

(Lección 12) Ejercicio en clase. N-4.

 Código: [RamanathanEX5-2v2.inp](#)

Ortogonalización de regresores mediante regresiones auxiliares. Abra el conjunto de datos **data3-7.gdt**, del libro de Ramanathan, sobre costes de mantenimiento de coches Toyota (*cost*), así como la edad de cada coche en semanas (*age*) y en número de millas recorridas (*miles*).

- (a) Estime el siguiente modelo “ModeloA”

$$cost_n = \alpha_1 + \alpha_2 \cdot age_n + \text{otros factores}_n$$

- (b) Realice una regresión auxiliar de *miles* sobre una constante y *age*. Guarde los residuos con el nombre *uhatMiles*. Piense qué interpretación tienen los residuos de esta regresión auxiliar.

- (c) Verifique que no hay correlación entre los residuos de esta regresión auxiliar (*uhatMiles*) y *age*. ¿Por qué?

- (d) Estime el siguiente modelo “ModeloA-Raux”

$$cost_n = \alpha_1 + \alpha_2 \cdot age_n + \alpha_3 \cdot uhatMiles_n + \text{otros factores}_n$$

y compare los resultados con el “ModeloA”. Realice un test de colinealidad para el modelo “ModeloA-Raux”. ¿Hay indicios de colinealidad?

- (e) Estime el siguiente modelo “ModeloB”

$$cost_n = \beta_1 + \beta_2 \cdot miles_n + \text{otros factores}_n$$

- (f) Realice una regresión auxiliar de *age* sobre una constante y *miles*. Guarde los residuos con el nombre *uhatAge*.
 (g) Estime el siguiente modelo “ModeloB-Raux”

$$cost_n = \beta_1 + \beta_2 \cdot miles_n + \beta_3 \cdot uhatAge_n + \text{otros factores}_n$$

y compare los resultados con el “ModeloB”. Realice un test de colinealidad para el modelo “ModeloB-Raux”. ¿Hay indicios de colinealidad?

- (h) Piense qué interpretación tiene el regresor *uhatAge* en esta regresión.
 (i) Estime el siguiente modelo “ModeloC”

$$cost_n = \gamma_1 + \gamma_2 \cdot age_n + \gamma_4 \cdot miles_n + \text{otros factores}_n$$

- (j) Compare los resultados de los cinco modelos.
 (k) Piense qué signos serían los esperados para los coeficientes y decida qué modelo prefiere.

(Lección 12) Ejercicio en clase. N-5.

■ CÓDIGO: [RamanathanAp5-4.inp](#)

Pobreza y sus determinantes (omisión de variables). Abra el conjunto de datos *data4-6.gdt*, del libro de Ramanathan, sobre el porcentaje de familias con una renta por debajo del nivel de pobreza (*povrate*) en los condados de California. Intentaremos explicar dicho porcentaje con las siguientes variables: *urb* es el porcentaje de población urbana; *famsize* es el número medio de personas por hogar, *unemp* es la tasa de desempleo, *highschl* es el porcentaje de población mayor de 25 que ha completado el “high school”, *college* es el porcentaje de la población mayor de 25 que ha completado 4 o más años de “college”, y *medinc* es la renta media familiar.

- (a) Observe la correlación entre las variables explicativas
- (b) Estime el modelo “ModeloA” con todas las variables explicativas.
- (c) Observe los elevados *p*-valores de algunas variables. Observe también el inesperado signo para *college*... pudiera ser debido a un problema de multicolinealidad.
- (d) Omita la variable menos significativa del modelo (la de mayor *p*-valor). Observe que prácticamente no hay cambios en el resultado.
- (e) Omita la siguiente variable menos significativa. Observe que ahora todas las variables son significativas, pero se mantiene el signo equivocado para la variable “college”.
- (f) Aunque la variable de renta familiar media (*medinc*) es la más significativa, cabe esperar que su efecto pueda ser aproximadamente explicado por las variables *highschl* y *college*. Pruebe a excluir dicha variable y compruebe que ahora el signo de *college* es adecuado. Esto sugiere la posibilidad de multicolinealidad.
- (g) Realice una regresión auxiliar de *medinc* sobre *famsize*, *unemp*, *highschl* y *college* y compruebe que dichas variables explican el 85% de la varianza de *medinc*. Quizá una solución es omitir dicha variable desde el principio, para poder captar adecuadamente el efecto de las demás variables.
- (h) Estime un modelo con todas las variables explicativas excepto *medinc*, y siga un procedimiento de eliminación secuencial de variables no significativas. ¿Es satisfactorio el modelo final?

LECCIÓN 13: Errores de especificación

- Novales (1993, Sección 3.11, pps. 100–103)
- Wooldridge (2006, Secciones 3.3 y 3.4, pps. 96–101 y 107–108)

30 Incorrecta especificación por omisión de variables explicativas

Sea el MLG

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + U \quad \text{con } \gamma \neq \mathbf{0}$$

donde $\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k]$ y $\mathbf{Z} = [\mathbf{Z}_1; \dots; \mathbf{Z}_p]$ son $k+p$ regresores.

Entenderemos con “*omitar variables*” a ignorar el subconjunto de regresores \mathbf{Z} y trabajar con el *incompleto* modelo

$$Y = \mathbf{X}\beta^* + U^*.$$

(Lección 13)

T-1 Errores de especificación: Omisión de variables

Sea el MLG

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + U \quad \text{con } \gamma \neq \mathbf{0}$$

donde $\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k]$ y $\mathbf{Z} = [\mathbf{Z}_1; \dots; \mathbf{Z}_p]$

1. Correcto: $Y = \mathbf{X}\beta + \underbrace{\mathbf{Z}\gamma + U}_{\text{("Completo")}}$
2. Incorrecto: $Y = \mathbf{X}\beta^* + \underbrace{U^*}_{\text{("Incompleto")}}$

F152

30.1 Esperanza del estimador

Sean \mathbf{X} , \mathbf{Z} y \mathbf{Y} *m.a.s* de \mathbf{X} , de \mathbf{Z} y de \mathbf{Y} respectivamente, donde $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + U$, que cumplen con los supuestos habituales: $\mathbb{E}(U | \mathbf{X}; \mathbf{Z}) = \mathbf{0}$; $\text{Var}(U | \mathbf{X}; \mathbf{Z}) = \sigma^2 \mathbf{I}$; y donde los $k+p$ regresores son linealmente independientes.

(Lección 13)

T-2 Estimación MCO con omisión de variables

Sea $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + U$ un *m.a.s* que cumple los supuestos.

Si omitimos \mathbf{Z} el estimador de β del modelo incompleto es:

$$\begin{aligned}\widehat{\beta}^* &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \mathbf{Z}\gamma + U) \\ &= \mathbf{I}\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}\gamma + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top U.\end{aligned}$$

$$\text{Entonces, } \mathbb{E}(\widehat{\beta}^* | \mathbf{X}) = \mathbf{I}\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}(\mathbf{Z} | \mathbf{X})\gamma.$$

Es decir, $\widehat{\beta}^*$ es insesgado cuando $\mathbf{Z} \perp \mathbf{X}$, pues en tal caso:

$$\mathbb{E}(\mathbf{Z} | \mathbf{X}) = \mathbf{0}.$$

F153

pues como U es ortogonal a los todos regresores, en particular es ortogonal a \mathbf{X} . Este estimador será insesgado si los regresores omitidos son ortogonales a los mantenidos en el modelo incompleto; es decir, cuando $\mathbb{E}(\mathbf{Z} | \mathbf{X}) = \mathbf{0}$; pues en tal caso

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}(\mathbf{Z} | \mathbf{X})\gamma = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{0}\gamma = \mathbf{0}.$$

Un ejemplo sencillo. Fijémonos en el siguiente MLG: $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \beta_3 \mathbf{Z} + \mathbf{U}$, es decir, $\mathbf{X} = [\mathbf{1}, \mathbf{X}, \mathbf{Z}]$.

(Lección 13)

T-3

Un ejemplo sencillo

Sea el MLG

$$\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \beta_3 \mathbf{Z} + \mathbf{U} \quad \text{con } \beta_3 \neq 0.$$

Si omitimos \mathbf{Z} y estimamos el modelo

$$\mathbf{Y} = \beta_1^* \mathbf{1} + \beta_2^* \mathbf{X} + \mathbf{U}^*$$

con $\mathbf{X} = [\mathbf{1}; \mathbf{X}]$ m.a.s de $\mathbf{X} = [\mathbf{1}; \mathbf{X}]$, entonces

$$E\left(\begin{pmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{pmatrix}\right) = E\left(E\left(\begin{pmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{pmatrix} \mid \mathbf{X}\right)\right) = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} E(\mathbf{Z})\beta_3 \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})}\beta_3 \end{pmatrix};$$

es decir

$$E(\widehat{\beta}_1^*) = \beta_1 + E(\mathbf{Z})\beta_3. \quad \text{y} \quad E(\widehat{\beta}_2^*) = \beta_2 + \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})}\beta_3,$$

Son insesgados solo si...

F154

P-1

(144)

Consecuentemente el estimador de β_1 será sesgado si la variable omitida tiene esperanza distinta de cero; y el estimador de β_2 será sesgado si la covarianza de la variable omitida con el regresor no constante es distinta de cero.

30.2 Varianza del estimador

(Lección 13)

T-4

Varianza del estimador MCO con el modelo incompleto

$$\text{Var}(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X}) \leq \text{Var}(\widehat{\boldsymbol{\beta}}_{(1:k)} \mid \mathbf{X}; \mathbf{Z})$$

Ojo! la mayoría de textos afirman: $\text{Var}(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X}) \leq \text{Var}(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}; \mathbf{Z})$.

F155

En cuanto a la varianza del estimador con el modelo incompleto, y teniendo en cuenta que

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^* - E(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X}) &= \overbrace{\widehat{\boldsymbol{\beta}}^*}^{\mathbf{I}\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}\boldsymbol{\gamma} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{U}} - \overbrace{E(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X})}^{(\mathbf{I}\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}))} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{Z}\boldsymbol{\gamma} - E(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}) + \mathbf{U}), \end{aligned}$$

y que $E(\mathbf{U} \mid \mathbf{X}; \mathbf{Z}) = \mathbf{0}$ (entonces \mathbf{U} y \mathbf{Z} son perpendiculares y consecuentemente $\text{Cov}(\mathbf{Z}\boldsymbol{\gamma}, \mathbf{U} \mid \mathbf{X}) = \mathbf{0}$); tenemos que

$$\begin{aligned} \text{Var}(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X}) &= E\left(\left[\widehat{\boldsymbol{\beta}}^* - E(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X})\right]^\top \left[\widehat{\boldsymbol{\beta}}^* - E(\widehat{\boldsymbol{\beta}}^* \mid \mathbf{X})\right] \mid \mathbf{X}\right) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E\left(\left[\mathbf{Z}\boldsymbol{\gamma} - E(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}) + \mathbf{U}\right]^\top \left[\mathbf{Z}\boldsymbol{\gamma} - E(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}) + \mathbf{U}\right] \mid \mathbf{X}\right) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Var}(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\text{Var}(\mathbf{U} \mid \mathbf{X})) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Var}(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\sigma_*^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Var}(\mathbf{Z}\boldsymbol{\gamma} \mid \mathbf{X}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} + \sigma_*^2 (\mathbf{X}^\top \mathbf{X})^{-1}; \end{aligned}$$

donde $\sigma_*^2 \mathbf{1} = \text{Var}(\mathbf{U} \mid \mathbf{X})$. Y como $\text{Var}(\mathbf{U} \mid \mathbf{X})$ es el cuadrado de la distancia entre \mathbf{U} y su proyección ortogonal $E(\mathbf{U} \mid \mathbf{X})$, esta distancia siempre será mayor o igual a la distancia entre \mathbf{U} y su proyección $E(\mathbf{U} \mid \mathbf{X}; \mathbf{Z})$ sobre el espacio $\mathcal{L}([\mathbf{X}; \mathbf{Z}])$ que contiene a $\mathcal{L}(\mathbf{X})$. Por tanto, en general

$$\text{Var}(\mathbf{U} \mid \mathbf{X}) = \sigma_*^2 \geq \sigma^2 = \text{Var}(\mathbf{U} \mid \mathbf{X}; \mathbf{Z}).$$

Ahora bien, *si se cumplen los supuestos*, y consecuentemente \mathbf{U} es ortogonal a todos los regresores, entonces ambas proyecciones son la variable $\mathbf{0}$, es decir, $E(\mathbf{U} \mid \mathbf{X}) = E(\mathbf{U} \mid \mathbf{X}; \mathbf{Z}) = \mathbf{0}$ y por tanto, *el cuadrado de la distancia de \mathbf{U} a*

dichas proyecciones no cambia: $\sigma_*^2 = \sigma^2$. En ese caso

$$\text{Var}(\widehat{\beta}^* | \mathbf{X}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Var}(\mathbf{Z}\gamma | \mathbf{X}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} + \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}; \quad (65)$$

Y ahora la cuestión es: ¿es la varianza del estimador MCO para el modelo incompleto, $\text{Var}(\widehat{\beta}^* | \mathbf{X})$, mayor o menor que la varianza para el modelo completo?

$$\text{Var} \left(\begin{array}{c|c} \widehat{\beta} \\ \hline \widehat{\gamma} \end{array} \middle| \mathbf{X}; \mathbf{Z} \right) = \sigma^2 \begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{Z} \\ \mathbf{Z}^\top \mathbf{X} & \mathbf{Z}^\top \mathbf{Z} \end{bmatrix}^{-1}.$$

(donde, como la matriz no es diagonal por bloques, tenemos que $\text{Var}(\widehat{\beta} | \mathbf{X}; \mathbf{Z}) \neq \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$).

Nótese que el vector del estimador MCO del modelo completo tiene $k + p$ componentes (k en $\widehat{\beta}$ y p en $\widehat{\gamma}$), pero lo que queremos comparar es la varianza del estimador de los k primeros parámetros del modelo completo $\text{Var}(\widehat{\beta} | \mathbf{X}; \mathbf{Z})$ con el estimador del modelo incompleto $\text{Var}(\widehat{\beta}^* | \mathbf{X})$ (donde el vector γ es ignorado). Y no puede ignorarse que en el modelo completo los parámetros β son estimados conjuntamente con los parámetros γ , mientras que en el modelo incompleto se ignora completamente el efecto de los regresores \mathbf{Z} (así que no está claro cómo comparar).

