

Sergio CAMIZ

EL MODELO LINEAL

LIMA - Abril-Mayo 2010

Prefación

Estas notas son de soporte al Modulo *Regresión* por el Curso de *Técnicas de predicción* que dicté a la Pontificia Universidad Católica de Perú en Lima en 2010.

Se trata de la revisión y traducción en español de unas notas que yo hice en 1997 en ocasión de un curso parecido en la Université des Sciences et Technologies de Lille (Francia). Estas notas se basan libremente sobre el libro de Guttman (1982) que sigue a parecerme una óptima referencia para tratar separadamente los aspectos matemáticos, estadísticos y inferenciales del modelo lineal.

Con esto quiero dejar un testigo de mi estadía como agradecimiento por la invitación que he recibido, deseando que podrá eser útil en el futuro. Agradezco también al Dr. Luis Valdivieso, por su paciencia empleada para revisar mi texto español.

*Sergio Camiz*¹

¹ Profesor de Matemática y Estadística, PhD.
Dipartimento di Matematica «Guido Castelnuovo», Sapienza Università di Roma.
Página Web: <http://www.camiz.net>
Correo electrónico: sergio.camiz@uniroma1.it.

Índice

Prefación	I
Índice	III
1. Análisis de datos, estadística y modelización	1
1.1 Introducción	1
1.2 La etapa exploratoria	2
1.3 La etapa confirmatoria	2
1.4 La modelización	3
1.5 ¿Qué cosa es la estadística?	4
2. El modelo lineal	7
2.1 Introducción	7
2.2 Análisis del modelo lineal	8
2.3 Estimación de los parámetros	10
2.4 La solución de los mínimos cuadrados	12
2.5 La recta de los mínimos cuadrados	14
2.6 La recta de los mínimos cuadrados para el origen	15
3. Propiedades estadísticas	17
3.1 Propiedades de los estimadores de los mínimos cuadrados	17
3.2 El teorema de Gauss - Markov	20
3.3 Estimación de σ^2	22
3.4 Análisis de varianza del modelo	24
3.5 El modelo lineal centrado en	24
3.6 La solución de los mínimos cuadrados	25
3.7 Las estadísticas de los estimadores	26
3.8 Análisis de los residuos	28
3.9 Análisis de varianza	29
3.10 Comparación de modelos	30
3.11 La calidad del modelo	30
4. Inferencia del modelo lineal	32
4.1 Hipótesis de normalidad	32
4.2 La falta de ajuste del modelo lineal	36

4.3	La predicción	41
5.	La regresión lineal múltiple	43
5.1	Introducción	43
5.2	Estimación de los parámetros	45
5.3	La solución en el caso de rango completo	48
5.4	Propiedades estadísticas de los estimadores de mínimos cuadrados	49
5.5	El análisis de varianza del modelo	51
6.	Inferencia	54
6.1	Distribuciones normales multivariadas	54
6.2	Test del modelo y de los parámetros	55
6.3	Partición de la regresión	57
6.4	Dos conjuntos ortogonales	58
6.5	Caso no ortogonal	60
6.6	Observación común	62
6.7	La eliminación del promedio	62
6.8	La falta de ajuste del modelo lineal	64
7.	Selección de modelos	68
7.1	Las matrices de covarianza y de correlación	68
7.2	El coeficiente de correlación parcial	69
7.3	Caso de tres variables	70
7.4	Caso general	72
7.5	Técnicas de regresión	74
7.6	El registro exhaustivo (all possible regression)	74
7.7	Método paso a paso: selección para adelante	74
7.8	La selección para atrás	75
7.9	La regresión paso a paso (stepwise)	76
8.	Ejemplos	77
8.1	Una regresión simple	77
8.2	Una aplicación de la regresión simple	77
8.3	Un ejemplo de regresión múltiple	81
	Bibliografía	86

1. Análisis de datos, estadística y modelización

1.1 Introducción

Cuando se encuentra un fenómeno, un panorama, un objeto que estudiar, la *observación* y, en seguida, el descubrimiento de lo que se necesita anotar, dependen muy fuertemente de la *cultura* del observador. Efectivamente, un niño al momento de su nacimiento no sabe que mirar: tal vez el ve pero no mira. Con el pasar del tiempo, las informaciones que recibe a través de los otros sentidos, la voz de su madre, su calor, su olor, el sabor de su leche, se asocian y forma la primera base cultural del niño: la idea de su madre. Entonces, después el niño será capaz de mirar a su madre, de buscarla, y podrá interesarse a enriquecer su cultura con otros elementos que el considera interesantes y útiles de observar.

En todo momento la observación depende de la cultura del observador, pero esta misma es una fuente de cultura, si bien el proceso no resulta siempre inmediato. Si en la vida cotidiana esto puede ser real, esto no lo es en un estudio científico. En este caso, lo que se observa tiene que ser agrupado con el contenido de otras observaciones y solo con una síntesis se pueden conseguir algunas verdades. Llamaremos de ahora en adelante *datos* a los pedazos de información empírica que se juntan en una observación.

¿Como hacer para conseguir un conocimiento científico, o sea cultura, empezando de los datos observados?. El proceso puede ser muy largo, porque precisa pasar a través de diversas etapas para conseguir una información que se pueda llamar cultura. Tomamos como comienzo lo que M. Carbon escribe en su notas:

«La statistique a pour objet le recueil de données pour leur analyse et leur interprétations»² (M. Carbon).

Aquí se encuentra una definición de la estadística que cobra toda la actividad relacionada a conceptos empíricos empezando con los datos. Claro que todo esto depende en ancha medida del fin de una aplicación particular, porque sin aplicación la estadística no tiene sentido. De otro lado, se puede considerar una aplicación como un momento de construcción cultural sobre una base empírica. Se trata entonces de la cultura, o sea de la disciplina en la cual está encuadrado el problema específico que se va tratar, que de un lado va a enriquecer la nueva experiencia y de otro permita mejor encuadrar el problema mismo y contribuir a la fijación de la dirección exacta para resolverlo.

² «La Estadística tiene como objeto la recolección de datos por su análisis y su interpretación».

En realidad, actualmente parece que la creación y el empleo de métodos numéricos e informáticos para tratar cualquier tipo de datos hace más complejo el hablar simplemente de estadística como cualquier tipo de tratamiento, considerando también que la difusión de las computadoras en todos los lugares permitió de aplicar estos métodos en la mayoría de los asuntos de estudio, incluso en donde los métodos estadísticos nunca fueron empleados antes.

Entonces, se puede considerar un estudio como organizado en tres etapas: *exploratoria*, *confirmatoria* y de *modelización*.

1.2 La etapa exploratoria

En la *etapa exploratoria* se empieza definiendo un cuadro de referencia y los objetivos del estudio. En esta etapa se empieza con una primera recolección de datos seguida de su acopio en la computadora. Al principio se emplean *estadísticas descriptivas*, o sea el conjunto de técnicas de síntesis de la información relativa a cada carácter independientemente de los otros, para examinar los datos, sobre todo del punto de vista de la calidad de la adquisición. En seguida, el objetivo de una *análisis exploratoria de datos* es extraer de esta base de datos más información, fuertemente sintetizada, para conseguir indicaciones sobre las estructuras subyacentes al fenómeno, que pueden causar efectos sobre los datos mismos.

En particular, se van buscando *factores*, o sea algún agente tanto *endógeno* que *exógeno*, que influencia al desarrollo del fenómeno observado. Así se podrán *ordenar* las observaciones segundo estos factores.

También es práctico de *clasificar* las observaciones según *particiones*. Esto permite reconocer su diversidad y tratarlos de manera apropiada. En particular, las observaciones que pertenecen a la misma clase tendrán caracteres parecidos, un hecho que permite de considerarlos como *tipos*, con caracteres propios, diferentes de los de las otras clases.

En esta etapa se analizan los datos con herramientas de análisis exploratorio de datos multidimensional, cuyo estudio es de ayudar el estudio de factores, a través métodos de transformación y proyección, y de proponer particiones basadas sobre relaciones de proximidad. En esta análisis de datos no se utilizan modelos estadísticos sino se emplean modelos *cognitivos* geométricos, donde la búsqueda de representaciones con propiedades optimales, dependiendo de criterios establecidos, guía hasta la solución.

1.3 La etapa confirmatoria

En la etapa confirmatoria se busca sobre todo de adquirir conocimientos sobre el fenómeno que se estudia, o sea conocer las relaciones causa / efecto que interesan del fenómeno mismo, como también los tipos diferentes de resultados que resultan de esas. Este estudio es fuertemente afectado para los resultados de la etapa precedente, porque se supone que ya se encontraron estructuras y relaciones posibles entre elementos: en particular ya se pudieron identificar factores y particiones posibles. No obstante, en esta etapa se buscan contestaciones mas *ciertas* y *serias* sobre las posibles hipótesis que se pensaron sobre la base de los conocimientos adquiridos en la etapa precedente. Así se trata de *estimar* valores, definir *intervalos de confianza* para esos, *testar* hipótesis, y sobre todo hacer *inferencia estadística*, o sea desplegar los resultados conseguidos a observaciones *que no se hicieron*, pero en todo compatible con las que se hicieron. En esta etapa es importante considerar un problema de *muestreo*, o sea elegir un conjunto limitado de observaciones que hacer, pero capaz de informar sobre un super-conjunto de informaciones *posibles*, que se indica como la *población de referencia*.

Hay tres elementos que contribuyen a definir la manera de buscar una solución a los problemas puestos en esta etapa: la *precisión* de los resultados, la *muestra* de observaciones y los *costos* de las faltas que se pueden hacer, debido a la precisión reducida, a un muestreo no adecuado; en fin a la falta de informaciones suficientes sobre el fenómeno debido a diversas causas difíciles de determinar.

Es por esta razón que se intenta de ligar la precisión de los resultados a los costos de las faltas posibles dependiendo de las muestras empleadas. De esta manera se intenta de garantizarse contra faltas o costos inaceptables. Sobre todo en esta etapa se hace *estadística* en el sentido mas autentico.

1.4 La modelización

En esta etapa se supone de conocer suficientemente bien el fenómeno y de poder representarlo formalmente a través de un sistema de ecuaciones matemáticas. Se trata de un modelo teórico, que explica suficientemente bien el fenómeno desde el punto de vista de las relaciones causales entre caracteres sin perder de vista los paradigmas teóricos que forman la disciplina de referencia.

Empleando estas ecuaciones se puede *simular* el fenómeno, o sea se pueden evaluar medidas observables en función de los factores empleados: así se pueden también preveer resultados de una situación específica. Debe de observar que a veces en la práctica no se espera un modelo teórico para hacer previsiones u estimaciones de los valores de un carácter si o no medible: a veces se limita a emplear modelos estadísticos que no pretenden explicar el fenómeno, sino de calcular los valores que se van buscando.

Ejemplo: si se encuentra una relación fuerte entre consumo de electricidad y radiación solar, se puede evaluar la radiación a través el consumo de electricidad, aunque el consumo de electricidad no tiene efecto sobre la

actividad del sol.

Para hacer modelos, se emplea sobre todo alguna *matemática*: pero, sin embargo análisis de datos y estadística tienen su importancia para la calibración de los modelos y la evaluación de los resultados.

Una vez que el modelo está diseñado, el estudio se puede decir concluido. Pero se puede empezar de nuevo para estudiar detalles que no fueron bastante considerados, mejorar la precisión de los resultados (luego reducir los costos), etc. Así se re-empieza siguiendo la tres etapas en una especie de estructura fractal de estudio. El modelo en tres etapas, aunque no siempre identificable, se puede emplear para comprender tanto la evolución de un gran campo de investigación como de una investigación pequeña en un contexto más general.

Es muy importante de enfatizar que el conocimiento del nivel del estudio es indispensable para decidir el comportamiento del investigador en relación con los datos. Efectivamente, solo el conocimiento de dichas etapas puede indicar al investigador la herramienta que debe emplear o no emplear: en particular la herramienta estadística no puede emplearse en la etapa exploratoria, porque a menudo la muestra no resulta adecuada a la aplicación de *test estadísticos*. Al contrario, la herramienta de análisis exploratoria no permite ninguna *inferencia estadística*, o sea de atribuir los resultados conseguidos a una población de referencia.

La secuencia de las tres etapas puede conducir a la construcción de un modelo sin los riesgos de utilizar un modelo definido *a priori*, sin una crítica eficaz (Benzécri *et al.*, 1982).

En nuestro estudio, será interesante reconocer que el modelo lineal tiene relaciones con todas las etapas. Efectivamente una regresión tiene un sentido exploratorio, para averiguar la relación que se puede suponer entre dos caracteres; confirmatorio, estimando dicha relación y testando los parámetros para poder hacer inferencia sobre estos; en fin, el modelo lineal constituye una parte muy larga de la modelización matemática.

1.5 ¿Qué cosa es la estadística?

Ya se vio que es la estadística segundo M. Carbon. Se puede considerar una definición un poco más extensa que no considere las diferencias estructurales de las tres etapas que se describieron en el párrafo precedente. Para comprender mejor la situación, citamos aquí algunas definiciones que pueden dar una idea de la estadística según autores diferentes.

«On peut définir la *statistique* comme l'ensemble des méthodes qui permettent, à partir de l'observation d'un phénomène aléatoire, d'obtenir des *informations* sur la probabilité associée à ce

phénomène.»³ (D. Bosq).

El carácter aleatorio del fenómeno es la traducción que ignora las leyes que gobiernan el fenómeno mismo. A este propósito, Bosq sugiere que

«...une étude préliminaire, ne tenant compte que des observations effectuées, peut se révéler intéressante.»⁴ (*ibid.*)

así separando el análisis de datos de la estadística.

«On se trouve devant un problème statistique si:

- on est confronté à des éventualités (en nombre fini ou infini) dont on sait que certaines sont vraies sans savoir lesquelles;
- on doit choisir une de ces éventualités ...
- ... en s'appuyant sur le résultat d'une expérience aléatoire, éventuellement à définir.»⁵ (Monfort, 1982).

Se trata de un enfoque *decisional* a la Estadística. Claro que no es el solo posible, porque un problema estadístico puede también ser concebido como un problema de información, sin depender de alguna decisión, con la ventaja de una presentación unificada de los problemas. Tratando de experiencias aleatorias se sitúa en el cuadro de la estadística inductiva, lejos del análisis de datos.