Un intento de comparación puede consistir en “ortogonalizar” los regresores del modelo completo y así poder estimar los parámetros β descontando el efecto de los regresores \mathbf{Z} (de manera similar a como se hizo en la clase de la lección anterior sobre multicolinealidad no estricta). Eso es equivalente a contemplar únicamente el efecto de \mathbf{X} que es ortogonal a \mathbf{Z} , es decir, proyectar sobre $\mathcal{L}(\mathbf{MX})$, donde $\mathbf{M} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$; y por tanto comparar $\text{Var}(\widehat{\beta}^* | \mathbf{X})$ con la matriz:⁸⁴

$$\text{Var}(\widehat{\beta} | \mathbf{MX}) = \sigma^2 (\mathbf{X}^\top \mathbf{MX})^{-1},$$

o sea

$$\text{Var}(\widehat{\beta} | \mathbf{MX}) = \sigma^2 (\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X})^{-1}. \quad (66)$$

Veamos primero que la varianza de la Ecuación (66) es siempre mayor o igual que el último sumando de la Ecuación (65). Para ello bastaría comprobar que la diferencia

$$\text{Var}(\widehat{\beta} | \mathbf{MX}) - \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

es semidefinida positiva, pero eso es equivalente a ver que la diferencia de sus inversas (con el orden cambiado) también es semidefinida positiva:

$$\begin{aligned} \left(\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \right)^{-1} - \left(\text{Var}(\widehat{\beta} | \mathbf{MX}) \right)^{-1} &= \sigma^{-2} \mathbf{X}^\top \mathbf{X} - \sigma^{-2} (\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}) \\ &= \sigma^{-2} \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} \end{aligned} \quad (67)$$

que efectivamente es semidefinda positiva ya que $(\mathbf{Z}^\top \mathbf{Z})^{-1}$ es definida positiva, por ser la inversa de la matriz definida positiva $\mathbf{Z}^\top \mathbf{Z}$ y, por tanto, $\mathbf{v}^\top \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} \mathbf{v} = (\mathbf{Z}^\top \mathbf{X} \mathbf{v})^\top (\mathbf{Z}^\top \mathbf{Z})^{-1} (\mathbf{Z}^\top \mathbf{X} \mathbf{v}) \geq 0$ por ser una suma de componentes al cuadrado.

Y ahora analicemos los posibles casos: en el caso de que el primer sumando de la Ecuación 65

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Var}(\mathbf{Z}\gamma | \mathbf{X}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

sea suficientemente pequeño, es decir, menor que la diferencia (67), entonces el estimador del modelo restringido tendrá una matriz de varianzas menor,

$$\text{Var}(\widehat{\beta}^* | \mathbf{X}) < \text{Var}(\widehat{\beta} | \mathbf{MX})$$

pero en caso contrario, será mayor

$$\text{Var}(\widehat{\beta}^* | \mathbf{X}) > \text{Var}(\widehat{\beta} | \mathbf{MX})$$

Así pues, cabe la posibilidad de que la precisión del estimador para algunos parámetros sea mayor y para otros menor, para todos mayor, o para todos menor. Es decir, a priori no se puede saber qué estimadores son más precisos al omitir regresores.⁸⁵

⁸⁴nótese que si las matrices \mathbf{X} y \mathbf{Z} son ortogonales, es decir, si $\mathbf{Z}^\top \mathbf{X} = \mathbf{0}$, entonces $\text{Var}(\widehat{\beta} | \mathbf{MX}) = \text{Var}(\widehat{\beta} | \mathbf{X}; \mathbf{Z})$, pues la matriz de más arriba sería diagonal por bloques.

⁸⁵Ojo! todos los libros de texto que he consultado afirman que el estimador del modelo incompleto siempre tiene una varianza menor

Estimación de σ^2 . Además, hay que advertir que al estimar σ^2 con el modelo incompleto, el estimador de la cuasivarianza muestral de los errores $(\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^*)/(N - k)$ del modelo incompleto tiene un sesgo al alza, pues

$$\widehat{\mathbf{e}}^* = \mathbf{MY} = \mathbf{M}(\mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{U}) = \mathbf{MZ}\gamma + \mathbf{MU}$$

donde $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$; y como

$$E(\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^*) = \gamma \mathbf{Z}^\top \mathbf{M} \mathbf{Z} \gamma + (N - k)\sigma^2$$

donde el primer sumando es positivo por ser suma de elementos al cuadrado, concluimos que el estimador $(\widehat{\mathbf{e}}^* \cdot \widehat{\mathbf{e}}^*)/(N - k)$ está sesgado por el término $\frac{\gamma \mathbf{Z}^\top \mathbf{M} \mathbf{Z} \gamma}{N - k} \geq 0$; que solo será cero si $\gamma = \mathbf{0}$, o si $\mathbf{MZ} = \mathbf{0}$.

(Lección 13)

T-5 Consecuencias de la omisión de variables del modelo

Si la variable omitida es ortogonal al resto de regresores

- No se introducen sesgos

☞ Código: [OmissionRegresorOrtogonal.inp](#)

Si la variable omitida no es ortogonal al resto de regresores

- **Sesgos:** Si no es ortogonal, se incumple el Supuesto 2.
Regresores no exógenos \Rightarrow MCO sesgado.
- Efecto indeterminado en la varianza... pero
- **Disminución de la varianza** si había elevada correlación con el resto de regresores (multicolinealidad).

☞ Código: [OmissionRegresorNoOrtogonal.inp](#)

☞ Código: [OmissionRegresorMulticol.inp](#)

F156

31 Incorrecta especificación por inclusión errónea de variables

(Lección 13)

T-6 Errores de especificación: Inclusión errónea variables

Sea el MLG

$$\mathbf{Y} = \mathbf{X}\beta^* + \mathbf{U}^*$$

donde $\mathbf{X} = [\mathbf{X}_1; \dots; \mathbf{X}_k]$ y sea $\mathbf{Z} = [\mathbf{Z}_1; \dots; \mathbf{Z}_p]$

1. Correcto: $\mathbf{Y} = \mathbf{X}\beta^* + \mathbf{U}^*$
2. Incorrecto: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{U}$

Sea $\mathbf{Y} = \mathbf{X}\beta^* + \mathbf{U}^*$ un *m.a.s* que cumple los supuestos

¿Qué cabe esperar respecto a la esperanza y varianza de los estimadores en el modelo incorrecto?

Si se cumplen los supuestos en el modelo incorrecto: $\hat{\gamma}$ insesgado.

¿Pero si $E(\mathbf{U} | \mathbf{X}; \mathbf{Z}) \neq \mathbf{0}$?

F157

o igual a la del estimador del modelo completo. Con la discusión de más arriba no encuentro un argumento que soporte lo que dicen los textos... pero además, los ejercicios de simulación desmienten que la varianza siempre sea menor o igual a la del estimador del modelo completo (véase la siguiente transparencia).

1. Correcto: $\mathbf{Y} = \beta_1^* \mathbf{1} + \beta_2^* \mathbf{X} + \mathbf{U}^*$

2. Incorrecto: $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \beta_3 \mathbf{Z} + \mathbf{U}$

¿Qué cabe esperar respecto a $\mathbb{E}(\widehat{\boldsymbol{\beta}} | \mathbf{X}; \mathbf{Z})$ y a $\text{Var}(\widehat{\boldsymbol{\beta}} | \mathbf{X}; \mathbf{Z})$ en el modelo incorrecto?

Si se cumplen los supuestos: $\widehat{\boldsymbol{\beta}}$ insesgado

(en concreto $\mathbb{E}(\widehat{\beta}_3) ?$) ¿Pero si $\mathbb{E}(\mathbf{U} | \mathbf{Z}) \neq \mathbf{0}$?

F158

Los estimadores MCO son insesgados si se cumplen los supuestos, por lo que $\mathbb{E}(\widehat{\beta}_3) = 0$.

EJERCICIO 2. ¿Qué efecto tiene la inclusión de regresores correlados con las perturbaciones sobre $\mathbb{E}(\widehat{\boldsymbol{\beta}} | \mathbf{X}; \mathbf{Z})$?

De 68 en la página 153, sabemos que *si se cumplen los supuestos*, para el modelo correcto:

$$\text{Var}(\widehat{\beta}_2^*) = \frac{\sigma_*^2}{\text{Var}(\mathbf{X})} \quad \text{donde } \sigma_*^2 = \text{Var}(\mathbf{U}^* | \mathbf{X}),$$

para el incorrecto (asumiendo $\mathbf{Z} \perp \mathbf{U}$) se puede demostrar que

$$\text{Var}(\widehat{\beta}_2) = \frac{\sigma^2}{(1 - R_{\mathbf{X}\mathbf{Z}}^2) \text{Var}(\mathbf{X})} \quad \text{donde } \sigma^2 = \text{Var}(\mathbf{U} | \mathbf{X}; \mathbf{Z}),$$

donde $R_{\mathbf{X}\mathbf{Z}}^2$ es el R^2 de la regresión lineal simple de \mathbf{X} sobre \mathbf{Z}

(el cuadrado de $\text{Corr}(\mathbf{X}, \mathbf{Z})$).

- ¿Qué pasa cuando $R_{\mathbf{X}\mathbf{Z}}^2 = 0$?
- ¿Qué pasa cuando $R_{\mathbf{X}\mathbf{Z}}^2$ es casi 1?
- ¿Qué varianza es mayor?

¿Y si no es cierto que $\mathbf{Z} \perp \mathbf{U}$?

P-9
(71)

F159

- Consideremos el caso en el que la variable \mathbf{Z} , incorrectamente incluida, es perpendicular a \mathbf{U} . Entonces el estimador es *insesgado*; pero además
 - si la correlación entre \mathbf{X} e \mathbf{Y} es nula la inclusión de la nueva variable no tendrá ningún efecto.
 - pero a medida que la correlación entre regresores crece en valor absoluto, la varianza del estimador se incrementa; disminuyendo consiguientemente la precisión de los estimadores.
- Consideremos el caso en el que la variable \mathbf{Z} , incorrectamente incluida, es NO perpendicular a \mathbf{U} . Entonces el estimador es *sesgado*;
 - pero a medida que la correlación entre regresores crece en valor absoluto, la varianza del estimador se incrementa; disminuyendo consiguientemente la precisión de los estimadores.
 - Aunque como $\mathcal{L}(\mathbf{X}) \subset \mathcal{L}(\mathbf{X}; \mathbf{Z})$ resulta que $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma_*^2 \geq \sigma^2 \text{Var}(\mathbf{U} | \mathbf{X}; \mathbf{Z})$.

Así, a priori es imposible saber cual de los efectos (correlación entre regresores o disminución de la varianza de las perturbaciones dominará).

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$ cumple supuestos y $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2$

Pero estimamos $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{U}$ con $\mathbf{Z} \perp \mathbf{X}$

Si $\mathbb{E}(\mathbf{U} | \mathbf{X}, \mathbf{Z}) = \mathbf{0}$ (es decir $\mathbf{Z} \perp \mathbf{U}$):

- $E(\hat{\beta}) = \beta$ y $E(\hat{\gamma}) = \mathbf{0}$.
- No hay efectos en la varianza

Si $\mathbb{E}(\mathbf{U} | \mathbf{Z}) \neq \mathbf{0}$ (es decir $\mathbf{Z} \not\perp \mathbf{U}$):

- $E(\hat{\beta}) = \beta$ y $E(\hat{\gamma}) \neq \mathbf{0}$.
- Disminución de la varianza por $\text{Var}(\mathbf{U} | \mathbf{X}, \mathbf{Z}) \leq \sigma^2$
(Test de hipótesis)

F160

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$ cumple supuestos y $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2$

Pero estimamos $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{U}$ con $\mathbf{Z} \not\perp \mathbf{X}$

Si $\mathbb{E}(\mathbf{U} | \mathbf{X}, \mathbf{Z}) = \mathbf{0}$ (es decir $\mathbf{Z} \perp \mathbf{U}$):

- $E(\hat{\beta}) = \beta$ y $E(\hat{\gamma}) = \mathbf{0}$.
- Incremento de la varianza por $\mathbf{Z} \not\perp \mathbf{X}$.
- Este caso se denomina *inclusión de variables irrelevantes*

Si $\mathbb{E}(\mathbf{U} | \mathbf{Z}) \neq \mathbf{0}$ (es decir $\mathbf{Z} \not\perp \mathbf{U}$):

- $E(\hat{\beta}) \neq \beta$ y $E(\hat{\gamma}) \neq \mathbf{0}$.
- Disminución de la varianza por $\text{Var}(\mathbf{U} | \mathbf{X}, \mathbf{Z}) \leq \sigma^2$
Incremento de la varianza por $\mathbf{Z} \not\perp \mathbf{X}$.

Efecto final?

☞ Código: ErrorDeEspecificacionPorInclusion.inp

F161

Los libros de texto que he consultado no realizan un estudio de casos como el de más arriba. Directamente afirman que la inclusión de *variables irrelevantes* tan solo supone un incremento de la varianza del estimador. Por ello, al caso en el que ocurre precisamente eso lo he denominado *inclusión de variables irrelevantes* (tal como lo denominan los libros). Pero he de advertir que es un nombre confuso, pues hay otros casos de incorrecta especificación por inclusión de variables que tienen efectos muy distintos (introducción de sesgos y/o disminución de la varianza).

Problemas de la Lección 13

Propiedades algebraicas del ajuste MCO

(L-13) PROBLEMA 1. Sea el MLG $\mathbf{Y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{X} + \beta_3 \mathbf{Z} + \mathbf{U}$ con $\beta_3 \neq 0$. Si se omite \mathbf{Z} y se estima el modelo

$$\mathbf{Y} = \beta_1^* \mathbf{1} + \beta_2^* \mathbf{X} + \mathbf{U}^*$$

con $\mathbf{X} = [\mathbf{1}; \mathbf{X}]$ m.a.s de $\mathbf{X} = [\mathbf{1}; \mathbf{X}]$; demuestre que

$$E\left(\begin{pmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{pmatrix}\right) = E\left(\mathbb{E}\left(\begin{pmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{pmatrix} \mid \mathbf{X}\right)\right) = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} E(\mathbf{Z})\beta_3 \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})}\beta_3 \end{pmatrix};$$

es decir, $E(\widehat{\beta}_1^*) = \beta_1 + E(\mathbf{Z})\beta_3$ y $E(\widehat{\beta}_2^*) = \beta_2 + \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})}\beta_3$.

Fin de los Problemas de la Lección 13

Fin de la lección

References

- Bujosa, M. (2022a). *Un Curso de Álgebra Lineal con notación asociativa y un módulo para Python.* <https://github.com/mbujosab> (self-published).
- URL <https://github.com/mbujosab/CursoDeAlgebraLineal>
- Bujosa, M. (2022b). *Un Curso de Álgebra Lineal con notación asociativa y un módulo para Python.* edición propia.
- URL <https://github.com/mbujosab/CursoDeAlgebraLineal>
- Golovina, L. I. (1980). *Algebra lineal y algunas de sus aplicaciones.* MIR, Moscú, second ed.
- Gujarati, D. N. (2003). *Basic Econometrics.* McGraw-Hill, fourth ed. ISBN 0-07-112342-3. International edition.
- Hayashi, F. (2000). *Econometrics.* Princeton University Press, Princeton, New Jersey. ISBN 0-691-01018-8.
- Johnston, J. and Dinardo, J. (2001). *Métodos de Econometría.* Vicens Vives, Barcelona, España, first ed. ISBN 84-316-6116-x.
- Mittelhammer, R. C. (1996). *Mathematical Statistics for Economics and Business.* Springer-Verlag, New York, first ed. ISBN 0-387-94587-3.
- Novales, A. (1993). *Econometría.* McGraw-Hill, second ed.
- Novales, A. (1997). *Estadística y Econometría.* McGraw-Hill, Madrid, first ed. ISBN 84-481-0798-5.
- Ramanathan, R. (2002). *Introductory Econometrics with applications.* South-Western, Mason, Ohio, fifth ed. ISBN 0-03-034186-8.
- Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data, second edition.* The MIT Press. MIT Press. ISBN 9780262296793.
- Wooldridge, J. M. (2006). *Introducción a la econometría. Un enfoque moderno.* Thomson Learning, Inc., second ed.