«Le matériel brut d'une investigation statistique est un ensemble d'observations: il s'agit de valeurs prises par des variables aléatoires X dont la distribution P_θ est au moins partiellement inconnue. On connaît du paramètre θ , qui caractérise la distribution, seulement qu'il appartient à un certain ensemble Ω , l'espace des paramètres. L'inférence statistique concerne les méthodes, qui utilisent le matériel des observations, pour obtenir des informations sur la distribution de X ou sur le paramètre θ .

La nécessité d'une analyse statistique dérive du fait que la distribution de X , et par conséquent les aspects de la situation sous-jacente le modèle mathématique, n'est pas connue.»⁶ (Lehmann, 1983).

³ «Se puede definir la *estadística* como el conjunto de métodos que permiten, comenzando para la observación de un fenómeno aleatorio, conseguir *informaciones* sobre la probabilidad asociada a dicho fenómeno.»

⁴ «...un estudio preliminar, que no considere que las observaciones efectuadas, puede revelarse interesante...»

⁵ «Se encuentra en frente de un problema estadístico si:

- se encuentra confrontados a eventualidades (en número finito o infinito) del cual se sabe que algunas son verdaderas sin saber cuales;
- hay que elegir una entre estas eventualidades ...
- ... apoyándose sobre el resultado de una experiencia aleatoria, que eventualmente hay que definir.»

⁶ «El material bruto de una investigación estadística es un conjunto de observaciones: se trata de valores tomados para algunas variables aleatorias X cuya distribución P_θ es a menos parcialmente desconocida. Del parámetro θ , que caracteriza la distribución, solo se conoce si pertenece a un dado conjunto Ω , el espacio de los parámetros. La inferencia estadística concierne los métodos, que utilizan el material de las observaciones, para conseguir informaciones sobre la distribución de X u sobre el parámetro θ .

La necesidad de un análisis estadístico deriva de lo hecho que la distribución de X , y por consiguiente de los aspectos de la situación subyacentes al modelo matemático, no es conocida.»

Acá todavía se sitúa muy lejos del análisis de datos, lo que nos indica que hay una diferencia muy fuerte entre los dos enfoques.

Efectivamente, hacer estadística puede significar para nosotros situarnos en la etapa confirmatoria de un estudio, cuando necesitamos confirmar o reprobar algunas hipótesis e inferir los resultados a las poblaciones de referencia donde las muestras estudiadas se extrajeron. Ya se observó que también en la construcción de un modelo matemático se emplean métodos estadísticos para evaluar y comparar la calidad de los modelos que se han construido.

El estudio del modelo lineal que vamos a hacer se va a desarrollar dividiendo muy claramente el tema cognitivo, de carácter exploratorio, que se hace construyendo un modelo matemático de la relación estudiada, del tema confirmatorio, o sea de la evaluación de la calidad de los resultados y de su fiabilidad, que se resuelve con métodos estadísticos.

2. El modelo lineal

2.1 Introducción

El *modelo lineal* o *modelo de regresión* es uno de los métodos más empleados en análisis confirmatoria de datos. Su campo de aplicación se encuentra por el interés de buscar si existe una relación de dependencia entre un carácter *respuesta* o *dependiente*, indicado por η , y otros caracteres independientes dichos *explicativos* o *predictivos* z_1, z_2, \dots, z_s , así que se puede escribir

$$\eta = f(z_1, z_2, \dots, z_s) \quad (1)$$

Ejemplo 1: en un proceso de producción, se puede imaginar que la cantidad producida depende del número de obreros empleados, de la electricidad consumada, de la cantidad de los diversos componentes adquiridos, etc.

Ejemplo 2: la velocidad al descolaje de un avión depende de su peso, del peso de la tripulación, del número de los pasajeros, de los equipajes, de la cantidad de gasolina. Asimismo, la gasolina que necesita depende del largo del viaje [teniendo en cuenta también la gasolina cargada...], de la altura y de la velocidad del vuelo.

Se dice que η sigue un modelo lineal si

$$\eta = f(z_1, z_2, \dots, z_s) = \sum_{i=1}^k \beta_i x_i(z_1, z_2, \dots, z_s) \quad (2)$$

donde los x_j son funciones solo de los z_i . Las cantidades β_j son *parámetros* en principio desconocidos que aparecen linealmente en la ecuación (2).

Ejemplo 3: $\eta = \alpha + \beta z$ es un modelo lineal, con $s = 1, k = 2, \beta_1 = \alpha, \beta_2 = \beta, x_1(z) = 1, x_2(z) = z, z_1 = z$. A veces se lo escribe $\eta = \alpha + \beta x$, con $x = x_2(z) = z$. Los parámetros $(\beta_1, \beta_2) = (\alpha, \beta)$ entran linealmente en el modelo.

Ejemplo 4: $\eta = \tau_0 + \tau_1 z_1 + \tau_2 z_2 + \dots + \tau_d z_d$ es un modelo lineal, poniendo $\beta_j = \tau_{j-1}, x_j = x_j(z) = z_{j-1}, j = 1, \dots, d + 1, s = 1, k = d + 1$. Los β_j entran linealmente en el modelo.

Ejemplo 5: una relación polinomial entre η y $z, \eta = \alpha + \tau_1 z^1 + \tau_2 z^2 + \dots + \tau_d z^d$ es un modelo lineal poniendo $\beta_j = \tau_{j-1}, x_j = x_j(z) = z^{j-1}, j = 1, \dots, d + 1, s = 1, k = d + 1$. Los β_j entran linealmente en el modelo.

Ejemplo 6: una relación polinomial entre η y z_1 y $z_2, \eta = \tau_{00} + \tau_{10} z_1^1 + \tau_{01} z_2^1 + \tau_{20} z_1^2 + \tau_{11} z_1 z_2 + \tau_{02} z_2^2 + \dots$ es un modelo lineal poniendo $\beta_1 = \tau_{00}, \beta_2 = \tau_{10}, \dots, x_1 = 1, x_2(z) = z_1, \dots$ etc. Los β_j entran linealmente en el modelo.

Ejemplo 7: $\eta = \alpha + \beta \sin 2\pi z$ es un modelo lineal, con $s = 1, k = 2, \beta_1 = \alpha, \beta_2 = \beta, x_1(z) = 1, x_2(z) = \sin 2\pi z, z_1 = z$. Los parámetros $(\beta_1, \beta_2) = (\alpha, \beta)$ entran linealmente en el modelo.

Ejemplo 8: al contrario, $\eta = \frac{e^{\beta_1 z_1} - e^{\beta_2 z_2}}{\beta_2 - \beta_1}$ no es un modelo lineal, porque los parámetros (β_1, β_2) no entran linealmente en el modelo.

Teniendo en cuenta los ejemplos y el hecho que cada función se puede aproximar por un polinomio según la fórmula de Taylor, se puede comprender la potencia del modelo lineal, aunque de esta manera se puede perder en simplicidad y a veces en calidad. Además, algunos modelos no lineales en los parámetros pueden volverse lineales a través alguna *transformación*.

Ejemplo 9: $\xi = \delta e^{\gamma z}$ no es lineal en γ , pero si se toman los logaritmos resulta $\log \xi = \log \delta + \gamma z$ así que se lo devuelve, poniendo $\eta = \log \xi$, $\beta_1 = \log \delta$, $\beta_2 = \gamma$, $x_1(z) = 1$, $x_2(z) = z$, etc.

Sin embargo, en estos casos uno tiene que ser muy cuidadoso y emplear reglas muy precisas cuando se hace inferencia a la población, en base a los resultados de una muestra.

2.2 Análisis del modelo lineal

Vemos ahora mas de cerca como se analiza el modelo lineal más sencillo, tal que

$$\eta = \alpha + \beta x \quad (3)$$

Ejemplo 10: Se pregunta como la temperatura hace variar el volumen de un gas. Entonces se hace una experimentación poniendo algún gas en un cilindro y se miden los volúmenes y_i correspondientes a las temperaturas x_i , $i = 1, 2, \dots, n$, antes fijadas.

Si se piensa que la relación entre y y x es lineal (a parte los errores experimentales) se puede decir que la *esperanza de y dado x* , $E(y|x)$ es la *función de regresión de y sobre x*

$$E(y | x) = \eta_x = \alpha + \beta x \quad (4)$$

bajo la condición que el *error experimental de medida de y* sea el mismo por cada x , que se lo expresa imponiendo que la *varianza de y* sea constante por cada x :

$$V(y | x) = \sigma^2, \forall x \quad (5)$$

La condición (5) se llama *hipótesis de homocedasticidad*.

Nota 1: por un valor dado de x , las correspondientes observaciones de y varían según una distribución D , por ejemplo la distribución normal N . La varianza de dicha distribución de y para x dada es indicada con σ^2 . Esto se puede sintetizar escribiendo

$$\text{Dado } x, y = D(\alpha + \beta x, \sigma^2) \quad (6)$$

desde que la condición (4) dice que la esperanza de y se encuentra sobre la recta de regresión $\alpha + \beta x$.

Nota 2: para otra x' , $x' \neq x$ se encuentra igualmente $y = D(\alpha + \beta x', \sigma^2)$, o sea, a parte la posición medida para las esperanzas de las distribuciones, las distribuciones de y para x y para x' son diferentes solo para su posición. Esto está representado en la Fig. 1.

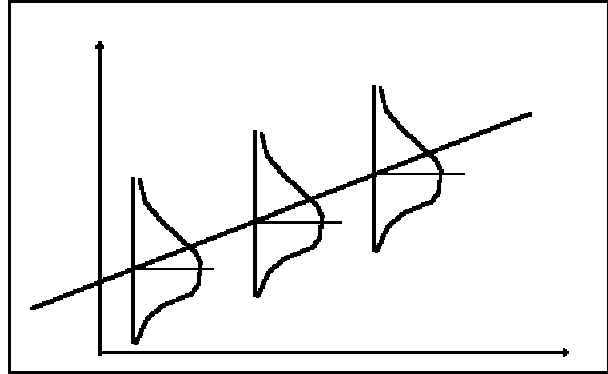


Fig. 1 - Distribución de las y_i por diferentes x_i .

Se sintetizan las dos notas con la escritura

$$y = \alpha + \beta x + \varepsilon, \quad \forall x \quad (7)$$

donde el error aleatorio ε tiene la misma distribución para cualquier x , con $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$. Se presupone también que la observación y_i correspondiente a x_i es independiente de la observación y_j correspondiente a x_j .

Entonces, el modelo de nuestro experimento es dado por:

$$\begin{cases} E(y | x) = \eta_x = \alpha + \beta x \\ V(y | x) = \sigma^2 \\ y_i \text{ y } y_j \text{ independientes por cada } i \neq j \end{cases} \quad (8)$$

Esta elección de modelo tal vez se hace porque: 1) ya se sabe que la relación es lineal; 2) en la región de las elecciones usuales de x la relación lineal es muy buena aproximación; 3) se busca una relación funcional entre y y x y se utiliza un modelo lineal como primera etapa de investigación. En todo caso encuentra en una etapa de análisis confirmatorio.

También el modelo se puede escribir así:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i \\ E(\varepsilon_i) = 0 \\ V(\varepsilon_i) = \sigma^2 \\ \varepsilon_i \text{ y } \varepsilon_j \text{ independientes por cada } i \neq j \end{cases} \quad (9)$$

Tiene que observar que algunos x_i pueden ser iguales, pero por lo menos dos tienen que ser diferentes por un asunto técnico que se estudiará en seguida, pero es evidente que no se podrá estimar una recta con solo una x_i . De otro lado no tiene sentido buscar relaciones funcionales entre x e y limitándose a una sola x . En realidad, para una relación que tenga sentido dos x_i no son suficientes, pero se puede establecer fácilmente un número mínimo. En la práctica, teniendo que inferir los resultados a una población de referencia, se establecerá el número de observaciones sobre el tamaño de la población y de los errores admitidos (ver, por ejemplo, Mood et al., 1974).

2.3 Estimación de los parámetros

Ahora se busca un método que permita estimar los parámetros desconocidos α y β en función de x , y y del error experimental σ^2 . Se puede proceder de la manera siguiente: supongamos que sobre una base cualquiera se estiman α , β con α^e , β^e . Esto implica que se estima la función de regresión (supuestamente lineal) como

$$\eta_{x_i}^e = \alpha^e + \beta^e x_i, \quad i = 1, 2, \dots, n \quad (10)$$

Se observa que los puntos sobre esta recta de regresión correspondientes a los valores tienen como coordenadas $(x_i, \eta_{x_i}^e)$ (Fig. 2), así que la recta tiene como ecuación

$$\eta^e = \alpha^e + \beta^e x \quad (11)$$

Tiene que observarse que con esta escritura no se pueden representar rectas verticales: pero esto no es un límite, porque dicha recta solo sería una indicación de independencia entre caracteres.

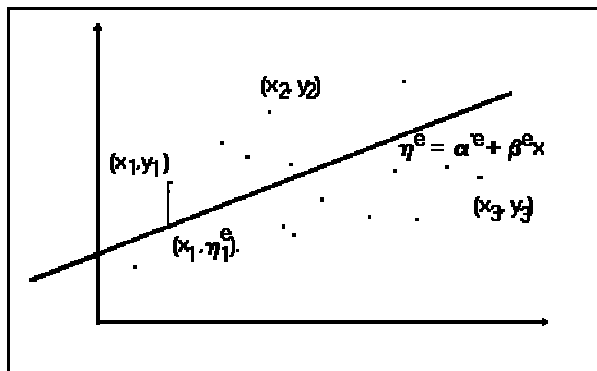


Fig. 2 - Puntos observados y recta de regresión.