Soluciones

(L-2) **Problema 1.** $\langle c|c\rangle = \langle a+b|a+b\rangle = \langle a|a+b\rangle + \langle b|a+b\rangle = \langle a|a\rangle + \underbrace{\langle a|b\rangle}_{=0} + \underbrace{\langle b|a\rangle}_{=0} + \langle b|b\rangle = \langle a|a\rangle + \langle b|b\rangle.$

□

(L-2) **Problema 2.** Como $\sigma_{xy} = \mu_{x \odot (y - \bar{y})}$ y como $\mu_{x \odot \bar{y}} = \mu_{x \odot 1 \mu_y} = (\mu_{x \odot 1})\mu_y = \mu_x \mu_y$,

$$\sigma_{xy} = \mu_{x \odot (y - \bar{y})} = \mu_{x \odot y} - \mu_{x \odot \bar{y}} = \mu_{x \odot y} - \mu_x \mu_y.$$

□

(L-2) **Problema 3.** Como $\bar{\hat{y}}$ es la proyección ortogonal de \hat{y} sobre $\mathbf{1}$ y la proyección ortogonal es una función lineal

$$\bar{\hat{y}} = \overline{(\hat{a}\mathbf{1} + \hat{b}\bar{x})} = \hat{a}\bar{\mathbf{1}} + \hat{b}\bar{x}.$$

□

(L-2) **Problema 4.** Como $\hat{y} = \hat{a}\mathbf{1} + \hat{b}\bar{x}$ y $\bar{\hat{y}} = \hat{a}\bar{\mathbf{1}} + \hat{b}\bar{x}$, tenemos que $\hat{y} - \bar{\hat{y}} = \hat{b}\bar{x} - \hat{b}\bar{x}$, así

$$\sigma_{\hat{y}}^2 = \|\hat{y} - \bar{\hat{y}}\|_s^2 = \|\hat{b}\bar{x} - \hat{b}\bar{x}\|_s^2 = \sigma_{(\hat{b}\bar{x})}^2 = \hat{b}^2(\sigma_x^2).$$

□

(L-2) **Problema 5.** Como $\hat{y} = \hat{a}\mathbf{1} + \hat{b}\bar{x}$, y como $(y - \bar{y})$ es ortogonal a los vectores constantes (los múltiplos de $\mathbf{1}$),

$$\sigma_{yy} = \mu_{((y - \bar{y}) \odot \hat{y})} = \mu_{((y - \bar{y}) \odot (\hat{a}\mathbf{1} + \hat{b}\bar{x}))} = \mu_{((y - \bar{y}) \odot (\hat{b}\bar{x}))} = \hat{b}\mu_{((y - \bar{y}) \odot \bar{x})} = \hat{b}(\sigma_{x\bar{x}}).$$

□

(L-2) **Problema 6.** Dividiendo ambos lados del sistema de ecuaciones normales por N tenemos

$$\begin{bmatrix} N^{-1}(\mathbf{1} \cdot \mathbf{1}) & N^{-1}(\mathbf{1} \cdot x) \\ N^{-1}(x \cdot \mathbf{1}) & N^{-1}(x \cdot x) \end{bmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} N^{-1}(\mathbf{1} \cdot y) \\ N^{-1}(x \cdot y) \end{pmatrix}; \text{ es decir } \begin{bmatrix} 1 & \mu_{(x)} \\ \mu_{(x)} & \mu_{(x^2)} \end{bmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \mu_{(y)} \\ \mu_{(x \odot y)} \end{pmatrix};$$

que resolvemos por eliminación (recordando que $\sigma_x^2 = \mu_{(x^2)} - \mu_x^2$ y que $\sigma_{xy} = \mu_{x \odot y} - \mu_x \mu_y$):

$$\left[\begin{array}{cc|c} 1 & \mu_{(x)} & -\mu_{(y)} \\ \mu_{(x)} & \mu_{(x^2)} & -\mu_{(x \odot y)} \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array} \right] \xrightarrow{\substack{[(-\mu_{(x)})\mathbf{1}+2] \\ [(\mu_{(y)})\mathbf{1}+3]}} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ \mu_{(x)} & \sigma_x^2 & -\sigma_{xy} \\ \hline 1 & -\mu_{(x)} & \mu_{(y)} \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array} \right] \xrightarrow{\left[\begin{array}{c} \frac{\sigma_{(x)y}}{\sigma_x^2} \\ 2+3 \end{array} \right]} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ \mu_{(x)} & \sigma_x^2 & 0 \\ \hline 1 & -\mu_{(x)} & -\frac{\sigma_{(x)y}\mu_{(x)}}{\sigma_x^2} + \mu_{(y)} \\ 0 & 1 & \frac{\sigma_{(x)y}}{\sigma_x^2} \\ \hline 0 & 0 & 1 \end{array} \right]$$

Por tanto, la combinación lineal de $\mathbf{1}$ y x más próxima a y es:

$$\hat{y} = \underbrace{\left(\mu_y - \mu_x \frac{\sigma_{(x)y}}{\sigma_x^2} \right)}_{\hat{a}} \mathbf{1} + \underbrace{\left(\frac{\sigma_{(x)y}}{\sigma_x^2} \right)}_{\hat{b}} x = \underbrace{\left(\mu_y - \mu_x \hat{b} \right)}_{\hat{a}} \mathbf{1} + \underbrace{\left(\frac{\sigma_{(x)y}}{\sigma_x^2} \right)}_{\hat{b}} x$$

□

(L-2) **Problema 7.** Entonces la condición de rango sobre la matriz de regresores \mathbf{X} no se cumpliría, pues la segunda columna x sería un múltiplo de la primera columna de unos: $x = c \mathbf{1}$.

En tal situación el sistema de ecuaciones normales se reduciría a:

$$\begin{bmatrix} (\mathbf{1} \cdot \mathbf{1}) & c(\mathbf{1} \cdot \mathbf{1}) \\ c(\mathbf{1} \cdot \mathbf{1}) & c^2(\mathbf{1} \cdot \mathbf{1}) \end{bmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} (\mathbf{1} \cdot \mathbf{y}) \\ c(\mathbf{1} \cdot \mathbf{y}) \end{pmatrix};$$

donde la segunda ecuación es c veces la primera, por lo que realmente tenemos más incógnitas que ecuaciones linealmente independientes (*situación de multicolinealidad perfecta*).

Además, la varianza de un vector constante, $\mathbf{x} = c\mathbf{1}$, es cero, por lo que $\sigma_{\mathbf{x}}^2 = 0$ y como $(\mathbf{y} - \bar{\mathbf{y}}) \perp \mathbf{1}$, también tenemos que $\sigma_{\mathbf{x}\mathbf{y}} = \mu_{((\mathbf{y} - \bar{\mathbf{y}}) \odot \mathbf{1})} = 0$; así que la expresión $\hat{b} = \frac{\sigma_{\mathbf{x}\mathbf{y}}}{\sigma_{\mathbf{x}}^2} = \frac{0}{0}$ carece de sentido.

□

(L-2) Problema 8. El sistema de ecuaciones normales es

$$\begin{bmatrix} (\mathbf{1} \cdot \mathbf{1}) & (\mathbf{1} \cdot \mathbf{x}) & (\mathbf{1} \cdot \mathbf{z}) \\ (\mathbf{1} \cdot \mathbf{x}) & (\mathbf{x} \cdot \mathbf{x}) & (\mathbf{x} \cdot \mathbf{z}) \\ (\mathbf{1} \cdot \mathbf{z}) & (\mathbf{z} \cdot \mathbf{x}) & (\mathbf{z} \cdot \mathbf{z}) \end{bmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix} = \begin{pmatrix} (\mathbf{1} \cdot \mathbf{y}) \\ (\mathbf{x} \cdot \mathbf{y}) \\ (\mathbf{z} \cdot \mathbf{y}) \end{pmatrix}, \text{ o dividiendo por N: } \begin{bmatrix} 1 & \mu_{\mathbf{x}} & \mu_{\mathbf{z}} \\ \mu_{\mathbf{x}} & \mu_{(\mathbf{x}^2)} & \mu_{\mathbf{x} \odot \mathbf{z}} \\ \mu_{\mathbf{z}} & \mu_{\mathbf{x} \odot \mathbf{z}} & \mu_{(\mathbf{z}^2)} \end{bmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix} = \begin{pmatrix} \mu_{\mathbf{y}} \\ \mu_{\mathbf{x} \odot \mathbf{y}} \\ \mu_{\mathbf{z} \odot \mathbf{y}} \end{pmatrix}.$$

□

(L-2) Problema 9. Resolvemos por eliminación simplificando las expresiones de varianzas y covarianzas a medida que aparecen:

$$\begin{array}{c} \left[\begin{array}{ccc|c} (\mathbf{1} \cdot \mathbf{1}) & (\mathbf{1} \cdot \mathbf{x}) & (\mathbf{1} \cdot \mathbf{z}) & -\mu_{\mathbf{y}} \\ (\mathbf{1} \cdot \mathbf{x}) & (\mathbf{x} \cdot \mathbf{x}) & (\mathbf{x} \cdot \mathbf{z}) & -\mu_{\mathbf{x} \odot \mathbf{y}} \\ (\mathbf{1} \cdot \mathbf{z}) & (\mathbf{z} \cdot \mathbf{x}) & (\mathbf{z} \cdot \mathbf{z}) & -\mu_{\mathbf{z} \odot \mathbf{y}} \\ \hline 1 & \mu_{\mathbf{x}} & \mu_{\mathbf{z}} & 0 \\ \mu_{\mathbf{x}} & \mu_{(\mathbf{x}^2)} & \mu_{\mathbf{x} \odot \mathbf{z}} & 0 \\ \mu_{\mathbf{z}} & \mu_{\mathbf{x} \odot \mathbf{z}} & \mu_{(\mathbf{z}^2)} & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right] \xrightarrow{\left[\begin{array}{c} \left[\begin{array}{c} (-\mu_{\mathbf{x}}) \mathbf{1+2} \\ (-\mu_{\mathbf{z}}) \mathbf{1+3} \\ (\mu_{\mathbf{y}}) \mathbf{1+4} \\ \left(-\frac{\sigma_{\mathbf{x}\mathbf{z}}}{\sigma_{\mathbf{x}}^2} \right) \mathbf{2+3} \\ \left(\frac{\sigma_{\mathbf{x}\mathbf{y}}}{\sigma_{\mathbf{x}}^2} \right) \mathbf{2+4} \\ \left(\frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{x}\mathbf{y}} - \sigma_{\mathbf{z}\mathbf{y}}\sigma_{\mathbf{x}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \right) \mathbf{3+4} \end{array} \right] } \\ \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ \mu_{\mathbf{x}} & \sigma_{\mathbf{x}}^2 & 0 & 0 \\ \mu_{\mathbf{z}} & (\sigma_{\mathbf{x}\mathbf{z}}) & -\frac{(\sigma_{\mathbf{x}\mathbf{z}})^2}{\sigma_{\mathbf{x}}^2} + \sigma_{\mathbf{z}}^2 & 0 \\ \hline 1 & -\mu_{\mathbf{x}} & \frac{(\sigma_{\mathbf{x}\mathbf{z}})\mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}^2} - \mu_{\mathbf{z}} & \frac{(\sigma_{\mathbf{x}\mathbf{z}})^2\mu_{\mathbf{y}} - (\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{x}\mathbf{y}}\mu_{\mathbf{z}} - (\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{z}\mathbf{y}}\mu_{\mathbf{x}} + \sigma_{\mathbf{x}\mathbf{y}}\mu_{\mathbf{x}}\sigma_{\mathbf{z}}^2 + \sigma_{\mathbf{z}\mathbf{y}}\mu_{\mathbf{z}}\sigma_{\mathbf{x}}^2 - \mu_{\mathbf{y}}\sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \\ 0 & 1 & -\frac{(\sigma_{\mathbf{x}\mathbf{z}})}{\sigma_{\mathbf{x}}^2} & \frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{z}\mathbf{y}} - \sigma_{\mathbf{x}\mathbf{y}}\sigma_{\mathbf{z}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \\ 0 & 0 & 1 & \frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{x}\mathbf{y}} - \sigma_{\mathbf{z}\mathbf{y}}\sigma_{\mathbf{x}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \\ \hline 0 & 0 & 0 & 1 \end{array} \right] \end{array}$$

Es decir, la combinación lineal de $\mathbf{1}$, \mathbf{x} y \mathbf{z} más próxima a \mathbf{y} es:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\left(\mu_{\mathbf{y}} - \mu_{\mathbf{x}} \hat{b} - \mu_{\mathbf{z}} \hat{c} \right)}_{\hat{a}} \mathbf{1} + \underbrace{\left(\frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{z}\mathbf{y}} - \sigma_{\mathbf{x}\mathbf{y}}\sigma_{\mathbf{z}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \right)}_{\hat{b}} \mathbf{x} + \underbrace{\left(\frac{(\sigma_{\mathbf{x}\mathbf{z}})\sigma_{\mathbf{x}\mathbf{y}} - \sigma_{\mathbf{z}\mathbf{y}}\sigma_{\mathbf{x}}^2}{(\sigma_{\mathbf{x}\mathbf{z}})^2 - \sigma_{\mathbf{x}}^2\sigma_{\mathbf{z}}^2} \right)}_{\hat{c}} \mathbf{z}.$$

□

(L-2) Problema 10. Si $\sigma_{\mathbf{x}\mathbf{z}} = 0$, la estimación \hat{b} en el modelo con tres regresores, $\hat{\mathbf{y}} = \hat{a}\mathbf{1} + \hat{b}\mathbf{x} + \hat{c}\mathbf{z}$, coincide exactamente con \hat{b} en el modelo lineal simple, $\hat{\mathbf{y}} = \hat{a}\mathbf{1} + \hat{b}\mathbf{x}$, donde no aparece el tercer regresor \mathbf{z} .

□

(L-2) Problema 11. Un coeficiente de correlación con valor absoluto igual a uno significa que hay una dependencia lineal entre los regresores, por lo que la condición sobre el rango de la matriz $\mathbf{X}^T \mathbf{X}$ deja de cumplirse; y por tanto el sistema de ecuaciones normales tiene infinitas soluciones.

En tal caso las expresiones (20) y (21) dejan de estar definidas (y por tanto también (19)). Veámoslo. Como

$$|\rho_{xz}| = \left| \frac{\sigma_{xz}}{\sigma_x \sigma_z} \right| = 1, \quad \text{entonces} \quad |\sigma_{xz}| = |\sigma_x \sigma_z|;$$

y por tanto $(\sigma_{xz})^2 = \sigma_x^2 \sigma_z^2$; así que los denominadores de las expresiones (20) y (21) son cero.

□

(L-3) Problema 1. Dado que la media es un operador lineal, $\mu_y = \mu_{(\hat{y} + \hat{e})} = \mu_{\hat{y}} + \mu_{\hat{e}} = \mu_{\hat{y}} + 0$.

□

(L-3) Problema 2. Sustituyendo en $\hat{y} \cdot y$ el vector y por su descomposición ortogonal y recordando que \hat{e} es ortogonal a \hat{y} tenemos: $\hat{y} \cdot y = \hat{y} \cdot (\hat{y} + \hat{e}) = \hat{y} \cdot \hat{y} + \hat{y} \cdot \hat{e} = \hat{y} \cdot \hat{y} + 0$.

□

(L-3) Problema 3. $SEC = \|\hat{y} - \bar{y}\|_u^2$, pero en este caso (Ecuación 10 en la página 17) los valores ajustados \hat{y} son el vector constante \bar{y} ; por tanto $SEC = \|\mathbf{0}\|_u^2 = 0$ y consecuentemente $R^2 = 0$. Es decir, un modelo que consiste únicamente en un constante, no tiene ninguna capacidad de “explicar” las variaciones de la variable dependiente.

□

(L-4) Problema 1. Como $E((\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \mathbf{1}))) = E((\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})) \cdot \mathbf{Y}) - E((\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})) \cdot \mathbb{E}(\mathbf{Y} | \mathbf{1}))$, pues la esperanza es una función lineal. Y como las componentes variables son ortogonales a las componentes constantes:

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = E((\mathbf{X} - \mathbb{E}(\mathbf{X} | \mathbf{1})) \cdot \mathbf{Y}) = E(\mathbf{X} \cdot \mathbf{Y} - \mathbb{E}(\mathbf{X}) \cdot \mathbf{1} \cdot \mathbf{Y}) = E(\mathbf{X} \cdot \mathbf{Y}) - E(\mathbf{X}) E(\mathbf{Y}),$$

ya que $\mathbf{1} \cdot \mathbf{Y} = \mathbf{Y}$ y que $E(\mathbf{X})$ es un número.

□

(L-4) Problema 2. Como en el ejercicio anterior, por ser la esperanza una función lineal, y las componentes variables ortogonales a las componentes constantes

$$\begin{aligned} \text{Cov}(a\mathbf{X} + b\mathbf{1}, c\mathbf{Y} + d\mathbf{1}) &= E((a\mathbf{X} + b\mathbf{1}) - \mathbb{E}(a\mathbf{X} + b\mathbf{1} | \mathbf{1})) \cdot (c\mathbf{Y} + d\mathbf{1}) \\ &= c E((a\mathbf{X} + b\mathbf{1}) - \mathbb{E}(a\mathbf{X} + b\mathbf{1} | \mathbf{1})) \cdot \mathbf{Y} && \text{por el mismo motivo} \\ &= c \text{Cov}((a\mathbf{X} + b\mathbf{1}), \mathbf{Y}) \\ &= c \text{Cov}(a\mathbf{X}, \mathbf{Y}) = ac \text{Cov}(\mathbf{X}, \mathbf{Y}) \end{aligned}$$

□

(L-4) Problema 3.

$$\begin{aligned} \text{Var}(Y | \mathbf{X}) &= \mathbb{E}((Y - \mathbb{E}(Y | \mathbf{X}))^2 | \mathbf{X}) \\ &= \mathbb{E}((Y - \mathbb{E}(Y | \mathbf{X})) \cdot (Y - \mathbb{E}(Y | \mathbf{X})) | \mathbf{X}) \\ &= \mathbb{E}(Y \cdot (Y - \mathbb{E}(Y | \mathbf{X})) | \mathbf{X}) \\ &= \mathbb{E}(Y^2 - Y \cdot \mathbb{E}(Y | \mathbf{X}) | \mathbf{X}) \\ &= \mathbb{E}(Y^2 | \mathbf{X}) - \mathbb{E}(Y | \mathbf{X}) \mathbb{E}(Y | \mathbf{X}) \\ &= \mathbb{E}(Y^2 | \mathbf{X}) - (\mathbb{E}(Y | \mathbf{X}))^2 \end{aligned}$$

□

(L-5) Problema 1. Primero una demostración ligada a la definición de esperanza condicional como proyección ortogonal: $\mathbb{E}(\mathbf{U}|\mathbf{X}) = \mathbf{0}$ implica que \mathbf{U} es ortogonal a $\mathcal{L}(\mathbf{X})$, pues $\mathbf{U} = \mathbf{U} - \mathbb{E}(\mathbf{U}|\mathbf{X})$; y por tanto \mathbf{U} es ortogonal a $\mathcal{L}(\mathbf{X}) \subset \mathcal{L}(\mathbf{X})$. Así, y en particular, para cada \mathbf{X}_j tenemos que $\langle \mathbf{U} | \mathbf{X}_j \rangle_\eta = \mathbb{E}(\mathbf{U} \mathbf{X}_j) = 0$.

Y ahora la demo habitual en los libros de Econometría: usando el *Teorema de las Esperanzas Iteradas* (Página 46):

$$\mathbb{E}(\mathbf{X}_j \mathbf{U}) = \mathbb{E}(\mathbb{E}(\mathbf{X}_j \mathbf{U} | \mathbf{X})) = \mathbb{E}(X_j \mathbb{E}(\mathbf{U} | \mathbf{X})) = \mathbb{E}(\mathbf{X}_j \mathbf{0}) = \mathbb{E}(\mathbf{0}) = 0, \quad \text{para } j = 1 : k.$$

o, análogamente, usando el producto de un sistema de variables aleatorias por una variable aleatoria (Definición 24):

$$\mathbb{E}(\mathbf{X}^\top \mathbf{U}) = \mathbb{E}(\mathbb{E}(\mathbf{X}^\top \mathbf{U} | \mathbf{X})) = \mathbb{E}(\mathbf{X}^\top \mathbb{E}(\mathbf{U} | \mathbf{X})) = \mathbb{E}(\mathbf{X}^\top \mathbf{0}) = \mathbb{E}(\mathbf{0}) = \mathbf{0} \in \mathbb{R}^k.$$

donde $\mathbf{0} = [\mathbf{0}; \mathbf{0}; \dots; \mathbf{0}]$ es un sistema de variables aleatorias nulas. Por tanto, $\mathbb{E}(\mathbf{j}(\mathbf{X}^\top \mathbf{U})) = \mathbb{E}(\mathbf{X}_j \mathbf{U}) = 0$.

□

(L-5) Problema 2. Como antes, primero una demo más algebraica: $\mathbb{E}(\mathbf{U} | \mathbf{X}) = \mathbf{0}$ implica que \mathbf{U} es ortogonal a $\mathcal{L}(\mathbf{X})$; luego en particular es ortogonal a $\mathbf{1}$, es decir, $\langle \mathbf{U} | \mathbf{1} \rangle_\eta = \mathbb{E}(\mathbf{U}) = 0$.

Ya ahora la demostración que alude al Tma. de Esperanzas Iteradas:

$$\begin{aligned} \mathbb{E}(\mathbf{U}) &= \mathbb{E}(\mathbb{E}(\mathbf{U} | \mathbf{X})) && \text{por el Tma de las esperanzas iteradas.} \\ &= \mathbb{E}(\mathbf{0}) = 0 && \text{puesto que } \mathbb{E}(\mathbf{U} | \mathbf{X}) = \mathbf{0}. \end{aligned}$$

□

(L-5) Problema 3. Como antes, primero una demo más algebraica. Como tanto \mathbf{X}_j como $\mathbb{E}(\mathbf{U} | \mathbf{1})$ pertenecen a $\mathcal{L}(\mathbf{X})$

$$\text{Cov}(\mathbf{U}, \mathbf{X}_j) = \mathbb{E}\left(\underbrace{[\mathbf{U} - \mathbb{E}(\mathbf{U} | \mathbf{1})]}_{\in \mathcal{L}(\mathbf{X})^\perp} \cdot \underbrace{[\mathbf{X}_j - \mathbb{E}(\mathbf{X}_j | \mathbf{1})]}_{\in \mathcal{L}(\mathbf{X})}\right) = 0$$

Y ahora otra demo distinta: sabemos que \mathbf{U} es ortogonal a los regresores, y también que $\mathbb{E}(\mathbf{U}) = 0$. Por tanto

$$\text{Cov}(\mathbf{U}, \mathbf{X}_j) = \mathbb{E}(\mathbf{U} \mathbf{X}_j) - \mathbb{E}(\mathbf{U}) \mathbb{E}(\mathbf{X}_j) = 0 - 0 \mathbb{E}(\mathbf{X}_j) = 0.$$

□

(L-5) Problema 4. Demostración:

$$\begin{aligned} \mathbb{E}(\mathbf{Y} | \mathbf{X}) &= \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U} | \mathbf{X}) && \text{por el Supuesto 1} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbb{E}(\mathbf{U} | \mathbf{X}) && \text{puesto que } \mathbb{E}(\mathbf{X}\boldsymbol{\beta} | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\boldsymbol{\beta} && \text{por el Supuesto 2.} \end{aligned}$$

□

(L-5) Problema 5. Razonemos a la inversa y veamos que: si \mathbf{X} es linealmente dependiente entonces $\mathbb{E}(\mathbf{X}^\top \mathbf{X})$ es singular.

Si existe un $\mathbf{c} \neq \mathbf{0}$ tal que $\mathbf{X}\mathbf{c} = \mathbf{0}$, entonces el producto de cada fila i -ésima de $\mathbb{E}(\mathbf{X}^\top \mathbf{X})$ por \mathbf{c} es cero:

$${}_{il} \mathbb{E}(\mathbf{X}^\top \mathbf{X}) \mathbf{c} = {}_{il} \mathbb{E}(\mathbf{X}^\top \mathbf{X} \mathbf{c}) = {}_{il} \mathbb{E}(\mathbf{X}^\top \mathbf{0}) = {}_{il} \mathbb{E}(\mathbf{0}^\top) = {}_{il} \mathbf{0} = 0,$$

donde $\mathbf{0} = [\mathbf{0}; \dots; \mathbf{0}]$. Por tanto $\mathbb{E}(\mathbf{X}^\top \mathbf{X}) \mathbf{c} = \mathbf{0}$.

□

$$(L-5) \text{ Problema 6(a)} \text{ La primera: } \frac{\mathbf{x}^\tau \mathbf{x}}{N} = \frac{1}{N} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \frac{1}{N} \begin{bmatrix} (\mathbf{1} \cdot \mathbf{1}) & (\mathbf{1} \cdot \mathbf{x}) \\ (\mathbf{x} \cdot \mathbf{1}) & (\mathbf{x} \cdot \mathbf{x}) \end{bmatrix} = \begin{bmatrix} 1 & m_{\mathbf{x}} \\ m_{\mathbf{x}} & m_{\mathbf{x}^2} \end{bmatrix};$$

... y la segunda... $\frac{\mathbf{x}^\tau \mathbf{y}}{N} = \frac{1}{N} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \frac{1}{N} \begin{pmatrix} (\mathbf{1} \cdot \mathbf{y}) \\ (\mathbf{x} \cdot \mathbf{y}) \end{pmatrix} = \begin{pmatrix} m_{\mathbf{y}} \\ m_{(\mathbf{x} \odot \mathbf{y})} \end{pmatrix}.$

□

(L-5) **Problema 6(b)** Resolviendo por eliminación Gaussiana (por columnas):

$$\left[\begin{array}{cc|c} 1 & m_{(\mathbf{x})} & -m_{(\mathbf{y})} \\ m_{(\mathbf{x})} & m_{(\mathbf{x}^2)} & 0 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array} \right] \xrightarrow{\begin{array}{l} [(-m_{(\mathbf{x})})^{1+2}] \\ [(-m_{(\mathbf{y})})^{1+3}] \end{array}} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ m_{(\mathbf{x})} & s_{(\mathbf{x})}^2 & -s_{(\mathbf{x} \mathbf{y})} \\ \hline 1 & -m_{(\mathbf{x})} & m_{(\mathbf{y})} \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array} \right] \xrightarrow{\begin{array}{l} \left[\left(\frac{s_{(\mathbf{x} \mathbf{y})}}{s_{(\mathbf{x})}^2} \right)^{2+3} \right] \\ \left[\left(\frac{s_{(\mathbf{x} \mathbf{y})}}{s_{(\mathbf{x})}^2} \right)^{2+3} \right] \end{array}} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ m_{(\mathbf{x})} & s_{(\mathbf{x})}^2 & 0 \\ \hline 1 & -m_{(\mathbf{x})} & -\frac{s_{(\mathbf{x} \mathbf{y})} m_{(\mathbf{x})}}{s_{(\mathbf{x})}^2} + m_{(\mathbf{y})} \\ 0 & 1 & \frac{s_{(\mathbf{x} \mathbf{y})}}{s_{(\mathbf{x})}^2} \\ \hline 0 & 0 & 1 \end{array} \right];$$

donde $s_{\mathbf{x}}^2 = m_{\mathbf{x}^2} - (m_{\mathbf{x}})^2$ y $s_{\mathbf{x} \mathbf{y}} = m_{(\mathbf{x} \odot \mathbf{y})} - m_{\mathbf{x}} m_{\mathbf{y}}$. Por tanto, $\hat{a} = m_{\mathbf{y}} - \frac{s_{\mathbf{x} \mathbf{y}}}{s_{\mathbf{x}}^2} m_{\mathbf{x}}$ y $\hat{b} = \frac{s_{\mathbf{x} \mathbf{y}}}{s_{\mathbf{x}}^2}$.