Un método para medir la calidad de la estimación consiste a buscar la distancia entre la recta (9) y los datos observados. Una manera de medir esta distancia consiste en la suma de los cuadrados de las distancias verticales, o sea el cuadrado de la desviación entre puntos observados (x_i, y_j) y los puntos sobre la recta correspondientes a los mismos x_i , cuyas coordenadas son $(x_i, \eta_{x_i}^e)$, lo que deja⁷:

⁷ De ahora en adelante, se escribirán las sumas \sum_i en lugar de $\sum_{i=1}^n$, porque son casi todas iguales,

$$SS_e = \sum_i (y_i - \eta^e_{x_i})^2 = \sum_i (y_i - \alpha^e - \beta^e x_i)^2$$

Sea ahora una otra estimación de la recta, dada para

$$SS_f = \sum_i (y_i - \eta^f_{x_i})^2 = \sum_i (y_i - \alpha^f - \beta^f x_i)^2$$

Es evidente que la elección entre (α^e, β^e) y (α^f, β^f) depende del tamaño de SS_e y SS_f ; si $SS_e < SS_f$, se va a elegir el par de estimadores (α^e, β^e) ; si $SS_e = SS_f$, la elección es indiferente; si $SS_e > SS_f$, se preferirá el par de estimadores (α^f, β^f) , así que entre la infinidad de rectas de regresión posibles (**Fig. 3**), se elegirá la estimación $(\hat{\alpha}, \hat{\beta})$ correspondiente a la recta

$$\hat{\eta} = \hat{\alpha} + \hat{\beta}x \quad (12)$$

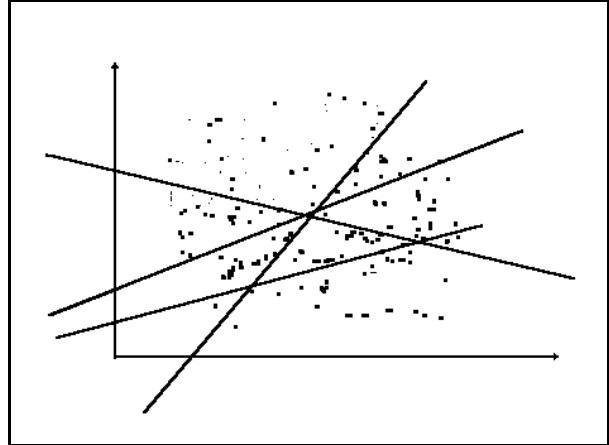


Fig. 3 - Algunas rectas de regresión posibles.

que resulte la mas cercana de los datos.

Se pueden considerar diferentes métodos para buscar dicha recta, pues se puede definir la distancia de la recta a los puntos de varias maneras. En la práctica, se emplea sobre todo el principio de los mínimos cuadrados, consistente en minimizar la suma de los cuadrados de las desviaciones de los puntos de la recta SS_e , por tanto a elegir $(\hat{\alpha}, \hat{\beta})$ como estimadores de (α, β) donde

$$SS_e = \min_{(\alpha, \beta)} SS_e(\alpha, \beta) = \min_{(\alpha, \beta)} \sum_i (y_i - \alpha - \beta x_i)^2 = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \quad (13)$$

Con esta condición, se obtiene la solución

extendidas a los n valores considerados. En seguida, las sumas mas frecuentes serán abreviadas con algunas S con indices. Así sera $S_x = \sum_i x_i$, $S_y = \sum_i y_i$; de manera análoga se indicaran las sumas de cuadrados con $S_{xx} = \sum x_i^2$, $S_{xy} = \sum_i x_i y_i$. Si se pusiera un punto sobre la variable, la suma es centrada alrededor de su promedio, o sea $S_{\bar{x}\bar{x}} = \sum_i (x_i - \bar{x})^2$, $S_{\bar{x}\bar{y}} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$. También sera útil en lo que sigue tener presente la relación $S_{\bar{x}\bar{y}} = S_{xy} - nS_x S_y$ y todas las que resultan. En particular, se observa que $S_{\bar{x}\bar{y}} = \sum_i (x_i - \bar{x})y_i = \sum_i (x_i y_i - n\bar{x}\bar{y}) = S_{xy}$.

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \end{cases} \quad (14)$$

que se puede escribir de manera sintetizada como:

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \frac{S_{xy}}{S_{xx}} \end{cases} \quad (15)$$

2.4 La solución de los mínimos cuadrados

Se van buscando α y β que minimizan SS_e :

$$SS_e = \min_{(\alpha, \beta)} SS_e(\alpha, \beta) = \min_{(\alpha, \beta)} \sum_i (y_i - \alpha - \beta x_i)^2 \quad (16)$$

con los x_i no todos iguales. Entonces, deben calcularse las derivadas parciales e igualarlas a cero:

$$\begin{cases} \frac{\partial SS_e(\alpha, \beta)}{\partial \alpha} = -2 \sum_i (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial SS_e(\alpha, \beta)}{\partial \beta} = -2 \sum_i (y_i - \alpha - \beta x_i)x_i = 0 \end{cases} \quad (17)$$

De este resulta el sistema lineal en α, β :

$$\begin{cases} n\hat{\alpha} + S_x \hat{\beta} = S_y \\ S_x \hat{\alpha} + S_{xx} \hat{\beta} = S_{xy} \end{cases} \quad (18)$$

Las ecuaciones (18) se llaman ecuaciones normales. Su solución existe en cuanto los x_i no son todos igual y por tanto

$$\det \begin{pmatrix} n & S_x \\ S_x & S_{xx} \end{pmatrix} = nS_{xx} - S_x^2 = nS_{\hat{x}\hat{x}} \neq 0 \quad (19)$$

La solución del sistema se encuentra con el método de Gauss:

$$\begin{cases} \hat{\alpha} = \frac{S_y}{n} - \frac{S_x}{n}\hat{\beta} = \bar{y} - \bar{x}\hat{\beta} \\ S_x(\bar{y} - \bar{x}\hat{\beta}) + S_{xx}\hat{\beta} = S_{xy} \end{cases} \quad (20)$$

$$\begin{cases} \hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} \\ \bar{y}S_x - \frac{1}{n}S_x^2\hat{\beta} + S_{xx}\hat{\beta} = S_{xy} \end{cases} \quad (21)$$

$$\begin{cases} \hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} \\ \hat{\beta} = \frac{S_{xy} - nS_x\bar{y}}{S_{xx} - nS_x^2} = \frac{S_{\hat{x}\hat{y}}}{S_{\hat{x}\hat{x}}} \end{cases} \quad (22)$$

Para averiguar que se trata de un mínimo, es suficiente de calcular el determinante Jacobiano de las derivadas segundas de $SS_e(\alpha, \beta)$, o sea

$$2 \begin{pmatrix} n & S_x \\ S_x & S_{xx} \end{pmatrix} = 2(nS_{xx} - S_x^2) = 2nS_{\hat{x}\hat{x}} > 0 \quad (23)$$

En realidad, el valor de $ss_e(\hat{\alpha}, \hat{\beta})$ estimado no solo es un mínimo *local*, sino también *global*. Efectivamente, si suponemos que tenemos otra estimación (α', β') , resulta que

$$\begin{aligned} SS_e(\alpha', \beta') &= \sum_i ((y - \alpha - \beta x_i) + (\alpha - \alpha') + (\beta - \beta')x_i)^2 \\ &\quad + 2 \sum_i (y - \alpha - \beta x_i)((\alpha - \alpha') + (\beta - \beta')x_i) \\ &= \sum_i (y - \alpha - \beta x_i)^2 + \sum_i ((\alpha - \alpha') + (\beta - \beta')x_i)^2 \end{aligned}$$

porque el doble producto de la segunda línea es cero (desarrollando, se encuentran las derivadas parciales de $SS_e(\alpha, \beta)$ que son cero). Entonces, se consigue el mínimo cuando en la última línea la segunda suma es cero, o sea cuando $\hat{\alpha} = \alpha'$ et $\hat{\beta} = \beta'$. ■

Claro que el termino *solución de los mínimos cuadrados* se puede entender porque se diferencié la suma $SS_e(\alpha, \beta)$ de los cuadrados y la solución buscada se consiguió para su mínimo.

2.5 La recta de los mínimos cuadrados

Conocida la estimación de los mínimos cuadrados $(\hat{\alpha}, \hat{\beta})$ de (α, β) , la estimación de $E(y|x) = \eta_x = \alpha + \beta x$ es representada por la *recta de los mínimos cuadrados*

$$\hat{\eta} = \hat{\alpha} + \hat{\beta}x \quad (24)$$

donde

$$\begin{cases} \hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} \\ \hat{\beta} = \frac{S_{xy}}{S_{xx}} \end{cases} \quad (25)$$

Vale señalar que si $x = \bar{x}$, resulta

$$\hat{\eta}_{\bar{x}} = \hat{\alpha} + \hat{\beta}\bar{x} = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\bar{x} = \bar{y} \quad (26)$$

y por eso la recta de los mínimos cuadrados pasa para el *baricentro* de los datos (\bar{x}, \bar{y}) .

Los puntos sobre la recta correspondiente a los valores x_i tienen como coordenadas $(x_i, \hat{\eta}_{x_i})$, por lo que se tiene que

$$SS_e = SS_e(\hat{\alpha}, \hat{\beta}) = \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \sum_i (y_i - \hat{\eta}_{x_i})^2 = \min_{(\alpha, \beta)} SS_e(\alpha, \beta) \quad (27)$$

La mínima suma SS_e , sobre cualquier elección de (α, β) , se llama *suma de los cuadrados de los residuos* (en inglés, *error sum of squares*, de allí la abreviación), y a

$$e_i = y_i - \hat{\eta}_{x_i} = y_i - \hat{\alpha} - \hat{\beta}x_i \quad (28)$$

e_i se llama *residuo* de y_i y puede considerarse como la cantidad residual que resulta por la

substitución del valor observado y_i con la estimación $\hat{\eta}_{x_i} = \alpha + \beta x_i$.

Observamos también que la suma de los residuos es cero:

$$\begin{aligned} S_e &= \sum_i (y_i - \alpha - \beta x_i) = \\ &= \sum_i (y_i - \bar{y} + \beta \bar{x} - \beta x_i) = \\ &= \sum_i (y_i - \bar{y}) - \beta \sum_i (x_i - \bar{x}) = 0 \end{aligned} \quad (29)$$

y en consecuencia, de $S_e = \sum_i (y_i - \alpha - \beta x_i) = \sum_i (y_i - \hat{\eta}_{x_i}) = 0$, resulta que $\bar{y}_i = \sum \hat{\eta}_{x_i} / n$, o sea que el *promedio de los valores estimados coincide con el promedio de los valores observados*.

2.6 La recta de los mínimos cuadrados para el origen

Pueden encontrarse situaciones en las cuales se necesita imponer que la recta de regresión pase por el origen. En este caso se utiliza un modelo lineal sin α , pero en el que la suma de los residuos no es necesariamente cero. No obstante se consigue el resultado siguiente:

Teorema: En el modelo lineal

$$y - \beta x = \varepsilon, \quad \forall x \quad (30)$$

la suma de los residuos $S_e = \sum_i (y_i - \beta x_i) = 0$ si $\bar{y} = \bar{x} = 0$.

Prueba. Resulta que

$$SS_e(\beta) = \sum_i (y_i - \beta x_i)^2 \quad (31)$$

con los x_i no todos iguales y se va buscando β tal que $SS_e(\beta)$ sea mínimo. Por esto se pone la derivada de SS_e en respecto de β igual a cero

$$\frac{dQ(\beta)}{d\beta} = -2 \sum_i (y_i - \beta x_i) x_i = 0 \quad (32)$$

donde se consigue la ecuación normal en β : $S_{xx} \hat{\beta} = S_{xy}$ y su solución

$$\beta = \frac{S_{xy}}{S_{xx}} \quad (33)$$

Esto implica que

$$\begin{aligned} S_e &= \sum_i (y_i - \hat{\eta}_{x_i}) = \sum_i (y_i - \beta x_i) = \sum_i \left(y_i - \frac{S_{xy}}{S_{xx}} x_i \right) \\ &= \sum_i y_i - \frac{S_{xy}}{S_{xx}} \sum_i x_i = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} = \bar{y} - \beta \bar{x} = \bar{y} - \frac{\sum_i \hat{\eta}_{\bar{x}}}{n} \end{aligned} \quad (34)$$

En este caso, no se ha dicho que S_e sea 0, sino cuando el origen es el baricentro de los puntos observaciones, o sea si ambas variables están centradas⁸. Sin esta condición resulta que no hay coincidencia entre el promedio de los y_i observados y el promedio de su estimadores, así que S_e indica precisamente su desviación. ■

⁸ El solo otro caso en el cual $S_e = 0$ es evidentemente cuando $\alpha = 0$ resulta sin restricción sino como resultado de la aplicación del modelo a un caso específico.