□

(L-5) **Problema 7.** Resolviendo por eliminación Gaussiana (por columnas):

$$\left[\begin{array}{ccc|c} 1 & m_{\mathbf{x}} & m_{\mathbf{z}} & -m_{\mathbf{y}} \\ m_{\mathbf{x}} & m_{(\mathbf{x}^2)} & m_{\mathbf{x} \odot \mathbf{z}} & -m_{\mathbf{x} \odot \mathbf{y}} \\ m_{\mathbf{z}} & m_{\mathbf{x} \odot \mathbf{z}} & m_{(\mathbf{z}^2)} & -m_{\mathbf{z} \odot \mathbf{y}} \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right] \xrightarrow{\begin{array}{l} [(-m_{\mathbf{x}})^{1+2}] \\ [(-m_{\mathbf{z}})^{1+3}] \\ [(m_{\mathbf{y}})^{1+4}] \\ \left[\left(-\frac{(s_{\mathbf{x} \mathbf{z}})}{s_{\mathbf{x}}^2} \right)^{2+3} \right] \\ \left[\left(\frac{s_{\mathbf{x} \mathbf{y}}}{s_{\mathbf{x}}^2} \right)^{2+4} \right] \\ \left[\left(\frac{(s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{x} \mathbf{y}} - s_{\mathbf{z} \mathbf{y}} s_{\mathbf{x}}^2}{(s_{\mathbf{x} \mathbf{z}})^2 - s_{\mathbf{x}}^2 s_{\mathbf{z}}^2} \right)^{3+4} \right] \end{array}} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ m_{\mathbf{x}} & s_{\mathbf{x}}^2 & 0 & 0 \\ m_{\mathbf{z}} & (s_{\mathbf{x} \mathbf{z}}) & -\frac{(s_{\mathbf{x} \mathbf{z}})^2}{s_{\mathbf{x}}^2} + s_{\mathbf{z}}^2 & 0 \\ \hline 1 & -m_{\mathbf{x}} & \frac{(s_{\mathbf{x} \mathbf{z}}) m_{\mathbf{x}}}{s_{\mathbf{x}}^2} - m_{\mathbf{z}} & \frac{(s_{\mathbf{x} \mathbf{z}})^2 m_{\mathbf{y}} - (s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{x} \mathbf{y}} m_{\mathbf{z}} - (s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{z} \mathbf{y}} m_{\mathbf{x}} + s_{\mathbf{x} \mathbf{y}} m_{\mathbf{x}} s_{\mathbf{z}}^2 + s_{\mathbf{z} \mathbf{y}} m_{\mathbf{z}} s_{\mathbf{x}}^2 - m_{\mathbf{y}} s_{\mathbf{x}}^2 s_{\mathbf{z}}^2}{(s_{\mathbf{x} \mathbf{z}})^2 - s_{\mathbf{x}}^2 s_{\mathbf{z}}^2} \\ 0 & 1 & -\frac{(s_{\mathbf{x} \mathbf{z}})}{s_{\mathbf{x}}^2} & \frac{(s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{z} \mathbf{y}} - s_{\mathbf{x} \mathbf{y}} s_{\mathbf{z}}^2}{(s_{\mathbf{x} \mathbf{z}})^2 - s_{\mathbf{x}}^2 s_{\mathbf{z}}^2} \\ 0 & 0 & 1 & \frac{(s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{x} \mathbf{y}} - s_{\mathbf{z} \mathbf{y}} s_{\mathbf{x}}^2}{(s_{\mathbf{x} \mathbf{z}})^2 - s_{\mathbf{x}}^2 s_{\mathbf{z}}^2} \\ \hline 0 & 0 & 0 & 1 \end{array} \right].$$

$$\hat{b} = \frac{(s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{z} \mathbf{y}} - s_{\mathbf{x} \mathbf{y}} s_{\mathbf{z}}^2}{(s_{\mathbf{x} \mathbf{z}})^2 - s_{\mathbf{x}}^2 s_{\mathbf{z}}^2}; \quad \hat{c} = \frac{(s_{\mathbf{x} \mathbf{z}}) s_{\mathbf{x} \mathbf{y}} - s_{\mathbf{z} \mathbf{y}} s_{\mathbf{x}}^2}{(s_{\mathbf{x} \mathbf{z}})^2 - s_{\mathbf{x}}^2 s_{\mathbf{z}}^2}; \quad \hat{a} = \mu_{\mathbf{y}} - \mu_{\mathbf{x}} \hat{b} - \mu_{\mathbf{z}} \hat{c}.$$

□

(L-5) **Problema 8.** Invirtiendo la matriz por el método de Gauss-Jordan

$$\left[\begin{array}{cc} 1 & E(\mathbf{X}) \\ E(\mathbf{X}) & E(\mathbf{X}^2) \\ 1 & 0 \\ 0 & 1 \end{array} \right] \xrightarrow{[(E(\mathbf{X}))\mathbf{1}+2]} \left[\begin{array}{cc} 1 & 0 \\ E(\mathbf{X}) & \text{Var}(\mathbf{X}) \\ 1 & -E(\mathbf{X}) \\ 0 & 1 \end{array} \right] \xrightarrow{[(-\frac{E(\mathbf{X})}{\text{Var}(\mathbf{X})})\mathbf{2}+\mathbf{1}]} \left[\begin{array}{cc} 1 & 0 \\ 0 & \text{Var}(\mathbf{X}) \\ \frac{E(\mathbf{X})^2+\text{Var}(\mathbf{X})}{\text{Var}(\mathbf{X})} & -E(\mathbf{X}) \\ -\frac{E(\mathbf{X})}{\text{Var}(\mathbf{X})} & 1 \end{array} \right] \xrightarrow{[(\frac{1}{\text{Var}(\mathbf{X})})\mathbf{2}]} \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ \frac{E(\mathbf{X})^2+\text{Var}(\mathbf{X})}{\text{Var}(\mathbf{X})} & -\frac{E(\mathbf{X})}{\text{Var}(\mathbf{X})} \\ -\frac{E(\mathbf{X})}{\text{Var}(\mathbf{X})} & \frac{1}{\text{Var}(\mathbf{X})} \end{array} \right].$$

Por tanto $\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{\text{Cov}(\mathbf{X}, \mathbf{X})} \left[\begin{array}{cc} E(\mathbf{X}^2) & -E(\mathbf{X}) \\ -E(\mathbf{X}) & 1 \end{array} \right]$.

□

(L-5) **Problema 9(a)** Invirtiendo la matriz por el método de Gauss-Jordan

$$\left[\begin{array}{ccc} 1 & E(\mathbf{X}) & E(\mathbf{Z}) \\ E(\mathbf{X}) & E(\mathbf{X}^2) & E(\mathbf{X}\mathbf{Z}) \\ E(\mathbf{Z}) & E(\mathbf{X}\mathbf{Z}) & E(\mathbf{Z}^2) \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right] \xrightarrow{\left[\begin{array}{c} [(-E(\mathbf{X}))\mathbf{1}+2] \\ [(-E(\mathbf{Z}))\mathbf{1}+3] \\ [(-\frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})})\mathbf{2}+3] \\ [(-\frac{E(\mathbf{X})}{\text{Var}(\mathbf{X})})\mathbf{2}+1] \\ [(-\frac{\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})})\mathbf{3}+1] \\ [(-\frac{\text{Cov}(\mathbf{X}, \mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})})\mathbf{3}+2] \\ [(-\frac{1}{\text{Var}(\mathbf{X})})\mathbf{2}] \\ [(-\frac{\text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})})\mathbf{3}] \end{array} \right]} \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 + 2 \text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) E(\mathbf{Z}) - E(\mathbf{X})^2 \text{Var}(\mathbf{Z}) - E(\mathbf{Z})^2 \text{Var}(\mathbf{X}) - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{Z}) + E(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \end{array} \right].$$

Por tanto la matriz de varianzas y covarianzas $\text{Var}(\hat{\boldsymbol{\beta}})$ es:

$$\sigma^2 \left[\begin{array}{ccc} \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 + 2 \text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) E(\mathbf{Z}) - E(\mathbf{X})^2 \text{Var}(\mathbf{Z}) - E(\mathbf{Z})^2 \text{Var}(\mathbf{X}) - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{Z}) + E(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{-\text{Cov}(\mathbf{X}, \mathbf{Z}) E(\mathbf{X}) + E(\mathbf{Z}) \text{Var}(\mathbf{X})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} & \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \end{array} \right].$$

□

(L-5) **Problema 9(b)** Basta fijarse en la segunda componente de la diagonal de matriz de varianzas y covarianzas del apartado anterior y operar

$$\begin{aligned} \text{Var}(\hat{\beta}_2) &= \frac{-\sigma^2 \text{Var}(\mathbf{Z})}{\text{Cov}(\mathbf{X}, \mathbf{Z})^2 - \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ &= \frac{\sigma^2 \text{Var}(\mathbf{Z})}{\left(1 - \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})^2}{\text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})}\right) \text{Var}(\mathbf{X}) \text{Var}(\mathbf{Z})} \\ &= \frac{\sigma^2}{\left(1 - \text{Corr}(\mathbf{X}, \mathbf{Z})^2\right) \text{Var}(\mathbf{X})} \\ &= \frac{\sigma^2}{\left(1 - R_{\mathbf{X}\mathbf{Z}}^2\right) \text{Var}(\mathbf{X})}. \end{aligned}$$

□

(L-6) **Problema 1.** La *cuasi-varianza* de \mathbf{Y} es

$$s_{\mathbf{Y}}^2 = \frac{\sum_{n=1}^N (\mathbf{Y}_n - m_{\mathbf{Y}})^2}{N-1}$$

Sea $\mu = E(\mathbf{Y})$ y $\sigma^2 = \text{Var}(\mathbf{Y})$. Ya sabemos que $E(m_{\mathbf{Y}}) = \mu$ y que $\text{Var}(m_{\mathbf{Y}}) = \frac{\sigma^2}{N}$. Además, por el Teorema de Pitágoras sabemos que $E(\mathbf{Y}^2) = \sigma^2 + \mu^2$; y por tanto

$$E\left(\sum \mathbf{Y}_n^2\right) = \sum E(\mathbf{Y}_n^2) = N(\sigma^2 + \mu^2).$$

Y por el mismo teorema,

$$E((m_{\mathbf{Y}})^2) = \text{Var}(m_{\mathbf{Y}}) + E(m_{\mathbf{Y}})^2 = \frac{\sigma^2}{N} + \mu^2.$$

Ahora ya podemos calcular la esperanza del numerador de $s_{\mathbf{Y}}^2$, es decir, de $E\left(\sum (\mathbf{Y}_n - m_{\mathbf{Y}})^2\right)$:

$$\begin{aligned} E\left(\sum (\mathbf{Y}_n - m_{\mathbf{Y}})^2\right) &= E\left(\sum (\mathbf{Y}_n^2 - 2\mathbf{Y}_n m_{\mathbf{Y}} + (m_{\mathbf{Y}})^2)\right) \\ &= E\left(\sum \mathbf{Y}_n^2 - 2m_{\mathbf{Y}} \sum \mathbf{Y}_n + \sum (m_{\mathbf{Y}})^2\right) \\ &= E\left(\sum \mathbf{Y}_n^2 - 2m_{\mathbf{Y}} N m_{\mathbf{Y}} + N(m_{\mathbf{Y}})^2\right) \\ &= E\left(\sum \mathbf{Y}_n^2 - N(m_{\mathbf{Y}})^2\right) \\ &= E\left(\sum \mathbf{Y}_n^2\right) - N E((m_{\mathbf{Y}})^2); \end{aligned}$$

sustituyendo por las expresiones de más arriba tenemos que

$$\begin{aligned} &= N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right) \\ &= (N-1)\sigma^2. \end{aligned}$$

Consecuentemente $E(s_{\mathbf{Y}}^2) = E\left(\frac{\sum_{n=1}^N (\mathbf{Y}_n - m_{\mathbf{Y}})^2}{N-1}\right) = \frac{1}{N-1} E\left(\sum_{n=1}^N (\mathbf{Y}_n - m_{\mathbf{Y}})^2\right) = \frac{N-1}{N-1} \sigma^2 = \sigma^2$.

□

(L-6) Problema 2. Veamos que es cierto para la componente de la fila i -ésima y columna j -ésima:

$$\begin{aligned} {}_{ij} \text{Var}(\mathbf{Q} \mathbf{Y})_{|j} &= \text{Cov}({}_{ij} \mathbf{Q} \mathbf{Y}, {}_{ij} \mathbf{Q} \mathbf{Y}) \\ &= \text{Cov}({}_{ij} \mathbf{Q} \mathbf{Y}, \mathbf{Y}(\mathbf{Q}^T)_{|j}) \quad \text{ya que } \mathbf{Y} \mathbf{Q}^T = \mathbf{Q} \mathbf{Y} \\ &= E([{}_{ij} \mathbf{Q} \mathbf{Y} - \mathbf{I} E({}_{ij} \mathbf{Q} \mathbf{Y})] [{}_{ij} \mathbf{Q} \mathbf{Y}^T]_{|j} - \mathbf{I} E({}_{ij} \mathbf{Q} \mathbf{Y}^T)_{|j})^T \\ &= E([{}_{ij} \mathbf{Q} \mathbf{Y} - {}_{ij} \mathbf{Q} \mathbf{I} E(\mathbf{Y})] [{}_{ij} \mathbf{Q} \mathbf{Y}^T]_{|j} - \mathbf{I} E(\mathbf{Y}) (\mathbf{Q}^T)_{|j})^T \\ &= E({}_{ij} \mathbf{Q} [\mathbf{Y} - \mathbf{I} E(\mathbf{Y})] [\mathbf{Y} - \mathbf{I} E(\mathbf{Y})]^T (\mathbf{Q}^T)_{|j}) \\ &= {}_{ij} \mathbf{Q} E([\mathbf{Y} - \mathbf{I} E(\mathbf{Y})] [\mathbf{Y} - \mathbf{I} E(\mathbf{Y})]^T) (\mathbf{Q}^T)_{|j} \\ &= {}_{ij} \mathbf{Q} \text{Var}(\mathbf{Y}) (\mathbf{Q}^T)_{|j}. \end{aligned}$$

□

(L-6) Problema 3. De las ecuaciones normales (Página 22) sabemos que $N^{-1} \mathbf{X}^T \mathbf{X} = \begin{bmatrix} (\mathbf{1} \cdot \mathbf{1}) & (\mathbf{1} \cdot \mathbf{x}) \\ (\mathbf{x} \cdot \mathbf{1}) & (\mathbf{x} \cdot \mathbf{x}) \end{bmatrix}$. Así que

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & m_{\mathbf{x}} \\ m_{\mathbf{x}} & m_{\mathbf{x}^2} \end{bmatrix}.$$

Calculemos la inversa de esta matriz

$$\begin{bmatrix} N^{-1} \mathbf{X}^T \mathbf{X} \\ \mathbf{I} \end{bmatrix} = \begin{bmatrix} 1 & m_{\mathbf{x}} \\ m_{\mathbf{x}} & m_{\mathbf{x}^2} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{[(m_{\mathbf{x}})^{-1} \mathbf{1} + \mathbf{2}]} \begin{bmatrix} 1 & 0 \\ m_{\mathbf{x}} & s_{\mathbf{x}}^2 \\ 1 & -m_{\mathbf{x}} \\ 0 & 1 \end{bmatrix} \xrightarrow{[(1/s_{\mathbf{x}}^2) \mathbf{2}]} \begin{bmatrix} 1 & 0 \\ m_{\mathbf{x}} & 1 \\ 1 & -m_{\mathbf{x}}/s_{\mathbf{x}}^2 \\ 0 & 1/s_{\mathbf{x}}^2 \end{bmatrix} \xrightarrow{[(-m_{\mathbf{x}})^{-1} \mathbf{2} + \mathbf{1}]} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 + m_{\mathbf{x}}^2/s_{\mathbf{x}}^2 & -m_{\mathbf{x}}/s_{\mathbf{x}}^2 \\ -m_{\mathbf{x}}/s_{\mathbf{x}}^2 & 1/s_{\mathbf{x}}^2 \end{bmatrix}.$$

Por tanto,

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{Ns_x^2} \begin{bmatrix} s_x^2 + m_x^2 & -m_x \\ -m_x & 1 \end{bmatrix} = \frac{1}{Ns_x^2} \begin{bmatrix} m_{x^2} & -m_x \\ -m_x & 1 \end{bmatrix},$$

pues $m_{x^2} = s_x^2 + m_x^2$ (T. Pitágoras).

Así pues, sustituyendo μ_x por $m_{\mathbf{x}}$ y σ_x^2 por $s_{\mathbf{x}}^2$, tenemos que la matriz de varianzas y covarianzas del estimador $\hat{\beta}$ con un *m.a.s.* de tamaño N es:

$$\text{Var}(\hat{\beta}) = E(\mathbf{X}^\top \mathbf{X})^{-1} = E\left(\frac{\sigma^2}{Ns_{\mathbf{x}}^2} \begin{bmatrix} s_{\mathbf{x}}^2 + m_{\mathbf{x}}^2 & -m_{\mathbf{x}} \\ -m_{\mathbf{x}} & 1 \end{bmatrix}\right); \quad \text{donde } \sigma^2 = \text{Var}(U).$$

Así pues, podemos deducir que

$$\widehat{\text{Var}}(\hat{a}) = \sigma^2 E\left(\frac{s_{\mathbf{x}}^2 + m_{\mathbf{x}}^2}{Ns_{\mathbf{x}}^2}\right) = \sigma^2 E\left(\frac{m_{\mathbf{x}^2}}{Ns_{\mathbf{x}}^2}\right) \quad \text{y} \quad \widehat{\text{Var}}(\hat{b}) = E\left(\frac{\sigma^2}{Ns_{\mathbf{x}}^2}\right). \quad (68)$$

(recuérdese que $m_{\mathbf{x}^2} = s_{\mathbf{x}}^2 + m_{\mathbf{x}}^2$). Además, ambos estimadores tienen covarianza

$$\widehat{\text{Cov}}(\hat{a}, \hat{b}) = E\left(\frac{-\sigma^2 m_{\mathbf{x}}}{Ns_{\mathbf{x}}^2}\right). \quad (69)$$

□

(L-6) Problema 4. En este caso seleccionamos la componente j -ésima del vector $\hat{\beta}$, por tanto

$$\text{Var}(\mathbf{v} \cdot \tilde{\beta} | \mathbf{X}) = \text{Var}(\tilde{\beta}_j | \mathbf{X}) \geq \text{Var}(\hat{\beta}_j | \mathbf{X}) = \text{Var}(\mathbf{v} \cdot \hat{\beta} | \mathbf{X}).$$

Es decir, el teorema de Gauss-Markov implica que la varianza del estimador de cada parámetro j -ésimo $\text{Var}(\tilde{\beta}_j | \mathbf{X})$ es mayor o igual que la del estimador MCO: $\text{Var}(\hat{\beta}_j | \mathbf{X})$.

□

(L-6) Problema 5. Por una parte: $\mathbf{P}\mathbf{X} = \mathbf{X} \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{\mathbf{I}} \mathbf{X} = \mathbf{X}$

Además \mathbf{P} es simétrica, ya que

$$\begin{aligned} \mathbf{P}^\top &= [\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]^\top \\ &= [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]^\top \mathbf{X}^\top && \text{pues } [\mathbf{XA}]^\top = \mathbf{A}^\top \mathbf{X}^\top \\ &= \mathbf{X} [(\mathbf{X}^\top \mathbf{X})^{-1}]^\top \mathbf{X}^\top && \text{idéntica regla de trasposición sobre el corchete} \\ &= \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{P} && \text{pues } (\mathbf{X}^\top \mathbf{X})^{-1} \text{ es simétrica.} \end{aligned}$$

Y también es idempotente:

$$\begin{aligned} \mathbf{P}\mathbf{P} &= \mathbf{P}\mathbf{XA} \\ &= \mathbf{XA} = \mathbf{P} && \text{pues } \mathbf{P}\mathbf{X} = \mathbf{X} \end{aligned}$$

□

(L-6) Problema 6. Por una parte

$$\mathbf{MX} = [\mathbf{I} - \mathbf{P}] \mathbf{X} = \mathbf{X} - \mathbf{P}\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}.$$

Y por otra

$$\mathbf{AM} = \mathbf{A} [\mathbf{I} - \mathbf{P}] = \mathbf{A} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{A} - \mathbf{A} = \mathbf{0}.$$

Además \mathbf{M} es simétrica puesto que

$$\mathbf{M}^T = [\mathbf{I} - \mathbf{P}]^T = \mathbf{I}^T - \mathbf{P}^T = \mathbf{I} - \mathbf{P} = \mathbf{M}$$

Y también es idempotente ya que

$$\mathbf{MM} = [\mathbf{I} - \mathbf{P}] [\mathbf{I} - \mathbf{P}] = \mathbf{I} - \mathbf{P} - \mathbf{P} + \mathbf{PP} = \mathbf{I} - \mathbf{P} - \mathbf{P} + \mathbf{P} = \mathbf{I} - \mathbf{P} = \mathbf{M}$$

□

(L-6) Problema 7. Basta recordar que $\widehat{\mathbf{e}} = \mathbf{MU}$ y emplear las propiedades de la matriz \mathbf{M} :

$$\widehat{SRC} = \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} = \mathbf{MU} \cdot \mathbf{MU} = \mathbf{UM}^T \mathbf{MU} = \mathbf{UMU}.$$

por ser \mathbf{M} simétrica e idempotente.

□

(L-6) Problema 8. Recordando que la traza es una operación lineal tenemos

$$\begin{aligned} \text{tr}(\mathbf{M}) &= \text{tr}\left(\underset{N \times N}{\mathbf{I}} - \underset{N \times N}{\mathbf{P}}\right) && \text{puesto que } \mathbf{M} \equiv \mathbf{I} - \mathbf{P} \\ &= \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{P}) && \text{puesto que traza es lineal} \\ &= N - \text{tr}(\mathbf{P}) \end{aligned}$$

Y

$$\begin{aligned} \text{tr}(\mathbf{P}) &= \text{tr}(\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) && \text{puesto que } \mathbf{P} \equiv \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{XA} \\ &= \text{tr}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}) && \text{puesto que } \text{tr}(\mathbf{XA}) = \text{tr}(\mathbf{AX}) \\ &= \text{tr}\left(\underset{k \times k}{\mathbf{I}}\right) = k \end{aligned}$$

□

(L-6) Problema 9. Sabemos que $\widehat{\mathbf{e}} = \mathbf{MU}$; por tanto

$$\begin{aligned} \mathbb{E}(\widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} | \mathbf{X}) &= \mathbb{E}(\mathbf{UMU} | \mathbf{X}) && \text{por ser } \mathbf{M} \text{ idempotente} \\ &= \mathbb{E}\left(\sum_{j=1}^N \left(\sum_{i=1}^N \mathbf{U}_i \cdot \mathbf{M}_{ij} \mathbf{U}_i^T\right) \cdot \mathbf{U}_j \mid \mathbf{X}\right) && \text{pues } \mathbf{UMU} = (\mathbf{U}\mathbf{M}) \cdot \mathbf{U} \\ &= \sum_{j=1}^N \sum_{i=1}^N \mathbb{E}\left(\mathbf{U}_i \cdot \mathbf{M}_{ij} \mathbf{U}_i^T \mid \mathbf{X}\right) && \text{pues la esperanza es un operador lineal} \\ &= \sum_{j=1}^N \sum_{i=1}^N (\mathbf{M}_{ij}) \mathbb{E}(\mathbf{U}_i \mathbf{U}_i^T \mid \mathbf{X}) && \text{pues } \mathbf{M} \in \mathbf{X} \\ &= \sigma^2 \sum_{i=1}^N (\mathbf{M}_{ii}) && \text{por el supuesto: } \text{Var}(\mathbf{U} \mid \mathbf{X}) = \sigma^2 \mathbf{I} \\ &= \sigma^2 \text{tr}(\mathbf{M}) = \sigma^2(N - k) \mathbf{1} && \text{por la Nota 5 (Pág. 82) y Proposición 14.1} \end{aligned}$$

□

(L-7) **Problema 1.** Demostración:

$$\begin{aligned} \frac{N-k}{\sigma^2} \hat{s}_{\hat{\mathbf{e}}}^2 &= \frac{N-k}{\sigma^2} (\hat{\mathbf{e}} \cdot \hat{\mathbf{e}})/(N-k) = \frac{1}{\sigma^2} \hat{\mathbf{e}} \cdot \hat{\mathbf{e}} = \frac{1}{\sigma} \hat{\mathbf{e}} \cdot \hat{\mathbf{e}} \frac{1}{\sigma} \\ &= \frac{1}{\sigma} \mathbf{U} \mathbf{M}^\top \mathbf{M} \mathbf{U} \frac{1}{\sigma} && \text{ya que } \hat{\mathbf{e}} = \mathbf{M} \mathbf{U} \\ &= \frac{1}{\sigma} \mathbf{U} \mathbf{M} \mathbf{U} \frac{1}{\sigma} \sim \chi_{(N-k)}^2 \end{aligned}$$

puesto que \mathbf{M} es idempotente, $\mathbf{U} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, por las **proposiciones 18.1** y **18.2** en la página 90 y la **Proposición 14.1** en la página 82.

□

(L-7) **Problema 2.** $E(\hat{s}_{\hat{\mathbf{e}}}^2) = \frac{\sigma^2}{N-k} \cdot (N-k) = \sigma^2$ y $\text{Var}(\hat{s}_{\hat{\mathbf{e}}}^2) = \left(\frac{\sigma^2}{N-k}\right)^2 \cdot 2(N-K) = 2\frac{\sigma^4}{(N-k)}$.

□

(L-7) **Problema 3.** Puesto que tanto $\hat{\beta}$ como $\hat{\mathbf{e}}$ tienen distribución normal, basta demostrar que ambas variables tienen covarianza nula. Como $(\hat{\beta} - \mathbf{I}\beta) = \mathbf{A}\mathbf{U}$ y $\hat{\mathbf{e}} = \mathbf{M}\mathbf{U}$, entonces

$$\begin{aligned} \text{Cov}(\mathbf{A}\mathbf{U}, \mathbf{M}\mathbf{U} \mid \mathbf{X}) &= \mathbf{A} \text{Var}(\mathbf{U} \mid \mathbf{X}) \mathbf{M}^\top && \text{por la Nota 8 (Página 86)} \\ &= \mathbf{A} \sigma^2 \mathbf{I} \mathbf{M}^\top && \text{por el supuesto: } \text{Var}(\mathbf{U} \mid \mathbf{X}) = \sigma^2 \mathbf{I} \\ &= \sigma^2 \mathbf{A} \mathbf{M} = \sigma^2 \mathbf{0} = \mathbf{0}, \end{aligned}$$

ya que $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, por lo que $\mathbf{A}\mathbf{X} = \mathbf{I}$ y $\mathbf{M} = \mathbf{I} - \mathbf{X}\mathbf{A}$. Así $\mathbf{A}\mathbf{M} = \mathbf{A}\mathbf{I} - \mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{I} - \mathbf{I}\mathbf{A} = \mathbf{0}$.

□

(L-7) **Problema 4.** Demostración.

$$\frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\sqrt{\hat{s}_{\hat{\mathbf{e}}}^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\sqrt{\sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}} \cdot \sqrt{\frac{\sigma^2}{\hat{s}_{\hat{\mathbf{e}}}^2}} = \frac{Z}{\sqrt{\frac{\hat{s}_{\hat{\mathbf{e}}}^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{1}{\sigma^2} \frac{\hat{\mathbf{e}} \cdot \hat{\mathbf{e}}}{N-k}}}$$

donde el numerador

$$Z = \frac{\hat{\beta}_j - \beta_j \mathbf{1}}{\sqrt{\sigma^2 [(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}},$$

es función de $(\hat{\beta} - \mathbf{I}\beta)$ y el denominador es función de $\hat{\mathbf{e}}$. Así pues, por la **Nota 9** en la página 90 y la **Proposición 18.4** en la página 90 el numerador y el denominador son independientes.

Además, el numerador tiene distribución $N(0, 1)$. Por tanto tenemos una $N(0, 1)$ dividida por la raíz cuadrada de un χ^2 dividida por sus grados de libertad; este cociente tiene distribución t de Student con $N - k$ grados de libertad.

□

(L-7) **Problema 5.** **Falso:** El test de hipótesis sólo puede indicar que con los datos observados no rechazamos la hipótesis nula H_0 con una probabilidad φ , es decir, asumiendo el riesgo de que, con probabilidad φ , no rechacemos H_0 aún siendo falsa (error tipo II).

□

(L-7) **Problema 6.** **Verdadero:** la región crítica asociada al nivel de significación del 10% contiene a la del 5%; por lo tanto si se ha “caído” en la región del 5%, entonces también se ha “caído” en la del 10%.

□

(L-7) **Problema 7.** $H_0 : \beta = 0$, frente a $H_1 : \beta > 0$. Nótese que, por t^a económica, lo más razonable es proponer un contraste de cola superior. La región crítica más conveniente es por tanto: $RC = \{\hat{\beta} > k\} = \left\{ \frac{\hat{\beta} - \beta}{\widehat{Dt}(\hat{\beta} | \mathbf{x})} > t_{33}^{(1-\alpha)} \right\}$,

donde $\widehat{\beta}$ es la estimación MCO y $t_{33}^{(1-\alpha)}$ es el valor tabulado de una t -Student con 33 grados de libertad que deja a la izquierda un área igual a $1 - \alpha$.

□

(L-8) Problema 1. La demostración es similar a la de la Proposición 18.5 (pág. 90)⁸⁶. Primero escribimos el estadístico en la forma

$$\mathcal{F} = \frac{\mathbf{W}/r}{\frac{1}{\sigma^2} \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}}/(N-k)}$$

donde $\mathbf{W} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) \left[\sigma^2 \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r})$.

Los pasos a demostrar son:

1. $\mathbf{W} \underset{H_0}{\sim} \chi_r^2$

(Bajo H_0 se verifica que $\text{Var}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r} \mid \mathbf{X}) = \sigma^2 \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top$ y puesto que $(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) \underset{H_0}{\sim} N$ aplicando la Proposición 18.2 (Pag. 90) queda demostrado.)

2. $(\widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} / \sigma^2) \underset{H_0}{\sim} \chi_{N-k}^2$.

(es la Proposición 18.3).

3. \mathbf{W} es independiente de $(\widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} / \sigma^2)$

(la demostración de este punto es idéntica al razonamiento empleado en la Proposición 18.5, pág. 90).

La demostración se completa con la Nota 10 en la página 96 de más arriba.