3. Propiedades estadísticas

3.1 Propiedades de los estimadores de los mínimos cuadrados⁹

El $\hat{\beta}$ se puede expresar como combinación lineal de los y_i , mas precisamente $\hat{\beta} = \sum_i c_i y_i$, donde $c_i = \frac{x_i - \bar{x}}{S_{xx}}$; es decir

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}} \quad (35)$$

Se puede observar que los c_i son constantes porque los x_i son valores preseleccionados. Estos tienen las propiedades siguientes:

$$\begin{aligned} \sum_i c_i &= 0 \\ \sum_i c_i^2 &= \frac{1}{S_{xx}} \\ \sum_i c_i x_i &= 1 \end{aligned} \quad (36)$$

Así también $\hat{\alpha}$ y $\hat{\eta}_x$ se pueden expresar como combinación lineal de los y_i , pues resulta que

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = \frac{\sum_i y_i}{n} - \sum_i c_i y_i = \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) y_i \quad (37)$$

$$\hat{\eta}_x = \hat{\alpha} + \hat{\beta} x = \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) y_i + \sum_i c_i y_i x = \sum_i \left(\frac{1}{n} + (x - \bar{x}) c_i \right) y_i \quad (38)$$

⁹ En seguida se utilizaran las relaciones siguientes. Para su prueba, verse por ejemplo Mood *et al.* (1974):

$$E(ax + b) = aE(x) + b$$

$$E(x \pm y) = E(x) \pm E(y)$$

$$E(xy) = E(x)E(y) + \text{cov}(x, y)$$

$$E(x^2) = (E(x))^2 + V(x)$$

$$V(ax + b) = a^2 V(x)$$

$$V(x \pm y) = V(x) + V(y) \pm 2 \text{cov}(x, y)$$

$$V(xy) = V(x)V(y) + E(x)E(y)$$

$$V(x^2) = (V(x))^2 + (E(x))^2$$

Este resultado nos permite calcular las esperanzas de $\hat{\alpha}$, $\hat{\beta}$ y $\hat{\eta}_x$ y su varianzas. Para $\hat{\beta}$ la esperanza vale

$$\begin{aligned}
 E(\hat{\beta} | x_1, \dots, x_n) &= \sum_i c_i E(y_i | x_1, \dots, x_n) \\
 &= \sum_i c_i E(y_i | x_i) \\
 &= \sum_i c_i \eta_i = \sum_i c_i (\alpha + \beta x_i) \\
 &= \alpha \sum_i c_i + \beta \sum_i c_i x_i = 0 + \beta 1 = \beta
 \end{aligned} \tag{39}$$

y, teniendo en cuenta la condición de homoscedasticidad, su varianza vale

$$\begin{aligned}
 V(\hat{\beta} | x_1, \dots, x_n) &= \sum_i c_i^2 V(y_i | x_1, \dots, x_n) \\
 &= \sum_i c_i^2 V(y_i | x_i) \\
 &= \sum_i c_i^2 \sigma^2 = \sigma^2 \sum_i c_i^2 = \frac{\sigma^2}{S_{\hat{x}\hat{x}}}
 \end{aligned} \tag{40}$$

Para $\hat{\alpha}$, su esperanza vale

$$\begin{aligned}
 E(\hat{\alpha} | x_1, \dots, x_n) &= \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) E(y_i | x_1, \dots, x_n) \\
 &= \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) E(y_i | x_i) \\
 &= \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) (\alpha + \beta x_i) \\
 &= \alpha \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) + \beta \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) x_i \\
 &= \alpha \left(\sum_i \frac{1}{n} - \bar{x} \sum_i c_i \right) + \beta \left(\sum_i \frac{x_i}{n} - \bar{x} \sum_i c_i x_i \right) \\
 &= \alpha (1 - 0) + \beta (\bar{x} - \bar{x}) = \alpha
 \end{aligned} \tag{41}$$

y su varianza

$$\begin{aligned}
 V(\hat{\alpha} | x_1, \dots, x_n) &= \sum_i \left(\frac{1}{n} - \bar{x} c_i \right)^2 V(y_i | x_i) \\
 &= \sigma^2 \sum_i \left(\frac{1}{n} - \bar{x} c_i \right)^2 \\
 &= \sigma^2 \left(\sum_i \frac{1}{n^2} - \frac{2\bar{x}}{n} \sum_i c_i + \bar{x}^2 \sum_i c_i^2 \right) \\
 &= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{n} \mathbf{0} + \frac{\bar{x}^2}{S_{xx}} \right) \\
 &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)
 \end{aligned} \tag{42}$$

En consecuencia, se puede también calcular la esperanza de $\hat{\eta}_x$ como

$$E(\hat{\eta}_x) = E(\hat{\alpha} + \hat{\beta}x) = E(\hat{\alpha}) + E(\hat{\beta})x = \alpha + \beta x = \eta_x \tag{43}$$

y su varianza por

$$\begin{aligned}
 V(\hat{\eta}_x) &= V(\hat{\alpha} + \hat{\beta}x) \\
 &= V\left(\sum_i \left(\frac{1}{n} + (x - \bar{x})c_i \right) y_i \right) \\
 &= \sum_i \left(\frac{1}{n} + (x - \bar{x})c_i \right)^2 V(y_i) \\
 &= \sigma^2 \sum_i \left(\frac{1}{n} + (x - \bar{x})c_i \right)^2 \\
 &= \sigma^2 \left(\sum_i \frac{1}{n^2} + (x - \bar{x})^2 \sum_i c_i^2 \right) \\
 &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)
 \end{aligned} \tag{44}$$

Nótese que en el caso $x = 0$, se encuentra que

$$E(\hat{\eta}_{x=0}) = \alpha \quad \text{et} \quad V(\hat{\eta}_{x=0}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = V(\hat{\alpha}) \quad (45)$$

Por lo que sigue será útil calcular también $E((\hat{\alpha} - \alpha)(\hat{\beta} - \beta)) = \text{cov}(\hat{\alpha}, \hat{\beta})$. Resulta

$$(\hat{\beta} - \beta) = \sum_i c_i (y_i - \eta_i) \quad \text{et} \quad (\hat{\alpha} - \alpha) = \sum_j \left(\frac{1}{n} - \bar{x} c_j \right) (y_j - \eta_j) \quad (46)$$

y por tanto

$$\begin{aligned} (\hat{\alpha} - \alpha)(\hat{\beta} - \beta) &= \sum_j \left(\frac{1}{n} - \bar{x} c_j \right) (y_j - \eta_j) \sum_i c_i (y_i - \eta_i) \\ &= \sum_i \left(\frac{1}{n} - \bar{x} c_i \right) c_i (y_i - \eta_i)^2 + \sum_i \sum_{i \neq j} \left(\frac{1}{n} - \bar{x} c_i \right) c_j (y_i - \eta_i)(y_j - \eta_j) \end{aligned} \quad (47)$$

Tomando las esperanzas, el primer cuadrado vale $(E(e_i))^2 + V(e_i) = \sigma^2$, mientras el ultimo termino vale cero, considerando la independencia de los y_i ; por tanto

$$\begin{aligned} \text{cov}(\hat{\alpha}, \hat{\beta}) &= E((\hat{\alpha} - \alpha)(\hat{\beta} - \beta)) = E\left(\sum_i \left(\frac{1}{n} - \bar{x} c_i \right) c_i (y_i - \eta_i)^2 \right) \\ &= \sigma^2 \sum_i \left(\frac{c_i}{n} - \bar{x} c_i^2 \right) + 0 = -\frac{\sigma^2 \bar{x}}{S_{xx}} \end{aligned} \quad (48)$$

Se puede preguntar si tiene sentido hacer una estimación de los mínimos cuadrados con esta covarianza entre estimadores. La respuesta es afirmativa, y se justifica con la propiedad de que estos estimadores tienen varianza mínima entre la clase de los estimadores lineales en y_i . Se trata del resultado siguiente.

3.2 El teorema de Gauss - Markov

Teorema. Dadas n pares de observaciones (x_i, y_i) , $i = 1, 2, \dots, n$, cuyos x_i son valores elegidos previamente y los y_i son medidas correspondientes e independientes para los cuales $E(y_i | x_i) = \alpha + \beta y_i$, $V(y_i | x_i) = \sigma^2$ por cada i ; sea $(\hat{\alpha}, \hat{\beta})$ la estimación de mínimos cuadrados de (α, β) , dada para el sistema (15). Entre todas las estimaciones lineales en los y_i para $\tau = \alpha_1 \alpha + \alpha_2 \beta$, la estimación de los mínimos cuadrados $\hat{\tau} = \alpha_1 \hat{\alpha} + \alpha_2 \hat{\beta}$ es la estimación de varianza mínima.

Prueba. Es evidente que $E(\hat{t}) = \tau$. Supongamos existente otra estimación de τ , *insesgada*, o sea sin margen de error, y lineal en los y_i , o sea $t = \sum d_i y_i$. Para la condición de sin margen de error resulta $E(t|x_1, x_2, \dots, x_n) = E(t) = \tau$, y por tanto

$$\begin{aligned} E(t) &= \sum_i d_i E(y_i | x_i) = \sum_i d_i (\alpha + \beta x_i) = \tau \\ &= \sum_i d_i \alpha + \sum_i d_i x_i \beta = a_1 \alpha + a_2 \beta \end{aligned}$$

por cada α, β . Entonces $\sum_i d_i = a_1$, $\sum_i d_i x_i = a_2$ y \hat{t} se pueden escribir como combinación lineal de los y_i , o sea

$$\begin{aligned} \hat{t} &= a_1 \alpha + a_2 \beta = \sum a_1 \left(\frac{1}{n} - \bar{x} c_i \right) y_i + \sum a_2 c_i y_i \\ &= \sum \left(\frac{a_1}{n} + (a_2 - a_1 \bar{x}) c_i \right) y_i \end{aligned} \tag{50}$$

Como los y_i son independientes, desarrollando resulta

$$\begin{aligned} V(\hat{t}) &= \sum \left(\frac{a_1}{n} + (a_2 - a_1 \bar{x}) c_i \right)^2 V(y_i) \\ &= \sigma^2 \left(\frac{a_1^2}{n} + \frac{(a_2 - a_1 \bar{x})^2}{S_{xx}} \right) \end{aligned} \tag{51}$$

Ahora se toman las desviaciones de las esperanzas, tanto de \hat{t} como de t :

$$\begin{aligned} \hat{t} - E(\hat{t}) &= \hat{t} - \tau = \sum \left(\frac{a_1}{n} + (a_2 - a_1 \bar{x}) c_i \right) (y_i - \eta_i) \\ t - E(t) &= t - \tau = \sum d_i (y_i - \eta_i) \end{aligned}$$

y, considerando que $\sum c_i d_i = \frac{\sum (x_i - \bar{x}) d_i}{S_{xx}} = \frac{\sum x_i d_i - \bar{x} \sum d_i}{S_{xx}} = \frac{a_2 - a_1 \bar{x}}{S_{xx}}$, se calcula su covarianza, teniendo en cuenta la independencia de los y_i como en (48). De esto resulta que

$$\begin{aligned}
\text{cov}(\hat{t}, t) &= E(\hat{t} - \tau)(t - \tau) = \sigma^2 \sum \left(\frac{a_1}{n} + (a_2 - a_1 \bar{x}) c_i \right) d_i \\
&= \sigma^2 \left(\frac{a_1^2}{n} + (a_2 - a_1 \bar{x}) \sum c_i d_i \right) \\
&= \sigma^2 \left(\frac{a_1^2}{n} + \frac{a_2 - a_1 \bar{x}}{S_{xx}} \right) = V(\hat{t})
\end{aligned} \tag{53}$$

El resultado es sorprendente. Luego se encuentra

$$\begin{aligned}
0 \leq V(t - \hat{t}) &= V(t) + V(\hat{t}) - 2\text{cov}(t, \hat{t}) \\
&= V(t) - V(\hat{t})
\end{aligned} \tag{54}$$

y por tanto $V(t) \geq V(\hat{t})$. ■

Corolario 1. ($a_1 = 0$, $a_2 = 1$). El estimador de los mínimos cuadrados de β es, entre todos los estimadores lineales en y_i , el de varianza mínima.

Corolario 2. ($a_1 = 1$, $a_2 = 0$). El estimador de los mínimos cuadrados de α es, entre todos los estimadores lineales en y_i , el de varianza mínima.

Corolario 3. ($a_1 = 1$, $a_2 = x$). El estimador de los mínimos cuadrados de η_x es, entre todos los estimadores lineales en y_i , el de varianza mínima.

3.3 Estimación de σ^2

Si el modelo es correcto, o sea si $E(y|x) = \eta_x$ es lineal en x , entonces se espera que los residuos $e_i = y_i - \hat{\eta}_{x_i}$ nos informen solo sobre los errores, o sea sobre σ^2 . Así calculamos la esperanza de los residuos e_i , o sea $E(SS_e)$. De este modo, resulta

$$\begin{aligned}
SS_e &= \sum_i (y_i - \hat{\eta}_{x_i})^2 = \sum_i (y_i - \alpha - \beta x_i)(y_i - \alpha - \beta x_i) \\
&= \sum_i (y_i - \alpha - \beta x_i)y_i - \alpha \sum_i (y_i - \alpha - \beta x_i) - \beta \sum_i (y_i - \alpha - \beta x_i)x_i
\end{aligned} \tag{55}$$

donde los dos últimos términos son cero porque son las derivadas parciales de SS_e , calculadas para (α, β) . Por tanto resulta que:

$$SS_e = S_{yy} - \alpha S_y - \beta S_{xy} = SS_t - SS_r \tag{56}$$

Esta ecuación dice que se puede calcular la suma de los cuadrados de los residuos empezando de las observaciones, de las estimaciones de los parámetros y de los términos de derecha de las ecuaciones normales. La suma SS_y de los cuadrados de los y_i se llama *suma de cuadrados total*, mientras que la suma SS_r de las estimaciones se llama *suma de cuadrados de la regresión*.

Como SS_r es combinación lineal de los segundos términos de las ecuaciones normales (18), substituyendolas con los primeros términos se puede escribir

$$SS_r = \alpha S_y + \beta S_{xy} = n\alpha^2 + 2S_x \alpha \beta + S_{xx} \beta^2 = \sum_i (\alpha + \beta x_i)^2 = \sum_i \hat{\eta}_{x_i}^2 = S_{\hat{\eta}\hat{\eta}} \quad (57)$$

lo que justifica el nombre de suma de cuadrados de la regresión. Es interesante observar que SS_r no es otro que la forma cuadrática que tiene como matriz la de los coeficientes en el sistema de ecuaciones normales:

$$(\alpha \ \beta) \begin{pmatrix} n & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad (58)$$

Calculamos ahora la esperanza de SS_r : sobre la base (57)

$$\begin{aligned} E(SS_r) &= n \left(\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_2}{S_{xx}} \right) + \alpha^2 \right) + 2n\bar{x} \left(-\frac{\sigma^2 \bar{x}}{S_{xx}} + \alpha \beta \right) + S_{xx} \left(\frac{\sigma^2}{S_{xx}} + \beta^2 \right) \\ &= 2\sigma^2 + (n\alpha^2 + 2S_x \alpha \beta + S_{xx} \beta^2) \end{aligned} \quad (59)$$

Observese que la presencia en $E(SS_r)$ de los dos parámetros α y β justifica el factor dos para $2\sigma^2$. De esto resultado, la esperanza $E(SS_e)$ se calcula fácilmente:

$$\begin{aligned} E(SS_e) &= \sum_i E(y_i^2) - SS_r \\ &= \sum_i (\sigma^2 + (\alpha + \beta x_i)^2) - E(SS_r) \\ &= n\sigma^2 + (n\alpha^2 + 2S_x \alpha \beta + S_{xx} \beta^2) - (2\sigma^2 + (n\alpha^2 + 2S_x \alpha \beta + S_{xx} \beta^2)) \\ &= (n-2)\sigma^2. \end{aligned} \quad (60)$$

El factor $n - 2$ se indica como los *grados de libertad* del error. Ahora, si se pone

$$MS_e = \frac{SS_e}{n-2} = \frac{\sum_i (y_i - \hat{\eta}_i)^2}{n-2}, \text{ es evidente que su esperanza vale}$$

$$E(MS_e) = \sigma^2 \quad (61)$$

lo que significa que MS_e es un estimador de σ^2 insesgado.