□

(L-8) Problema 2(a) La restricción escrita en forma matricial es: $\mathbf{R}\boldsymbol{\beta} = [0, 3, -1, 0] (\beta_1, \beta_2, \beta_3, \beta_4) = 7$; además $r = 1$ y $k=4$. Por una parte $s_{\widehat{\mathbf{e}}}^2 = \frac{SRC}{N-k} = 40/21$; por otra parte

$$\begin{aligned} \mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top &= (0, 3, -1, 0) \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & v_{44} \end{bmatrix} \begin{pmatrix} 0 \\ 3 \\ -1 \\ 0 \end{pmatrix} \\ &= 9v_{22} - 6v_{23} + v_{33} = 9 \cdot 4 - 6 \cdot 2 + 6 = 30 \end{aligned}$$

Así pues, puesto que $\frac{\mathbf{R}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}}|\mathbf{X})} \underset{H_0}{\sim} t_{N-k}$ (de la Nota 12 en la página 98)

$$\widehat{\mathcal{T}} = \frac{\mathbf{R}\widehat{\boldsymbol{\beta}} - b}{\widehat{\text{Dt}}(\mathbf{R}\widehat{\boldsymbol{\beta}})} = \frac{3\widehat{\beta}_2 - \widehat{\beta}_3 - 7}{\sqrt{\frac{40}{21} \cdot 30}} = 0.198 < t_{N-k}^{(0.975)} = 2.08$$

pues $t_{21}^{(0.975)} = 2.08$ (dos colas $\alpha = 0.05$). **El nivel crítico p (p -value):** t_{21} : área a la derecha de 0.198 es 0.422475 (valor a dos colas: 0.844949; complemento = 0.155051).

□

(L-8) Problema 2(b) Ahora, $r = 2$ y $\mathbf{R} = \begin{bmatrix} 0 & 3 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mathbf{r} = \begin{pmatrix} 7 \\ 0 \end{pmatrix}$

El estadístico a utilizar es el ratio-F: $\mathcal{F} = \frac{(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{I}\mathbf{r}) / r}{s_{\widehat{\mathbf{e}}}^2}$;

⁸⁶pude encontrar la demostración completa en Hayashi (2000, pp. 41)

Los resultados parciales son

$$\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top = \begin{bmatrix} 30 & -50 \\ -50 & 10 \end{bmatrix}$$

$$[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} = \begin{bmatrix} -0.0045455 & -0.0227273 \\ -0.0227273 & -0.0136364 \end{bmatrix} \quad \text{y} \quad \mathbf{R}\hat{\beta} - \mathbf{r} = \begin{pmatrix} 1.5 \\ -1.5 \end{pmatrix}$$

Por tanto $\hat{\mathcal{F}} = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r}) [\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) / r}{\hat{s}^2}$

$$= (1.5, -1.5) \begin{bmatrix} -0.0045455 & -0.0227273 \\ -0.0227273 & -0.0136364 \end{bmatrix} \begin{pmatrix} 1.5 \\ -1.5 \end{pmatrix} \cdot \frac{1}{2} \cdot \frac{21}{40} = 0.016108$$

Si realizamos el contraste al 5% de significación tenemos: $F_{2,21}^{(0.95)} = 3.47$ ($\alpha = 0.05$).

□

(L-9) Problema 1. STC es común, puesto que la variable dependiente \mathbf{y} es la misma sea el modelo restringido o no. Por tanto

$$\frac{SRC^* - SRC}{SRC} = \frac{(1 - R^2^*)STC - (1 - R^2)STC}{(1 - R^2)STC} = \frac{R^2 - R^2^*}{1 - R^2}$$

□

(L-9) Problema 2(a) Para el modelo lineal simple tenemos

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{1} \cdot \mathbf{1} & \mathbf{1} \cdot \mathbf{x} \\ \mathbf{x} \cdot \mathbf{1} & \mathbf{x} \cdot \mathbf{x} \end{bmatrix} = \begin{bmatrix} N & Nm_x \\ Nm_x & Nm_{x^2} \end{bmatrix}$$

y aplicando la eliminación Gaussiana por columnas podemos calcular la inversa:

$$\begin{bmatrix} N & Nm_x \\ Nm_x & Nm_{x^2} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{[(-m_x)\mathbf{1}+2]} \begin{bmatrix} N & 0 \\ Nm_x & Ns_x^2 \\ 1 & -m_x \\ 0 & 1 \end{bmatrix} \xrightarrow{[(\frac{1}{N}s_x^2)\mathbf{2}]} \begin{bmatrix} 1 & 0 \\ m_x & 1 \\ \frac{1}{N} & \frac{-m_x}{Ns_x^2} \\ 0 & \frac{1}{Ns_x^2} \end{bmatrix} \xrightarrow{[(-m_x)\mathbf{2}+\mathbf{1}]} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{s_x^2+m_x^2}{Ns_x^2} & \frac{-m_x}{Ns_x^2} \\ \frac{-m_x}{Ns_x^2} & \frac{1}{Ns_x^2} \end{bmatrix},$$

así

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{Ns_x^2} \begin{bmatrix} (s_x^2 + m_x^2) & -m_x \\ -m_x & 1 \end{bmatrix}.$$

Si la restricción es $\beta_1 = 0$, es decir $\mathbf{R}\beta = [1 \ 0] \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \mathbf{0} = \mathbf{r}$, entonces $\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top = \left[\frac{s_x^2 + m_x^2}{Ns_x^2} \right]$ y la estimación MCRL queda así

$$\begin{aligned} \widehat{\beta}^* &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} - \frac{1}{Ns_x^2} \begin{bmatrix} (s_x^2 + m_x^2) & -m_x \\ -m_x & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \left[\frac{s_x^2 + m_x^2}{Ns_x^2} \right]^{-1} (\widehat{\beta}_1 - 0,) \\ &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} - \frac{1}{Ns_x^2} \begin{bmatrix} s_x^2 + m_x^2 \\ -m_x \end{bmatrix} \begin{bmatrix} \frac{Ns_x^2}{s_x^2 + m_x^2} \end{bmatrix} (\widehat{\beta}_1,) = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} - \begin{bmatrix} 1 \\ -\frac{m_x}{s_x^2 + m_x^2} \end{bmatrix} (\widehat{\beta}_1,) \\ &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} - \left(-\frac{\widehat{\beta}_1}{s_x^2 + m_x^2} \widehat{\beta}_1 \right) = \begin{pmatrix} 0 \\ \widehat{\beta}_2 + \frac{m_x}{s_x^2 + m_x^2} \widehat{\beta}_1 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \widehat{\beta}_2 + \frac{m_x}{m_{x^2}} \widehat{\beta}_1 \end{pmatrix}; \end{aligned}$$

pues $m_{x^2} = s_x^2 + m_x^2$ (T. Pitágoras).

Nótese que si en la estimación MCO resultara que $\widehat{\beta}_1 = 0$ (que se cumpliera la restricción), entonces $\widehat{\beta}^* = \widehat{\beta}$

□

(L-9) Problema 2(b) Por otra parte, si la restricción es $\beta_2 = 0$, es decir $\mathbf{R}\beta = [0 \ 1] \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \mathbf{0} = \mathbf{r}$; entonces $\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top = \frac{1}{Ns_x^2}$ y la estimación MCRL queda así

$$\begin{aligned}\widehat{\beta}^* &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} - \frac{1}{Ns_x^2} \begin{bmatrix} (s_x^2 + m_x^2) & -m_x \\ -m_x & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \left[\frac{1}{Ns_x^2} \right]^{-1} (\widehat{\beta}_2 - 0,) \\ &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} - \frac{1}{Ns_x^2} \begin{bmatrix} -m_x \\ 1 \end{bmatrix} [Ns_x^2] (\widehat{\beta}_2,) = \begin{pmatrix} \widehat{\beta}_1 + m_x \widehat{\beta}_2 \\ 0 \end{pmatrix}\end{aligned}$$

De nuevo, si en la estimación MCO resultara que $\widehat{\beta}_2 = 0$ (que se cumpliera la restricción), entonces $\widehat{\beta}^* = \widehat{\beta}$

□

(L-9) Problema 3. Vamos a emplear el estadístico \mathcal{F} expresado mediante sumas residuales (47) para contrastar la hipótesis.

Para ello estimamos primero por MCO:

$$\text{Por una parte tenemos } (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 10 & 4 & -2 \\ 4 & 8 & 0 \\ -2 & 0 & 6 \end{bmatrix}^{-1}; \text{ y por otra } \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 3 \\ -3 \\ -2 \end{pmatrix}.$$

Así pues,

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} \frac{23}{44} \\ \frac{-28}{44} \\ \frac{-7}{44} \end{pmatrix};$$

y puesto que $SRC = \widehat{\mathbf{e}} \cdot \widehat{\mathbf{e}} = \mathbf{y} \cdot \mathbf{y} - \widehat{\beta} \mathbf{X}^\top \mathbf{y}$ (de T^a de Pitágoras y $\widehat{\mathbf{y}} \cdot \widehat{\mathbf{y}} = \widehat{\mathbf{y}} \cdot \mathbf{y}$), entonces

$$SRC = 10 - \left(\frac{23}{44}, \frac{-28}{44}, \frac{-7}{44} \right) \begin{pmatrix} 3 \\ -3 \\ -2 \end{pmatrix} = 10 - \left(\frac{23}{44} \cdot 3 + \frac{28}{44} \cdot 3 + \frac{7}{44} \cdot 2 \right) = 6.20$$

Ahora estimamos por MCR:

$$\widehat{\beta}^* = \widehat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{r} - \mathbf{R}\widehat{\beta}) = \begin{pmatrix} \frac{4}{7} \\ \frac{-4}{7} \\ \frac{-1}{7} \end{pmatrix}$$

donde $\mathbf{R} = (1, 1, 0)$ y $\mathbf{r} = 0$.

De nuevo $SRC^* = \mathbf{y} \cdot \mathbf{y} - \widehat{\beta}^* \mathbf{X}^\top \mathbf{y}$, y por tanto

$$SRC^* = 10 - \left(\frac{4}{7} \cdot 3 + \frac{4}{7} \cdot 3 + \frac{1}{7} \cdot 2 \right) = 6.28$$

Empleando (47) obtenemos:

$$\begin{aligned}\widehat{\mathcal{F}} &= \frac{N - k}{r} \frac{SRC^* - SRC}{SRC} \\ &= \frac{100}{1} \frac{6.28 - 6.20}{6.20} = 1.29 < F_{1,100}^{(0.95)} = 3.94\end{aligned}$$

Hay una segunda forma de obtener SRC^* estimando por MCO el modelo con las restricciones incorporadas: Si sustituimos la restricción ($\beta_1 + \beta_2 = 0$) en el modelo tenemos:

$$\mathbf{Y} = \beta_1(\mathbf{X}_1 - \mathbf{X}_2) + \beta_3\mathbf{X}_3 + \mathbf{U} = \beta_1\mathbf{X}_1^* + \beta_3\mathbf{X}_3 + \mathbf{U},$$

donde $\mathbf{X}_1^* = \mathbf{X}_1 - \mathbf{X}_2$.

Ahora

$$\mathbf{X}^{*\top}\mathbf{X}^* = \begin{bmatrix} 10 & -2 \\ -2 & 6 \end{bmatrix} \quad \text{y} \quad \mathbf{X}^{*\top}\mathbf{y} = \begin{pmatrix} 6 \\ -2 \end{pmatrix}$$

puesto que

$$\begin{aligned} \sum(x_{n1}^*)^2 &= \sum(x_{n1})^2 + \sum(x_{n2})^2 - 2\sum(x_{n1}x_{n2}) &= 10 + 8 - 2 \cdot 4 = 10 \\ \sum(x_{n1}^*x_{n3}) &= \sum(x_{n1}x_{n3}) - \sum(x_{n2}x_{n3}) &= -2 - 0 = -2 \\ \sum(x_{n1}^*y_n) &= \sum(x_{n1}y_n) - \sum(x_{n2}y_n) &= 3 - (-3) = 6 \end{aligned}$$

Así pues, $\widehat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{y} = \begin{pmatrix} 4/7 \\ -1/7 \end{pmatrix}$.

De nuevo $SRC^* = \mathbf{y} \cdot \mathbf{y} - \widehat{\boldsymbol{\beta}}^*\mathbf{X}^*\mathbf{y}$, y

$$SRC^* = 10 - \left(\frac{4}{7}6 + \frac{1}{7}2\right) = 6.28;$$

que es el mismo resultado que obtuvimos empleando $\widehat{\boldsymbol{\beta}}^*$.

□

(Lección 10) Ejercicio en clase. N- 1(e) 10,2096

□

(Lección 10) Ejercicio en clase. N- 1(f) 287,609

□

(Lección 10) Ejercicio en clase. N- 1(g) No. Los p -valores son muy elevados

□

(Lección 10) Ejercicio en clase. N- 1(h) La varianza parece crecer en función de la renta, algo que no debería pasar bajo el supuesto de homocedasticidad.

□

(Lección 10) Ejercicio en clase. N- 2(e) En el primer caso la elasticidad es creciente, en el caso del modelo Lin-Log la elasticidad decrece con el tamaño de la vivienda (como cabría esperar).

□

(Lección 10) Ejercicio en clase. N- 2(f) Aumento del precio = $\widehat{\beta}_2 * 0.01 \approx 3000$ dólares.

□

(Lección 10) Ejercicio en clase. N- 3(b) El modelo final resulta ser:

$$\widehat{\text{L_WAGE}} = 7,02337 + 0,0236809 \text{EXPER} + 0,00502251 \text{sq_EDUC}$$

(0,092457)	(0,0061404)	(0,0011710)
------------	-------------	-------------

$$T = 49 \quad \bar{R}^2 = 0,3334 \quad F(2, 46) = 13,004 \quad \hat{\sigma} = 0,25534$$

(Desviaciones típicas entre paréntesis)

□

(Lección 10) Ejercicio en clase. N- 3(c) Aproximadamente un incremento salarial de 2.368%. Si lo calculamos con más precisión ($\exp(\beta_2) - 1$), el incremento salarial esperado es de 2.40%

□

(Lección 10) Ejercicio en clase. N- 3(e) Puesto que el modelo final es aproximadamente

$$\ln \widehat{WAGE} = 7,023 + 0,024\text{EXPER} + 0,005\text{EDUC}^2$$

derivando con respecto a $EDUC$ tenemos que $\frac{1}{WAGE} \cdot \frac{\partial WAGE}{\partial EDUC} = 2 \cdot 0,005 \cdot EDUC$ es decir

$$\frac{\Delta WAGE}{WAGE} = 0,01 \cdot EDUC \cdot \Delta EDUC$$

Si el incremento en la educación es un año ($\Delta EDUC = 1$), la respuesta es aproximadamente un 1% en el primer caso y un 7% en el segundo.