3.4 Análisis de varianza del modelo

Resulta que la suma de cuadrados total se comparte en dos,

$$SS_t = SS_r + SS_e \quad (62)$$

una de las cuales informa sobre los parámetros de la función de regresión y la otra sobre los errores. Los elementos de esta partición se representan en una *tabla de análisis de varianza* como:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Regresión	2	SS_r	$SS_r / 2$	$\sigma^2 + (n\alpha^2 + 2S_x\alpha\beta + S_{xx}\beta^2) / 2$
Error	$n - 2$	SS_e	$SS_e / (n - 2)$	σ^2
Total	n	SS_t		

Se pueden comprender los grados de libertad como sigue: antes de ver los datos, los n y_i son libres en el espacio \mathbb{R}^n . La solución de los mínimos cuadrados a las ecuaciones normales son dos restricciones para calcular α y β , que se encuentra en un sub-espacio de \mathbb{R}^n de dimensión 2. Por esto, las cantidades $e_i = y_i - \eta_i$ se encuentra en el sub-espacio complemento de dimensión $n - 2$, solución de las ecuaciones normales (18).

La tabla de análisis de varianza nos muestra que las esperanzas de los cuadrados medios de la regresión y de los errores coinciden si $\alpha = \beta = 0$. Por tanto se pueden utilizar para saber como contestar la pregunta «¿serían $\alpha = \beta = 0$?».

3.5 El modelo lineal centrado en \bar{x}

En el análisis que se presentó en los párrafos precedentes, los residuos del modelo lineal empleado pueden ser empleados para saber si el modelo tiene algunos parámetros diferente de cero o no. Al contrario, a menudo el interés es concentrado sobre solo un parámetro, o sea β , y en este caso se vuelve a una versión del modelo lineal modificado por

respecto a los sistemas (8), (9), (125) que consiste en el *centrar* los x_i alrededor de su promedio. Una vez introducida la variable centrada $w_i = x_i - \bar{x}$, con $\sum_i w_i = 0$, se puede reescribir el modelo como sigue:

$$\left\{ \begin{array}{l} y_i = \alpha + \beta \bar{x} + (\beta x_i - \beta \bar{x}) + \varepsilon_i = \varphi + \beta w_i + \varepsilon_i \\ E(\varepsilon_i) = 0 \\ V(\varepsilon_i) = \sigma^2 \\ \varepsilon_i \text{ y } \varepsilon_j \text{ independientes por cada } i \neq j \end{array} \right. \quad (63)$$

3.6 La solución des los mínimos cuadrados

Se busca la solución de (63) con el método de los mínimos cuadrados minizando $SS_e(\varphi, \beta)$. Derivando parcialmente, se consigue

$$\left\{ \begin{array}{l} \frac{\partial SS_e(\varphi, \beta)}{\partial \varphi} = -2 \sum_i (y_i - \varphi - \beta w_i) = 0 \\ \frac{\partial SS_e(\varphi, \beta)}{\partial \beta} = -2 \sum_i (y_i - \varphi - \beta w_i) w_i = 0 \end{array} \right. \quad (64)$$

luego, teniendo en cuenta el centrado de los w_i deja las ecuaciones normales en φ, β , se simplifican a

$$\left\{ \begin{array}{l} n \hat{\varphi} = S_y \\ S_{ww} \hat{\beta} = S_{wy} \end{array} \right. \quad (65)$$

cuya solución inmediata es

$$\left\{ \begin{array}{l} \hat{\varphi} = \frac{S_y}{n} = \bar{y} \\ \hat{\beta} = \frac{S_{wy}}{S_{ww}} \end{array} \right. \quad (66)$$

Claro que empezando de los estimadores que resultan de (66) se ve que esta solución

coincide con la del modelo precedente:

$$\begin{cases} \alpha = \bar{y} - \beta \bar{x} \\ \beta = \frac{S_{xy}}{S_{xx}} \end{cases} \quad (67)$$

3.7 Las estadísticas de los estimadores

Se emplean las técnicas ya empleadas en este capítulo para definir las esperanzas y las varianzas de los estimadores $\hat{\phi}$, $\hat{\beta}$ y $\hat{\eta}_x$. Aquí también se expresa $\hat{\beta}$ como combinación

lineal de los y_i , $\hat{\beta} = \sum_i c_i y_i$, donde $c_i = \frac{w_i}{S_{ww}} = \frac{x_i - \bar{x}}{S_{xx}}$, y por tanto tiene las propiedades (36). Así resulta que

$$\begin{aligned} \hat{\eta}_{wi} &= \hat{\phi} + \hat{\beta} w_i = \sum_i \left(\frac{y_i}{n} \right) + \sum_i c_i y_i w_i \\ &= \sum_i \left(\frac{1}{n} + w_i c_i \right) y_i \\ &= \sum_i \left(\frac{1}{n} + (x_i - \bar{x}) c_i \right) y_i = \hat{\eta}_{xi} \end{aligned} \quad (68)$$

y por tanto que $E(\hat{\beta}) = \beta$ y $V(\hat{\beta}) = \frac{\sigma^2}{S_{ww}} = \frac{\sigma^2}{S_{xx}}$. Luego, la esperanza de $\hat{\phi}$ vale:

$$E(\hat{\phi}) = E(\bar{y} | x_i) = \frac{\sum_i (\phi + \beta w_i)}{n} = \frac{\sum_i \phi}{n} + \frac{\beta}{n} \sum_i w_i = \phi \quad (69)$$

y que su varianza vale $V(\hat{\phi}) = V(\bar{y}) = \frac{\sigma^2}{n}$. Finalmente, se tiene

$$E(\hat{\eta}_w) = E(\hat{\phi} + \hat{\beta} x) = E(\hat{\phi}) + E(\hat{\beta}) w = \phi + \beta w = \eta_w \quad (70)$$

y su varianza es

$$\begin{aligned}
 V(\hat{\eta}_w) &= V(\hat{\phi} + \beta w) = V\left(\sum_i \left(\frac{1}{n} + w c_i\right) y_i\right) \\
 &= \sum_i \left(\frac{1}{n} + w c_i\right)^2 V(y_i) = \sigma^2 \left(\frac{1}{n} + \frac{w^2}{S_{ww}}\right) \\
 &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)
 \end{aligned} \tag{71}$$

Si ahora se calcula la covarianza entre $\hat{\phi}$ y β , resulta

$$(\beta - \beta) = \sum_i c_i (y_i - \eta_i) \quad \text{y} \quad (\hat{\phi} - \phi) = \frac{\sum_j (y_j - \eta_j)}{n} \tag{72}$$

Siguiendo (48) se tiene que

$$\begin{aligned}
 (\hat{\phi} - \phi)(\beta - \beta) &= \frac{1}{n} \sum (y_j - \eta_j) \sum_i c_i (y_i - \eta_i) \\
 &= \frac{1}{n} \sum_i c_i (y_i - \eta_i)^2 + \frac{1}{n} \sum_i \sum_{i \neq j} c_j (y_i - \eta_i) (y_j - \eta_j)
 \end{aligned} \tag{73}$$

Tomando las esperanzas, el primero cuadrado vale $(E(e_i))^2 + V(e_i) = \sigma^2$ y el último termino vale cero, por el tema de la independendencia de los y_i ; por tanto resulta

$$\begin{aligned}
 cov(\hat{\phi}, \beta) &= E((\hat{\phi} - \phi)(\beta - \beta)) \\
 &= \frac{1}{n} E\left(\sum_i c_i (y_i - \eta_i)^2\right) = \frac{1}{n} \sum_i c_i E((y_i - \eta_i)^2) \\
 &= \frac{\sigma^2}{n} \sum_i c_i = 0
 \end{aligned} \tag{75}$$

Entonces, al contrario de α y β , los estimadores $\hat{\phi}$ y β no son correlacionados.

3.8 Análisis de los residuos

Ahora calculamos la esperanza de los residuos e_i , o sea $E(SS_e)$. Si se desarrolla $SS_e = \sum_i (y_i - \hat{\eta}_{x_i})^2$ se encuentra

$$SS_e = SS_t - SS_r = S_{yy} - \hat{\phi}S_y - \beta S_{wy} \quad (76)$$

donde la suma de los cuadrados de la regresión resulta compartida en dos, una parte dependiendo solo de $\hat{\phi}$ y la otra solo de β . Acordamos aquí que estas dos partes pueden escribirse de manera diferente, o sea $\hat{\phi}S_y = n\bar{\bar{y}}^2 = n\bar{y}^2$ y $\beta S_{wy} = \beta S_{xy} = \beta^2 S_{xx}$. Se pueden calcular también las esperanzas y resulta que respectivamente

$$\begin{aligned} E(\hat{\phi}S_y) &= E(n\bar{y}^2) = n(E\bar{y}^2 + V(\bar{y})) = n\bar{y}^2 + n\frac{\sigma^2}{n} = \hat{\phi}S_y + \sigma^2 \\ E(\beta S_{wy}) &= E(\beta^2 S_{ww}) = S_{ww}(E(\beta)^2 + V(\beta)) \\ &= S_{ww}\left(\beta^2 + \frac{\sigma^2}{S_{ww}}\right) = \beta^2 S_{ww} + \sigma^2 = \beta S_{wy} + \sigma^2 \end{aligned} \quad (77)$$

Es fácil ahora mostrar que la suma de las esperanzas es igual a la esperanza de SS_r del modelo en α y β . Precisamente, considerando que $S_{xx} = S_{\hat{x}\hat{x}} + \frac{S_x^2}{n}$, se consigue

$$\begin{aligned} E(SS_r) &= E(\hat{\phi}S_y + \beta S_{wy}) = 2\sigma^2 + \hat{\phi}S_y + \beta S_{wy} \\ &= 2\sigma^2 + n\left(\alpha + \beta\frac{S_x}{n}\right)^2 + \beta^2 S_{\hat{x}\hat{x}} \\ &= 2\sigma^2 + n\alpha^2 + 2S_x\alpha\beta + \frac{S_x^2}{n}\beta^2 + \beta^2 S_{\hat{x}\hat{x}} \\ &= 2\sigma^2 + n\alpha^2 + 2S_x\alpha\beta + S_{xx}\beta^2 \end{aligned} \quad (78)$$

3.9 Análisis de varianza

Al final, se pueden organizar estos resultados en una tabla de análisis de varianza donde se separan los efectos de $\hat{\varphi}$ y $\hat{\beta}$:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
$\hat{\varphi}$	1	$\hat{\varphi}S_y$		$\varphi S_y + \sigma^2$
$\hat{\beta}$	1	$\hat{\beta}S_{wy}$		$\beta S_{wy} + \sigma^2$
Error	$n - 2$	SS_e	$SS_e / (n - 2)$	σ^2
Total	n	SS_t		

Entonces, si la afirmación « $\beta = 0$ » es verdadera, resulta que $E(\hat{\beta}S_{wy}) = E(MS_e) = \sigma^2$. A menudo, cuando no interesa φ , conviene poner la tabla de forma diferente, o sea centrando los y_i también alrededor de su promedio, y resulta

$$\begin{aligned}
 SS_e &= SS_t - SS_r = S_{yy} - \hat{\varphi}S_y - \hat{\beta}S_{wy} = \\
 &= S_{yy} - nS_y^2 - \hat{\beta}S_{wy} = S_{\ddot{y}\ddot{y}} - \hat{\beta}S_{wy}
 \end{aligned}
 \tag{79}$$

y empleando la suma de los cuadrados centrados $SS_{\ddot{t}} = S_{\ddot{y}\ddot{y}}$ y $SS_r(\beta) = \hat{\beta}S_{wy}$ se consigue finalmente

$$SS_{\ddot{t}} = SS_r(\beta) + SS_e
 \tag{80}$$

En la tabla siguiente, se sintetizan estos resultados.

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
$\hat{\beta}$	1	$SS_r(\beta) = \hat{\beta}S_{wy}$		$\beta S_{wy} + \sigma^2$
Error	$n-2$	SS_e	$SS_e / (n - 2)$	σ^2
Total	$n-1$	$SS_{\ddot{t}}$		

3.10 Comparación de modelos

Se puede definir $SS_r(\beta) = \beta' S_{wy}$ como la *suma de los cuadrados de la regresión adjunta debido a β* , porque, escribiendo $SS_e = SS_e(\alpha, \beta)$ se encuentra

$$SS_r(\beta) = SS_i - SS_e(\alpha, \beta) \quad (81)$$

Se puede pensar en la ecuación (81) así: suponemos que los y_i son independientes, tales que $E(y_i) = \eta_i = \alpha$, y que $V(y_i) = \sigma^2$. En consecuencia del teorema de Huygens se encuentra la solución de mínimos cuadrados $\hat{\alpha} = \bar{y}$ se donde resulta

$$SS_e(\alpha) = SS_i - \sum_i (y_i - \bar{y})^2 = \min_{\alpha} \sum_i (y_i - \alpha)^2 \quad (82)$$

Además se tiene que $E(\hat{\alpha}) = E(\bar{y}) = \alpha$ y $V(\hat{\alpha}) = E(\sum_i (y_i - \bar{y})^2) = (n-1)\sigma^2$ donde se puede escribir

$$SS_r(\beta) = SS_e(\alpha) - SS_e(\alpha, \beta) \quad (83)$$

lo cual muestra que la suma de los cuadrados debido a β es en efecto la parte de suma de cuadrados de la regresión que resulta incluyendo β en el modelo. Es evidente que, si $\beta \neq 0$, con su inclusión en el modelo, la suma de los cuadrados del modelo se reduce, porque resulta que

$$SS_e(\alpha, \beta) = SS_e(\alpha) - SS_r(\beta) \quad (84)$$

3.11 La calidad del modelo

Como claramente $SS_i > SS_e$, se puede escribir

$$\frac{SS_e}{SS_i} = 1 - \frac{SS_r(\beta)}{SS_i} = 1 - r^2 \quad (85)$$

r^2 se llama el *coeficiente de determinación*: $0 \leq r^2 \leq 1$. Dado que $SS_e = (1 - r^2)SS_i$, si r^2 vale cero se tiene que $SS_e = SS_i$. En esto caso no tiene sentido de buscar β , porque el error que resulta es el mismo que resultaría quedándose con la recta de regresión $\eta_x = \varphi = \bar{y}$.