□

(Lección 10) Ejercicio en clase. N- 4(c) El segundo (Log-Lin) parece tener una ligera ventaja ($rsq = 0,369558$)

□

(Lección 10) Ejercicio en clase. N- 4(d) El segundo de nuevo parece mostrar una ligerísima ventaja según los estadísticos de selección de modelos.

□

(Lección 10) Ejercicio en clase. N- 5(b) El modelo resultante tras eliminar secuencialmente las variables es:

$$\begin{aligned} \widehat{1_BUSTRAVL} &= 45,8457 - 4,730081\text{INCOME} + 1,820371\text{POP} - 0,9709971\text{LANDAREA} \\ &\quad (9,6141) \quad (1,0212) \quad (0,23573) \quad (0,20681) \\ T &= 40 \quad \bar{R}^2 = 0,6087 \quad F(3,36) = 21,220 \quad \hat{\sigma} = 0,72412 \\ &\quad (\text{Desviaciones típicas entre paréntesis}) \end{aligned}$$

□

(Lección 10) Ejercicio en clase. N- 5(c) Los estadísticos t y sus niveles de significación marginales (p -valores) son:

$t_{Inc} = 3,65267$; $p_{Inc} = 0,00081935$

$t_{Pop} = 3,48009$; $p_{Pop} = 0,00133085$

$t_{Land} = -0,140241$; $p_{Land} = 0,889252$

No se puede rechazar que la elasticidad sea unitaria para *landarea*. Para las otras variables, la demanda es claramente elástica.

□

(Lección 11) Ejercicio en clase. N- 2(a) Un hombre gana en media 525.6 dólares más que una mujer. Cada año adicional de experiencia supone casi 20 dólares más de salario.

□

(Lección 11) Ejercicio en clase. N- 2(b) Un hombre gana en media casi un 29% más que una mujer. Cada año adicional de experiencia supone casi un 1.3% más de salario.

□

(Lección 11) Ejercicio en clase. N- 2(c) Un hombre gana en media casi un 26% más que una mujer. Cada año adicional de experiencia supone casi un 2% más de salario. Cada año adicional de educación supone un 6% más de salario.

Nótese como incluir la educación en el modelo aumenta la significatividad de la variable experiencia con los datos de esta muestra.

□

Ejercicio 1. El modelo restringido es

$$Y = \beta_1 \mathbb{1} + \beta_2 X + U^*,$$

y el modelo sin restringir (general)

$$Y = \beta_1^A \mathbb{1}_A + \beta_1^B \mathbb{1}_B + \beta_2^A X \cdot \mathbb{1}_A + \beta_2^B X \cdot \mathbb{1}_B + U,$$

La hipótesis nula contiene dos restricciones:

$$H_0 : \beta_1^A = \beta_1^B \text{ y conjuntamente } \beta_2^A = \beta_2^B$$

por tanto el contraste es necesariamente un contraste F con hipótesis alternativa:

$$H_1 : \text{la nula es falsa.}$$

De nuevo tenemos dos opciones

1. Por sumas residuales: estimando ambos modelos y empleando el estadístico de la ecuación 47 en la página 106 del tema 2.
2. Por sustitución: $\mathbb{1}_B = \mathbb{1} - \mathbb{1}_A$ en el modelo sin restringir:

$$\begin{aligned} Y &= \beta_1^B \mathbb{1} + (\beta_1^A - \beta_1^B) \mathbb{1}_A + \beta_2^B X + (\beta_2^A - \beta_2^B) X \cdot \mathbb{1}_A + U \\ &= \beta_1^B \mathbb{1} + \alpha \mathbb{1}_A + \beta_2^B X + \delta X \cdot \mathbb{1}_A + U \end{aligned}$$

en cuyo caso la hipótesis nula se transforma en

$$H_0 : \begin{pmatrix} \alpha \\ \delta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

es decir, un contraste de significación conjunta de α y β (compárese con el contraste de Chow de cambio estructural de la página 107)

□

(Lección 11) Ejercicio en clase. N- 3(b) La salida de Gretl es:

Contraste de Chow de diferencia estructural con respecto a D90
 $F(9, 82) = 6,90351$ con valor p 0,0000

□

(Lección 11) Ejercicio en clase. N- 3(d) El modelo resultante es:

$$\begin{aligned} \widehat{WLFP} &= 47,6366 + 0,00477939 YF - 0,00405375 D90YF + 0,275070 EDUC \\ &\quad (6,5784) \quad (0,00073395) \quad (0,00068212) \quad (0,045506) \\ &\quad - 1,06141 UE - 0,569355 D90UE - 0,207293 MR + 0,126361 D90MR \\ &\quad (0,24559) \quad (0,32722) \quad (0,10489) \quad (0,050976) \\ &\quad + 0,281618 DR - 0,0784652 URB - 0,111495 WH \\ &\quad (0,13370) \quad (0,020624) \quad (0,024242) \\ &\quad T = 100 \quad \bar{R}^2 = 0,8423 \quad F(10, 89) = 53,871 \quad \hat{\sigma} = 2,1919 \\ &\quad \text{(Desviaciones típicas entre paréntesis)} \end{aligned}$$

□

(Lección 11) Ejercicio en clase. N- 3(e) El modelo reproduce más de un 85% de la varianza de $WLFP$. Los coeficientes de $EDUC$, DR , URB , y WH tienen los signos esperados y dichos signos no cambian de un periodo a otro. El efecto matrimonio es menor en 1990 que en 1980. Un 1% de incremento en MR reduce la participación

en un 0.207% en 1980 y sólo un 0.081% en 1990, lo que sugiere que en 1990 más mujeres mantuvieron su actividad laboral tras el matrimonio. El efecto “desaliento” debido a la tasa de desempleo es estadísticamente distinto en los dos años estudiados, habiendo aumentado en 1990 respecto a 1980. El cambio más llamativo es para el efecto del salario mediano de las mujeres... este cambio es difícil de explicar y probablemente se debe a la elevada correlación entre YF y $D90YF$. El test de multicolinealidad **vif** también alerta del problema.

□

(Lección 11) Ejercicio en clase. N- 4(d) El grupo de referencia son los hombres solteros (los que no aparecen explícitamente en el modelo).

Los hombres casados ganan aproximadamente un 21.3% más que los solteros (manteniendo fijos los niveles de educación, experiencia y antigüedad).

Como la variable dependiente está en logaritmos, el cálculo exacto del efecto de la variable ficticia es $100 * (\exp(\beta_2) - 1) = 23,6983\%$.

Del mismo modo, las mujeres casadas ganan un 19.8% menos que los hombres solteros; y las mujeres solteras ganan un 11% menos que los hombres solteros (calcule los porcentajes exactos).

□

(Lección 11) Ejercicio en clase. N- 4(e) Que tendíamos multicolinealidad perfecta, pues la suma de las cuatro variables sería siempre uno (como el término constante).

□

(Lección 11) Ejercicio en clase. N- 4(f) El intervalo al 95% es $[-0,014924 ; 0,19076]$ por lo que no es significativa la diferencia.

□

(Lección 11) Ejercicio en clase. N- 4(g) Si se incluye *singmale* y se excluye *marrfem* (que ahora será el grupo de referencia), el parámetro para *singfem* es significativa sólo al 10% (pero no al 5%).

□

(Lección 11) Ejercicio en clase. N- 4(h) $100 * (\exp(\beta_4) - 1) = 9.1898$, es decir aproximadamente un 9.2% más para las solteras.

El intervalo va de $100 * (\exp(c1) - 1)$ a $100 * (\exp(c2) - 1)$, donde $c1$ y $c2$ son los límites del intervalo al 95% para β_4 del último modelo; es decir $[-1,4813 ; 21,017]$.

□

(Lección 12) Ejercicio en clase. N- 3(i) Los errores absolutos medios de predicción (MAPE) son:

```
mapeA = 227,75136  
mapeB = 278,19832  
mapeC = 48,240362
```

□

(Lección 12) Ejercicio en clase. N- 4(b) *uhatMiles* es la parte de las millas recorridas que no se puede aproximar por la edad del coche (algo así como un indicador de la intensidad de uso del coche).

□

(Lección 12) Ejercicio en clase. N- 4(h) El regresor *uhatAge* reflejará es la parte del coste de mantenimiento debido al mero paso del tiempo (pues ya hemos descontado el uso del coche). Es una medida por el deterioro de coche por el simple paso del tiempo. Interpretar el efecto que tiene *uhatAge* sobre el coste de mantenimiento más natural que tenía *uhatMiles*.

□

(Lección 12) Ejercicio en clase. N- 4(k) En el cuarto modelo *ModeloB-Raux* todos los regresores (incluido *uhatAge*) tiene una interpretación natural; además, en ese modelo los regresores son significativos y con los signos esperados.

Nótese que como tanto el *ModeloC* como los dos con regresiones auxiliares emplean la misma información (las mismas variables), los tres tiene estadísticos de selección de modelos idénticos. El ajuste es igual de bueno en los tres casos, aunque la interpretación de resultados no es igual.

□

Ejercicio 2. Denominemos \mathbf{W} a la matriz cuyas primeras k columnas son las columnas de \mathbf{X} y cuyas p restantes columnas son las columnas de \mathbf{Z} , es decir, $\mathbf{W} = [\mathbf{X}; \mathbf{Z}]$. Calculando la esperanza condicionada tenemos

$$\mathbb{E}(\widehat{\boldsymbol{\beta}} | \mathbf{W}) = \boldsymbol{\beta} + (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \underbrace{\mathbb{E}(\mathbf{U} | \mathbf{W})}_{\neq \mathbf{0}}$$

por lo que las estimaciones serán sesgadas en general. Nótese que lo importante no es si las variables \mathbf{Z} son relevantes o no en el modelo (véase la última transparencia de la lección), lo crucial es que sean, o no sean, ortogonales a las perturbaciones.

Por otra parte, la matriz de varianzas y covarianzas del vector de perturbaciones sera

$$\text{Var}(\mathbf{U} | \mathbf{W}) = \mathbb{E}([\mathbf{U} - \mathbb{E}(\mathbf{U} | \mathbf{W})]^\top [\mathbf{U} - \mathbb{E}(\mathbf{U} | \mathbf{W})] | \mathbf{W}) = \boldsymbol{\Sigma} \neq \sigma^2 \mathbf{I}.$$

donde $\boldsymbol{\Sigma}$ es una matriz simétrica y definida positiva, pero en general ni siquiera diagonal. Por ello el Teorema de Gauss-Markov deja de tener efecto (observe como para su demostración se empleó el hecho de que $\text{Var}(\mathbf{U} | \mathbf{X}) = \sigma^2 \mathbf{I}$, una condición que no es válida si $\mathbb{E}(\mathbf{U} | \mathbf{W}) \neq \mathbf{0}$).

□

(L-13) Problema 1. Es fácil verificar que

$$\frac{1}{(\mathbf{X} \cdot \mathbf{X})(\mathbf{1} \cdot \mathbf{1}) - (\mathbf{1} \cdot \mathbf{X})^2} \begin{bmatrix} (\mathbf{X} \cdot \mathbf{X}) & -(\mathbf{1} \cdot \mathbf{X}) \\ -(\mathbf{1} \cdot \mathbf{X}) & (\mathbf{1} \cdot \mathbf{1}) \end{bmatrix} \text{ es la inversa de } \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} (\mathbf{1} \cdot \mathbf{1}) & (\mathbf{1} \cdot \mathbf{X}) \\ (\mathbf{X} \cdot \mathbf{1}) & (\mathbf{X} \cdot \mathbf{X}) \end{bmatrix}.$$

Así,

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} \beta_3 = \frac{1}{(\mathbf{X} \cdot \mathbf{X})(\mathbf{1} \cdot \mathbf{1}) - (\mathbf{1} \cdot \mathbf{X})^2} \begin{bmatrix} (\mathbf{X} \cdot \mathbf{X}) & -(\mathbf{1} \cdot \mathbf{X}) \\ -(\mathbf{1} \cdot \mathbf{X}) & (\mathbf{1} \cdot \mathbf{1}) \end{bmatrix} \begin{bmatrix} \mathbf{1} \cdot \mathbf{Z} \\ \mathbf{X} \cdot \mathbf{Z} \end{bmatrix} \beta_3$$

y por tanto, su valor esperado es

$$\begin{aligned} \mathbb{E}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} \beta_3\right) &= \frac{1}{N^2 \mathbb{E}(\mathbf{X}^2) - N^2 \mathbb{E}(\mathbf{X})^2} \begin{bmatrix} N \mathbb{E}(\mathbf{X}^2) & -N \mathbb{E}(\mathbf{X}) \\ -N \mathbb{E}(\mathbf{X}) & N \end{bmatrix} \begin{bmatrix} N \mathbb{E}(\mathbf{Z}) \\ N \mathbb{E}(\mathbf{X}) \mathbb{E}(\mathbf{Z}) \end{bmatrix} \beta_3 \\ &= \frac{1}{\mathbb{E}(\mathbf{X}^2) - \mathbb{E}(\mathbf{X})^2} \begin{bmatrix} \mathbb{E}(\mathbf{X}^2) & -\mathbb{E}(\mathbf{X}) \\ -\mathbb{E}(\mathbf{X}) & 1 \end{bmatrix} \begin{bmatrix} \mathbb{E}(\mathbf{Z}) \\ \mathbb{E}(\mathbf{X}) \mathbb{E}(\mathbf{Z}) \end{bmatrix} \beta_3 \\ &= \frac{1}{\text{Var}(\mathbf{X})} \begin{bmatrix} \mathbb{E}(\mathbf{X}^2) \mathbb{E}(\mathbf{Z}) - \mathbb{E}(\mathbf{X})^2 \mathbb{E}(\mathbf{Z}) \\ -\mathbb{E}(\mathbf{X}) \mathbb{E}(\mathbf{Z}) - \mathbb{E}(\mathbf{X}) \mathbb{E}(\mathbf{Z}) \end{bmatrix} \beta_3 \\ &= \frac{1}{\text{Var}(\mathbf{X})} \begin{bmatrix} \text{Var}(\mathbf{X}) \mathbb{E}(\mathbf{Z}) \\ \text{Cov}(\mathbf{X}, \mathbf{Z}) \end{bmatrix} \beta_3 = \begin{pmatrix} \mathbb{E}(\mathbf{Z}) \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})} \end{pmatrix} \beta_3 \end{aligned}$$

Consecuentemente, el valor esperado del estimador del modelo incompleto es

$$\mathbb{E}\left(\begin{pmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{pmatrix}\right) = \mathbb{E}\left(\mathbb{E}\left(\begin{pmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{pmatrix} \mid \mathbf{X}\right)\right) = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \mathbb{E}(\mathbf{Z}) \\ \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})} \end{pmatrix} \beta_3;$$

es decir $\mathbb{E}(\widehat{\beta}_1^*) = \beta_1 + \mathbb{E}(\mathbf{Z}) \beta_3$ y $\mathbb{E}(\widehat{\beta}_2^*) = \beta_2 + \frac{\text{Cov}(\mathbf{X}, \mathbf{Z})}{\text{Var}(\mathbf{X})} \beta_3$.

□