Si al contrario $r^2 \rightarrow 1$, SS_e se reduce progresivamente y el modelo siempre es más eficaz; si por otro lado $r^2 = 1$, entonces $SS_e = 0$ y por tanto todas las observaciones se encuentran *sobre la recta de regresión*: en este caso la relación entre x y y es funcional.

r^2 se puede escribir en diferentes formas, por ejemplo como

$$r^2 = \frac{SS_r(\hat{\beta})}{SS_i} = \frac{\hat{\beta}S_{wy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} \quad (86)$$

donde su raíz cuadrada

$$r = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (87)$$

r se llama el *coeficiente de correlación* y varía entre -1 y 1. Considerando el signo del numerador en el último miembro de (87), r es negativo cuando los x_i y los y_i tienen una variación opuesta.

Vale observar que, conociendo previamente r , resulta

$$\begin{cases} \alpha = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = r \sqrt{\frac{S_{yy}}{S_{xx}}} \end{cases} \quad (88)$$

entonces, del punto de vista del cálculo, α y $\hat{\beta}$ solo dependen de los promedios de las variables y de su coeficiente de correlación. Debido a la simetría de r se puede calcular al mismo tiempo la regresión de y sobre x y la de x sobre y . Claro que la existencia de dichas rectas y la calidad del modelo no son suficientes para afirmar la existencia de relaciones causales entre las variables, ni tampoco el sentido de dichas relaciones.

4. Inferencia del modelo lineal

4.1 Hipótesis de normalidad

Hasta ahora solo se hizo matemática, o sea se hicieron cálculos, siempre posibles, que nos dieron estimaciones de los parámetros del modelo así como los residuos. Pero no sabemos como utilizar las estadísticas que resultan de la estimación de mínimos cuadrados del modelo (2).

Para poder utilizarlas, se necesita una hipótesis adicional, o sea se necesita modificar la condición (6) precisando la distribución D . Se impondrá que la distribución de y dado x sea *normal e independiente*, solo así podremos efectuar los test admitidos bajo esta hipótesis. También se puede imponer que los y sean non correlacionados, por que si dos variables aleatorias tienen una distribución conjunta normal bivariada y no son correlacionadas, por consecuencia son independientes. Esta condición se sintetiza escribiendo

$$\text{para } x, y = \text{DNI}(\alpha + \beta x, \sigma^2) \quad (89)$$

donde *DNI* indica *distribución normal independiente*. Bajo estas condiciones se tiene el siguiente teorema:

Teorema. Dado un conjunto de variables aleatorias $y_i, i = 1, \dots, n$ de distribución normal e independientes (89) para x_i dado, la distribución conjunta del estimador de mínimos cuadrados $(\hat{\alpha}, \hat{\beta})$ de (α, β) es la distribución normal bivariada, cuya función de densidad viene dada por:

$$f(\alpha, \beta) = \frac{\sqrt{nS_{xx}}}{2\pi\sigma^2} e^{-\frac{T(\alpha, \beta)}{2\sigma^2}} \quad (90)$$

dond

$$T(\alpha, \beta) = n(\hat{\alpha} - \alpha)^2 + 2S_x(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) + S_{xx}(\hat{\beta} - \beta)^2 \quad (91)$$

Esto implica que

$$\begin{aligned}
 E(\hat{\alpha}) &= \alpha, & V(\hat{\alpha}) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \\
 E(\hat{\beta}) &= \beta, & V(\hat{\beta}) &= \frac{\sigma^2}{S_{xx}}
 \end{aligned}
 \tag{92}$$

y que $(\hat{\alpha}, \hat{\beta})$ tienen una distribución independiente de SS_e , cuya distribución es

$$SS_e = \sigma^2 \chi^2_{n-2} \tag{93}$$

Igualmente se es independiente de SS_e la distribución de $MS_e = SS_e / (n-2)$.

Corolario. La variable aleatoria $T(\hat{\alpha}, \hat{\beta})$ definida en (91) tiene una distribución $T = \sigma^2 \chi^2_2$ y es independiente de SS_e y MS_e . Por tanto

$$F = \frac{T(\hat{\alpha}, \hat{\beta})/2}{SS_e/(n-2)} = F_{2, n-2} \tag{94}$$

tiene una distribución de Fisher-Snedecor con 2 y $n-2$ grados de libertad.

Definamos ahora

$$\hat{\tau} = a_1 \hat{\alpha} + a_2 \hat{\beta} \tag{95}$$

Corolario. La variable aleatoria $\hat{\tau}$ tiene una distribución normal

$$\hat{\tau} = N \left(\tau, \sigma^2 \left(\frac{a_1^2}{n} + \frac{(a_2 - a_1 \bar{x})^2}{S_{xx}} \right) \right) \tag{96}$$

independiente de SS_e y MS_e . Par tanto, introduciendo el *desvío estándar* de $\hat{\tau}$ (en inglés *standard deviation, SD*),

$$SD(\hat{\tau}) = \sqrt{V(\hat{\tau})} = \sqrt{MS_e} \sqrt{\frac{a_1^2}{n} + \frac{(a_2 - a_1 \bar{x})^2}{S_{xx}}} \tag{97}$$

y considerando que MS_e es un estimador de σ^2 , se encuentra que

$$\frac{\hat{\tau} - \tau}{SD(\hat{\tau})} = t_{n-2} \quad (98)$$

sigue una ley t de student con $n - 2$ grados de libertad. Fijado un nivel de probabilidad π se puede no aceptar la hipótesis

$$H_0 : \tau = \tau_0$$

si resulta que

$$\frac{\hat{\tau} - \tau_0}{SD(\hat{\tau})} > t_{n-2; \pi/2} \quad (99)$$

y definir el intervalo de confianza de τ a nivel de $1 - \pi$ como

$$\left\{ \tau \mid \hat{\tau} - SD(\hat{\tau})t_{n-2; \pi/2} \leq \tau \leq \hat{\tau} + SD(\hat{\tau})t_{n-2; \pi/2} \right\} \quad (100)$$

Esto teorema y su dos corolarios nos permiten de utilizar las estadísticas que se construyeron en el capítulo precedente para testar la calidad de los resultados.

En base al teorema (93) se puede hacer un test sobre SS_e y construir un intervalo de confianza para σ^2 . Así resulta

$$\left\{ \sigma^2 \mid \frac{SS_e}{\chi^2_{n-2; \pi/2}} \leq \sigma^2 \leq \frac{SS_e}{\chi^2_{n-2; 1-\pi/2}} \right\} \quad (101)$$

El primero corolario se emplea para testar la hipótesis

$$H_0 : \alpha = \alpha_0 \text{ y } \beta = \beta_0$$

que implica que la distribución muestrada al punto $x = x_i$ tiene como promedio $\alpha_0 + \beta_0 x_i$. Si H_0 es verdadera, si se puede emplear el test

$$F = \frac{T(\alpha_0, \beta_0)/2}{MS_e} \quad (102)$$

por que en este caso resulta $F = F_{2, n-2}$. Por tanto no se acepta H_0 a nivel de significatividad π si $F \geq F_{2, n-2, \pi}$. En particular, para testar la hipótesis

$$H_0 : \alpha = 0 \text{ y } \beta = 0$$

el test

$$F = \frac{T(0,0)/2}{MS_e} = \frac{MS_r}{MS_e} \quad (103)$$

se basa en los valores que se encuentran en la tabla de análisis de varianza.

Como

$$\frac{T(\alpha_0, \beta_0)/2}{MS_e} = F_{2, n-2}$$

la *región de confianza* para (α, β) es

$$\{(\alpha, \beta) \mid T(\alpha, \beta) \leq 2MS_e F_{2, n-2; \pi}\} \quad (104)$$

que es una elipse centrada en $(\hat{\alpha}, \hat{\beta})$.

El segundo corolario puede ser utilizado para testar las hipótesis sobre cualquier combinación lineal de los parámetros y en particular sobre ellos mismos.

Así supongamos que deseamos testar β : en este caso se pone $a_1 = 0$ y $a_2 = 1$ en (95) y

se obtiene

$$SD(\hat{\beta}) = \sqrt{\frac{MS_e}{S_{\hat{x}\hat{x}}}}$$

donde

$$\frac{\hat{\beta} - \beta_0}{SD(\hat{\beta})} > t_{n-2; \pi/2}$$

y

$$\{\beta \mid \hat{\beta} - SD(\hat{\beta})t_{n-2; \pi/2} \leq \beta \leq \hat{\beta} + SD(\hat{\beta})t_{n-2; \pi/2}\}$$

Para testar α se pone $a_1 = 1$ y $a_2 = 0$ en (95) donde resulta

$$SD(\hat{\alpha}) = \sqrt{MS_e \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{\hat{x}\hat{x}}} \right)}$$

y se consigue

$$\frac{\hat{\alpha} - \alpha_0}{SD(\hat{\alpha})} > t_{n-2; \pi/2}$$

y
$$\{\alpha \mid \hat{\alpha} - SD(\hat{\alpha})t_{n-2;\pi/2} \leq \alpha \leq \hat{\alpha} + SD(\hat{\alpha})t_{n-2;\pi/2}\}$$

Ahora, para testar η_x se pone $a_1 = 1$ y $a_2 = x$ en (95) y por tanto el desvío estándar de $\hat{\eta}$ vale

$$DS(\hat{\eta}) = \sqrt{MS_e \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{\hat{x}\hat{x}}} \right)},$$

su estimación vale
$$\frac{\hat{\eta} - \eta_0}{SD(\hat{\eta})} > t_{n-2;\pi/2}$$

y el intervalo de confianza es

$$\{\eta \mid \hat{\eta} - SD(\hat{\eta})t_{n-2;\pi/2} \leq \eta \leq \hat{\eta} + SD(\hat{\eta})t_{n-2;\pi/2}\}$$

Se puede observar que el ancho del intervalo de confianza crece si x se deja del promedio.

Vale observar que podría ser que ambos α y β se encuentra cada uno en su intervalo de confianza pero que el par no se encuentra en el elipse de la región de confianza. En este caso tiene que pedirse cual es el objetivo del estudio: si se va buscando β , entonces solo interesa esto y por tanto su intervalo de confianza; si al contrario es de interés α y β , será necesario limitarse dentro de la región de confianza (104).

4.2 La falta de ajuste del modelo lineal

En todo lo que precede, se hace la hipótesis que la relación entre x y y era conocida como lineal o era buena una aproximación lineal. No obstante en algunas situaciones es importante de comprobar que la relación entre x es y y es lineal y se muestra aquí un método para averiguar lo que se llama la *falta de ajuste* del modelo lineal.

La idea se basa en el hecho que por cada x_i , la recta de regresión pasa por el punto $\eta_{x_i} = E(y_i \mid x_i)$ que corresponde al promedio de los y que se encuentran en correspondencia del valor x_i , mientras si la relación que se trata como lineal no es tal, ninguna recta va pasar para todos los promedios. En consecuencia MS_e , el estimador de la varianza disponible σ^2 , que depende del modelo empleado, mediando los desvíos con respecto a puntos diferentes del promedio, va sobreestimar la varianza.

Entonces se trata de estimar la varianza de los y_i de una otra manera y comparar los dos. Para una medida de la varianza de los y_i se necesitan por lo menos uno de los x_i correspondientes a por lo menos dos medidas y_{i1} y y_{i2} , aunque para una buena estimación claro que sería preferible conocer diversos valores y_{ij} por cada x_i .

Supongamos por tanto que hemos elegido $m > 3$ valores $x_i, i = 1, 2, \dots, m$ y por cada uno haber medido n_i valores $y_{ij}, j = 1, 2, \dots, n_i$ con al menos un $n_i > 1$. Los estimadores de mínimos cuadrados se pueden calcular como siempre, mientras las fórmulas se pueden escribir de manera diferente.

De hecho, si $\sum_{i=1}^m n_i = n$ se tiene

$$\bar{x} = \frac{\sum_{i=1}^m n_i x_i}{n} \quad \text{y} \quad \bar{y} = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij}}{n} = \frac{\sum_{i=1}^m n_i \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}}{n} = \frac{\sum_{i=1}^m n_i \bar{y}_i}{n}$$

y la solución de mínimos cuadrados resulta

$$\begin{cases} \alpha = \bar{y} - \beta \bar{x} \\ \beta = \frac{s_{\hat{x}\hat{y}}}{s_{\hat{x}\hat{x}}} = \frac{\sum_{i=1}^m (x_i - \bar{x}) \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) + \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}) \right)}{s_{\hat{x}\hat{x}}} = \frac{\sum_{i=1}^m n_i (x_i - \bar{x}) (\bar{y}_i - \bar{y})}{s_{\hat{x}\hat{x}}} \end{cases} \quad (105)$$

por que $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$ en cuanto a desvíos del promedio y $s_{\hat{x}\hat{x}} = \sum_{i=1}^m n_i (x_i - \bar{x})^2$. Por tanto la suma de los cuadrados de los residuos vale

$$SS_e = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{\eta}_i)^2 \quad (106)$$

Intuitivamente se comprende que si la regresión no es lineal, los residuos deberían contener de la información sobre esto, ya que los y_{ij} informen sobre los valores *verdaderos* mientras $\hat{\eta}$ solo informa sobre la linealidad. Por esto SS_e tiene que informar también sobre el desvío a la linealidad de la función verdadera y será más grande de σ^2 . Como se repitieron algunas medidas por los mismos x_i se esta en condición de medir σ^2 y por tanto de comprobar su diferencia con SS_e . Para esto se hace una *análisis de varianza a una via* sobre los datos, compuestos en m grupos de n_i observaciones bajo la asunción que las esperanzas de los y_i en cada grupo resultan de la recta de regresión

$$E(y_{ij} | x_i) = \alpha + \beta x_i, \quad i = 1, 2, \dots, m$$

si bien se duda que estas sean diferentes. En dicho análisis de varianza el *término de error* es la suma de los cuadrados *intra* SS_W .

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Inter	$m - 1$	$SS_B = \sum_{i=1}^m n_i (\bar{y}_i - \bar{y})^2$	$SS_B / (m - 1)$	
Intra	$n - m = \sum (n_j - 1)$	$SS_W = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$SS_W / (n - m)$	σ^2
Total	$n - 1 = \sum n_j - 1$	$SS_T = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$		

Se observa que en el análisis de varianza la relación $SS_T = SS_B + SS_W$ resulta

$$\begin{aligned} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^m \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}))^2 = \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \\ &+ 2 \sum_{i=1}^m (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m n_i (\bar{y}_i - \bar{y})^2 \end{aligned}$$

donde la última suma de la tercera línea vale cero pues es una suma de desvíos al promedio.

Es simple comprender que SS_W , suma de cuadrados *intra* solo informa sobre σ^2 , por que es formada de sumas de desvíos al promedio en cada grupo. Como en cada grupo los y_i tienen una distribución independiente e idéntica, con promedio $\eta_{x_i} = E(y_i | x_i)$ y varianza

$V(y_{ij}) = \sigma^2$, resulta¹⁰

$$E(SS_W) = E\left(\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2\right) = \sigma^2 \sum_{i=1}^m (n_i - 1) = \sigma^2 (n - m) \quad (108)$$

Por tanto $MS_W = SS_W / (n - m)$ es un estimador insesgado de σ^2 , que no depende del modelo.

Ahora, como ya se hizo en la regresión lineal, se puede compartir SS_B , con $m - 1$ grados de libertad, en dos partes, una $SS_r(\beta)$ que resulta de la regresión, y una otra, que se va indicar con SS_M con $m - 2$ grados de libertad, que vale

$$SS_M = SS_B - SS_r(\beta) \quad (109)$$

Se puede mostrar que la esperanza de SS_M , que vale

$$E(SS_M) = (m - 2)\sigma^2$$

si la hipótesis de linealidad es correcta, en el caso de falta de ajuste lineal vale

$$E(SS_M) = (m - 2)\sigma^2 + \Lambda^2$$

o sea $\Lambda^2 = 0$ si la hipótesis de linealidad es correcta. Por tanto se puede reconstruir la tabla de análisis de varianza como sigue:

¹⁰ En general, si n variables aleatorias x_i son independientes, con una misma esperanza ξ y una misma varianza σ^2 por cada i , se tiene $\sum E(x_i^2) = \sum ((E(x_i))^2 + V(x_i)) = n(\xi^2 + \sigma^2)$ et $E(x_i x_j) = E(x_i)E(x_j)$, y por tanto

$$\begin{aligned} E\left(\sum (x_i - \bar{x})^2\right) &= \sum E(x_i^2) - \frac{1}{n} E\left(\sum x_i\right)^2 = \\ &= n(\xi^2 + \sigma^2) - \frac{1}{n} \left(\sum E(x_i)^2 + \sum_{i \neq j} E(x_i)E(x_j) \right) = \\ &= n(\xi^2 + \sigma^2) - \frac{1}{n} (n(\xi^2 + \sigma^2) + (n^2 - n)\xi^2) = \\ &= (n - 1)\sigma^2 \end{aligned}$$

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Inter Regresión	1	$SS_r(\beta)$		$\sigma^2 + \beta^2 S_{xx} + g(\Lambda)^\alpha$
Inter falta	$m - 2$	$SS_M = SS_B - SS_r(\beta)$	$MS_M = SS_M / (m - 2)$	$\sigma^2 + \Lambda^2 / (m - 2)$
Intra	$n - m$	$SS_W = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$MS_W = SS_W / (n - m)$	σ^2
Total	$n - 1$	$SS_T = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$		

Para testar el ajuste del modelo, bajo la hipótesis de normalidad de la distribución de los y_{ij} nos basamos en el hecho que si el modelo es lineal la proporción entre MS_M e MS_W tiene una distribución F con $m-2$ y $n-m$ grados de libertad. Por tanto se puede no aceptar la hipótesis de linealidad a nivel de significatividad π si

$$F_M = \frac{MS_M}{MS_W} > F_{m-2, n-m; \pi}$$

y aceptarla en caso contrario.

Se puede leer la tabla de análisis de varianza de la regresión considerando compartir SS_e en lugar de SS_B , ya que resulta

$$\begin{aligned} SS_T &= SS_r(\beta) + SS_e \\ SS_e &= SS_M + SS_W \end{aligned} \tag{113}$$

Si la hipótesis de linealidad no es aceptada, *cualquier* relación no lineal puede ser comprobada.

La falta de ajuste puede ser testada con dicho procedimiento cuando se tienen observaciones repetidas. En los otros casos, es necesario observar la distribución de los residuos sobre gráficos de dispersión en respecto a x e $\hat{\eta}$: si se distribuyen regularmente en una cinta alrededor de la recta horizontal $e = 0$, se pueden aceptar las hipótesis hechas, en particular la homoscedasticidad y la linealidad.

4.3 La predicción

A menudo el objetivo de un estudio experimental es el de recolectar datos para pronosticar un valor y de la variable criterio cuando la variable explicativa x tiene un valor particular de interés. Claro que se supone que los parámetros del modelo ya fueron estimados y las hipótesis necesarias ya fueron testadas, incluso el ajuste.

Por tanto se piensa en un modelo lineal

$$\square (y \mid x) = \eta_x = \alpha + \beta x \quad (114)$$

cuyos parámetros ya fueron estimados. Así resulta que

$$\hat{\eta}_x = \hat{\alpha} + \hat{\beta} x \quad (115)$$

Ahora se quiere pronosticar y , o sea se quiere estimar un nuevo valor de y cuando $x = x_0$, donde y será medido independientemente de los datos utilizados en la regresión. Para esto se va buscando un predictor adecuado \tilde{y}_{x_0} de una respuesta y y observada en correspondencia del valor $x = x_0$, sobre la base de la información contenida en la muestra $\{(x_i, y_i) \mid i = 1, 2, \dots, n\}$. Parece razonable utilizar como predictor

$$\tilde{y}_{x_0} = \hat{\alpha} + \hat{\beta} x_0 \quad (116)$$

en tanto que su esperanza, porque y es independiente de la muestra, es

$$E(\tilde{y}_{x_0} \mid (x_i, y_i), i=1, 2, \dots, n) = E(y \mid x_0) = \eta_{x_0} = \alpha + \beta x_0 \quad (117)$$

El pensamiento que conduce a la elección de \tilde{y}_{x_0} como predictor se basa en el hecho que se busca un desvío entre valor pronosticado y observado mínimo. En fórmulas

$$\begin{aligned} E((\tilde{y}_{x_0} - y)^2 \mid (x_i, y_i), i=1, 2, \dots, n) &= \\ E((\tilde{y}_{x_0} - \eta_{x_0})^2 \mid (x_i, y_i), i=1, 2, \dots, n) + E((y - \eta_{x_0})^2 \mid x_0) &= \\ = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) & \end{aligned} \quad (118)$$

dado que el producto cruzado es cero por el tema de la independencia de y y de \tilde{y}_{x_0} . Por tanto, del teorema de Gauss el primero término es mínimo entre todos los predictores lineal

en los y_i , y vale como $V(\eta_{x_0})$, y el segundo claramente vale σ^2 .

El desvío estándar del predictor es $SD(\tilde{y}_{x_0}) = \sqrt{MS_e \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$, luego bajo

la hipótesis que dado $x = x_0$ la variable y es distribuida normalmente, se puede testar la hipótesis nula $H_0: \tilde{y}_{x_0} = y_0$ a través de

$$\frac{\tilde{y}_{x_0} - y_0}{SD(\tilde{y}_{x_0})} > t_{n-2; \pi/2}$$

y el intervalo de confianza es

$$\left\{ y \mid \tilde{y}_{x_0} - SD(\tilde{y}_{x_0}) t_{n-2; \pi/2} \leq y \leq \tilde{y}_{x_0} + SD(\tilde{y}_{x_0}) t_{n-2; \pi/2} \right\}$$

Se observa que el intervalo de confianza del predictor es más ancho del del estimador, por el hecho que en el predictor hay que considerar también la varianza de los y_i .

5. La regresión lineal múltiple

5.1 Introducción

Se estudió hasta ahora el caso de la regresión lineal simple, o sea del modelo lineal

$$y = \alpha + \beta x = \varepsilon, \quad \forall x \quad (119)$$

por el cual se intenta de explicar la variación de un carácter respuesta y para un solo regresor x . Seguidamente se quiere estudiar el modelo que expresa la relación de dependencia entre un carácter respuesta y y algunos regresores, de manera que se puede suponer una relación del tipo

$$\eta = f(z_1, z_2, \dots, z_s; \theta_1, \theta_2, \dots, \theta_t) \quad (120)$$

donde los z_1, z_2, \dots, z_s son los regresores y los $\theta_1, \theta_2, \dots, \theta_t$ son los parámetros del modelo. En lo que sigue la mayoría de los objetos homogéneos tratados serán sintetizados por vectores (denotados en negrita) o matrices. Por tanto la relación que tratamos se escribe por

$$\eta = f(\mathbf{z}; \boldsymbol{\theta}) \quad (121)$$

Con esta representación, los regresores se vuelven componentes del vector \mathbf{z} y los parámetros componentes del vector $\boldsymbol{\theta}$. Así, el *modelo de regresión lineal múltiple* es un modelo en el cual los parámetros aparecen linealmente en la ecuación (121), que se vuelve

$$\eta = f(\mathbf{z}; \boldsymbol{\theta}) = \sum_{j=1}^k \beta_j x_j(\mathbf{z}) = \boldsymbol{\beta}' \mathbf{x}(\mathbf{z}) \quad (122)$$

donde las componentes x_j del vector \mathbf{x} solo son función de las componentes z_h del vector \mathbf{z} . Esto significa que, empezando de los regresores \mathbf{z} verdaderos, hay libertad de transformarlos a través de cualquier función antes de incluirlos en el modelo a condición que ningún parámetro a estimar entre en las transformaciones. Las componentes β_j del vector $\boldsymbol{\beta}$ son los *parámetros* que aparecen en la ecuación (122), en principio desconocidos. Claro que los vectores $\boldsymbol{\beta}$ y \mathbf{x} tienen la misma dimensión k mientras la dimensión de \mathbf{z} puede ser cualquiera.

Ejemplo 1: $\eta = \tau_0 + \tau_1 z_1 + \tau_2 z_2 + \dots + \tau_d z_d$ es un modelo lineal, tomándose $\beta_j = \tau_{j-1}$, $x_j = x_j(\mathbf{z}) = z_{j-1}$, $j = 1, \dots, d+1$, $s = 1$, $k = d+1$. Se trata del tipo más común de modelo lineal, donde los β_j entran linealmente. El término β_1 es el término *constante*, correspondiente al parámetro α del modelo lineal simple.

Ejemplo 2: una relación polinomial entre η y \mathbf{z} , $\eta = \alpha + \tau_1 z^1 + \tau_2 z^2 + \dots + \tau_d z^d$ es un modelo lineal, tomándose $\beta_j = \tau_{j-1}$, $x_j = x_j(\mathbf{z}) = z^{j-1}$, $j = 1, \dots, d+1$, $s = 1$, $k = d+1$. Los β_j entran linealmente en el modelo.

Ejemplo 3: una relación polinomial entre η y z_1 e z_2 , $\eta = \tau_{00} + \tau_{10} z_1^1 + \tau_{01} z_2^1 + \tau_{20} z_1^2 + \tau_{11} z_1 z_2 + \tau_{02} z_2^2 + \dots$ es un modelo lineal, tomándose $\beta_1 = \tau_{00}$, $\beta_2 = \tau_{10}$, ..., $x_1 = 1$, $x_2(z) = z_1$, ...etc. Los β_j entran linealmente en el modelo.

Se puede tener interés de estudiar una relación lineal entre un carácter y algunos otros, porque: 1) a menudo se sabe que existen influencias lineales entre caracteres que se quiere explorar; 2) aunque la función f no es lineal, resulta que hay pequeñas regiones en el dominio de rango de \mathbf{x} donde la relación puede ser aproximada para una relación lineal; 3) no se conoce la forma de f , pero se empieza aproximandola con funciones polinomiales, por supuesto lineales en los parámetros.

Como en la regresión simple, aquí se estudiarán tres aspectos.

1) Estimación de los parámetros

Bajo la condición que el modelo (121) es aceptable para el estudio de las relaciones entre caracteres que interesan, se trata de *estimar los parámetros* β del modelo. Esto se hace con una experimentación: se empieza de n conjuntos de k valores $(x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ de los k regresores (eventualmente resultando de transformaciones de los valores de los $z_{i1}, z_{i2}, \dots, z_{is}$ definidos) y se observan valores y_i , $i = 1, 2, \dots, n$ en correspondencia de cada conjunto j . Claro que cada valor y_i es afectado de un error experimental, de manera que tiene que escribirse, para cada i

$$y_i = \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i \quad (123)$$

donde resulta que el error se encuentra como desvío del valor observado al modelo, o sea, para cada i

$$\varepsilon_i = y_i - \eta_i \quad (124)$$

Se hace la hipótesis que los errores son errores experimentales aleatorios independientes y que por cada \mathbf{x}_i se tiene la misma distribución con media 0 y varianza σ^2 ; así que se puede escribir

$$\left\{ \begin{array}{l} y_i = \eta_i + \varepsilon_i = \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i \\ E(y_i | \mathbf{x}_i) = \eta_i \\ V(y_i | \mathbf{x}_i) = \sigma^2 \\ y_i \text{ y } y_j \text{ independientes para cada } i \neq j \end{array} \right. \quad (125)$$

o, igualmente,

$$\left\{ \begin{array}{l} y_i = \eta_i + \varepsilon_i \\ E(\varepsilon_i) = 0 \\ V(\varepsilon_i) = \sigma^2 \\ \varepsilon_i \text{ y } \varepsilon_j \text{ independientes para cada } i \neq j \end{array} \right. \quad (126)$$

Existen varias razones para admitir la presencia de un error experimental, pero se indica en particular que esto puede depender de la influencia sobre el carácter y de otros factores que no están incluidos en los regresores x_j . Esto significa que el modelo (122) es un abreviado del modelo

$$y_i = \sum_{j=1}^k \beta_j x_{ij} + \sum_{l=1}^h \beta_l x_{il} \quad (127)$$

en el cual están incluidos otros regresores. Como los x_{il} van cambiando de manera desconocida en los experimentos pero influyendo sobre los y_i , si h es bastante grande, el teorema del límite central asegura la normalidad de los errores.

2) Calidad del modelo

Empezando con la estimación de los parámetros, el interés en el estudio de un modelo particular se encuentra en la evaluación de su calidad, o sea comprobar si el modelo elegido es adecuado para describir la relación entre regresores y respuesta.

3) Predicción

Un tercero aspecto interesante del estudio consiste en la posibilidad de poder pronosticar, sobre la base de los valores y_i observados en un rango de los x_{ij} , las posibles respuestas a otros valores x_{0j} que se encuentra en el mismo rango.

En lo que sigue se examinarán estas etapas.

5.2 Estimación de los parámetros

Se quiere estimar los parámetros del modelo lineal (8), (9), (125). Así, se hicieron n observaciones, con $n \gg k$, de valores del carácter respuesta y_i correspondiendo a valores prefijos de los regresores $x_{i1}, x_{i2}, \dots, x_{ik}$, $i = 1, 2, \dots, n$. Entonces, cada valor y_i se puede considerar, en el espacio \mathbb{R}^n , como una componente del vector respuesta \mathbf{y} , con n componentes; mientras que los valores x_{ij} tienen que considerarse por un lado como componentes de k vectores \mathbf{x}_j de dimensión n , cada uno compuesto con todos los valores

considerados en las n observaciones por el j -ésimo factor, y por otro lado como las componentes de los n vectores \mathbf{x}'_i , de dimensión k , cada uno correspondiente al conjunto de valores de los regresores en la i -ésima observación. Por tanto se trata de una matriz X con n filas y k columnas. Similarmente los términos de error ε_i se pueden representar con un vector $\boldsymbol{\varepsilon}$ con n componentes, de modo que el modelo se puede escribir

$$\begin{cases} \mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ E(\mathbf{y}|X) = \boldsymbol{\eta} \quad \text{ou} \quad E(\boldsymbol{\varepsilon}) = \mathbf{0} \\ V(\mathbf{y}) = V(\boldsymbol{\varepsilon}) = \sigma^2 I \end{cases} \quad (128)$$

Aquí, el vector $\boldsymbol{\eta} = X\boldsymbol{\beta}$ tiene n dimensiones y la independencia de los errores resulta, en el caso de la normalidad, de la tercera ecuación en el modelo (128).

Observación: En la mayoría de los modelos es importante guardar un parámetro correspondiente a un término constante. A este parámetro, normalmente β_1 , se asocia en X el primero vector columna $\mathbf{x}_1 = (1, 1, \dots, 1)'$, así que en lugar de k regresores en realidad se tienen $k-1$.

Ahora, las k columnas de la matriz X como vector del espacio \mathbb{R}^n generan un sub-espacio $S \subset \mathbb{R}^n$, cuya dimensión es a lo más $k < n$. Este espacio toma el nombre de *espacio solución, de regresión, de los parámetros, o de estimación*. Su dimensión es k solo si los regresores son *linealmente independientes* y en este caso se dice que el sistema es *de rango completo*. Resulta que el vector $\boldsymbol{\eta} = X\boldsymbol{\beta}$ es un vector del sub-espacio S , es decir que se quiere *estimar el vector \mathbf{y} en \mathbb{R}^n para un vector $\boldsymbol{\eta}$ de S* . Por esto se busca, en estas condiciones, *cual es la mejor estimación posible*, o sea cual es el mejor estimador de \mathbf{y} para un vector de S , como combinación lineal de vectores-columnas de X .

Desde el punto de vista geométrico, como el espacio \mathbb{R}^n es Euclidiano, la solución más natural es estimar \mathbf{y} por su *proyección ortogonal* $\boldsymbol{\eta} = X\boldsymbol{\beta}$ sobre S .

Se sabe bien geoméricamente que la solución dada para la proyección siempre existe y es única. Esta se consigue empleando un *operador ortogonal de proyección* sobre S , denotado por ρ_S . La condición de ortogonalidad se expresa diciendo que, dado cualquier vector \mathbf{y} , el vector $\mathbf{y} - \rho_S \mathbf{y}$, es ortogonal a cada vector de S . En consecuencia cada vector de \mathbb{R}^n se encuentra compartido en dos componentes $\mathbf{y} = \rho_S \mathbf{y} + (\mathbf{y} - \rho_S \mathbf{y})$, así que el espacio también resulta compartido como *suma directa* de sub-espacios ortogonales. Si se escribe $\mathbf{y} - \rho_S \mathbf{y} = (I - \rho_S)\mathbf{y} = \rho_{S^\perp} \mathbf{y} = \mathcal{E}_S \mathbf{y}$ resulta por tanto $\mathbf{y} = \rho_S \mathbf{y} \oplus (\mathbf{y} - \rho_S \mathbf{y}) = \rho_S \mathbf{y} \oplus \mathcal{E}_S \mathbf{y}$. En lo que sigue se va omitir el índice del proyector. Es evidente que cada proyector ortogonal es *idempotente*, o sea $\rho \circ \rho = \rho$ y por tanto $\rho \circ \mathcal{E} = \rho \circ (I - \rho) = \rho - \rho \circ \rho = \mathbf{0}$.

Teorema. Un proyector ortogonal es representado para una matriz A *idempotente* y

simétrica.

Prueba: En efecto, de la relación $\varrho \circ (I - \varrho) = 0$, si $A\mathbf{y} = \varrho\mathbf{y}$ se consigue, para cada \mathbf{x}, \mathbf{y} en \mathbb{R}^n

$$\mathbf{0} = (A\mathbf{y})'(\mathbf{x} - A\mathbf{x}) = \mathbf{y}'A'\mathbf{x} - \mathbf{y}'A'A\mathbf{x}$$

Por tanto $A' = A'A$. Transponiendo ambos miembros resulta $A = A'A$, donde $A = A'$ y la idempotencia como consecuencia.

En el caso, $\mathbb{R}^n = S \oplus S^\perp$, donde S^\perp es el *complemento ortogonal de S* en \mathbb{R}^n , la proyección ortogonal de \mathbf{y} en S es $\varrho_S\mathbf{y} = \hat{\boldsymbol{\eta}} = X\hat{\boldsymbol{\beta}}$, donde resulta $\mathbf{y} = \hat{\boldsymbol{\eta}} + \mathbf{e}$ con $\mathbf{y} - \hat{\boldsymbol{\eta}} = (I - \varrho_S)\mathbf{y} = \mathbf{e} \in S^\perp$. Entonces S^\perp se llama el *espacio de los errores* o de los *residuos*.

Teorema. El punto $\hat{\boldsymbol{\eta}}$ es el punto dentro de S más cercano a \mathbf{y} .

Prueba: Sea $\mathbf{s} = X\boldsymbol{\theta} \in S$ un punto cualquiera. Entonces el triángulo $\mathbf{y}\hat{\boldsymbol{\eta}}\mathbf{s}$ es rectangular e \mathbf{ys} es su hipotenusa, entonces su longitud es mas grande que la del cateto $\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\eta}}$.

De otro lado, se puede utilizar el teorema de Pitágoras y medir el cuadrado de la longitud del segmento \mathbf{ys} como suma de cuadrados de los catetos. Por esto resulta que el mínimo

$$\begin{aligned} SS_e &= \min (\mathbf{y} - \mathbf{s})'(\mathbf{y} - \mathbf{s}) = \min ((\mathbf{y} - \hat{\boldsymbol{\eta}})'(\mathbf{y} - \hat{\boldsymbol{\eta}}) + (\mathbf{s} - \hat{\boldsymbol{\eta}})'(\mathbf{s} - \hat{\boldsymbol{\eta}})) = \\ &= \min ((\mathbf{y} - X\hat{\boldsymbol{\beta}})'(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (X(\boldsymbol{\theta} - \hat{\boldsymbol{\beta}}))'(X(\boldsymbol{\theta} - \hat{\boldsymbol{\beta}}))) \end{aligned}$$

se consigue cuando el segundo término es cero, o sea cuando $\boldsymbol{\theta} = \hat{\boldsymbol{\beta}}$.

Se puede calcular $\hat{\boldsymbol{\beta}}$ considerando que $\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\eta}} \in S^\perp$, porque así \mathbf{e} es ortogonal a cada vector de S , donde $X'\mathbf{e} = X'(\mathbf{y} - \hat{\boldsymbol{\eta}}) = (X'\mathbf{y} - X'X\hat{\boldsymbol{\beta}}) = \mathbf{0}$. Así resulta

$$X'X\hat{\boldsymbol{\beta}} = X'\mathbf{y} \tag{131}$$

que constituye el sistema de *ecuaciones normales* cuya solución numérica dejaría la estimación $\hat{\boldsymbol{\beta}}$ a través del *método de mínimos cuadrados*.

Las ecuaciones normales pueden ser conseguidas directamente buscando el mínimo

$$SS_e = \min_{\theta} (\mathbf{y} - \mathbf{X}\theta)'(\mathbf{y} - \mathbf{X}\theta) = \min_{\theta} (\mathbf{y}'\mathbf{y} + 2\theta'\mathbf{X}'\mathbf{y} + \theta'\mathbf{X}'\mathbf{X}\theta) \quad (132)$$

a través del cálculo de las derivadas parciales

$$\frac{\partial SS_e}{\partial \theta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\theta = \mathbf{0} \quad (133)$$

donde se consigue el mismo sistema (131) pero donde se indica con $\hat{\beta}$ su solución.

Para conseguir la solución (131) tiene que distinguirse el caso de *rango completo*, en el cual la dimension de S es k , igual al número de regresores, del caso de rango no lleno, o sea cuando $\dim S < k$. Este caso se produce si algún regresor es linealmente dependiente de otros. Si se puede borrarlo, se vuelva al caso de rango completo, sino sera necesario de resolver el problema de otra manera. En nuestro estudio, nos limitaremos al caso de rango completo, pero precisando que: 1) *la solución siempre existe y siempre es única en cualquier situación*, porque siempre se trata de una proyección ortogonal; 2) además su expresión algébrica en el caso de rango no completo puede no ser única, y por esto hay que utilizar técnicas numéricas específicas; 3) normalmente, los programas no hacen distinción entre las dos situaciones porque fueron realizados de manera de tratar sin dificultad ambas condiciones.

5.3 La solución en el caso de rango completo

En el caso de rango completo, como $\dim S = k$, la matriz $\mathbf{X}'\mathbf{X}$ es invertible y por tanto la solución es

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

donde

$$\hat{\eta} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathcal{O}\mathbf{y}$$

Es evidente que la matriz $\mathcal{O} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es simétrica e idempotente, y como $\hat{\eta}$ se encuentra en S , $\hat{\eta}$ es la proyección ortogonal de \mathbf{y} sobre S . En consecuencia, el vector $\mathbf{e} = \mathbf{y} - \hat{\eta} = (\mathbf{I} - \mathcal{O})\mathbf{y} = \mathcal{E}\mathbf{y}$ es la proyección ortogonal de \mathbf{y} sobre el espacio de residuos S^\perp que entonces tiene dimensión $n - k$.

Se consiguen los resultados siguientes:

$\hat{\beta}$ es el estimador de mínimos cuadrados de β , que minimiza la distancia entre \mathbf{y} y su estimador en el espacio de los regresores S ;

$\hat{\boldsymbol{\eta}}$ *Proyección ortogonal de \mathbf{y} sobre S* es el vector de los *valores ajustados* para la regresión: como proyección ortogonal de \mathbf{y} sobre S , es el punto de S más cercano de \mathbf{y} ;

\mathbf{e} *proyección ortogonal de \mathbf{y} sobre S^\perp* es el vector de los *residuos*, ortogonal a $\hat{\boldsymbol{\eta}}$;

$SS_r = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}}$ es la norma del vector $\hat{\boldsymbol{\eta}}$, que vale

$$SS_r = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\boldsymbol{\rho}\mathbf{y}$$

y que puede verse también como el producto escalar $\mathbf{y}'\hat{\boldsymbol{\eta}}$;

$SS_e = \mathbf{e}'\mathbf{e}$ es la norma del vector \mathbf{e} , y resulta

$$SS_e = \mathbf{e}'\mathbf{e} = \mathbf{y}'\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}\mathbf{y} = \mathbf{y}'\boldsymbol{\mathcal{E}}\mathbf{y}$$

Este puede ver también como el producto escalar $\mathbf{y}'\mathbf{e}$.

Como $\boldsymbol{\rho}$ y $\boldsymbol{\mathcal{E}}$ son simétricos e idempotentes, sus trazas son iguales a sus rangos (pues, al ser idempotentes, sus autovalores solo pueden valer un o cero). Entonces resulta que

$$r(\boldsymbol{\rho}) = \text{tr}(\boldsymbol{\rho}) = \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}) = \text{tr}(\mathbf{X}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = \text{tr}(\mathbf{I}_k) = k$$

y

$$r(\boldsymbol{\mathcal{E}}) = \text{tr}(\boldsymbol{\mathcal{E}}) = \text{tr}(\mathbf{I}_n - \boldsymbol{\rho}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\boldsymbol{\rho}) = n - k$$

5.4 Propiedades estadísticas de los estimadores de mínimos cuadrados¹¹

Sabiendo que $E(\mathbf{y}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, resulta todavía que

$$E(\hat{\boldsymbol{\beta}}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$E(\hat{\boldsymbol{\eta}}) = E(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}E(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}$$

$$E(\mathbf{e}) = E(\mathbf{y} - \hat{\boldsymbol{\eta}}) = E(\mathbf{y}) - E(\hat{\boldsymbol{\eta}}) = \boldsymbol{\eta} - \boldsymbol{\eta} = \mathbf{0}$$

y, considerando que $V(\mathbf{y}) = \sigma^2 \mathbf{I}$ resulta

¹¹ En lo que sigue, tiene que acordarse que si \mathbf{x} es un vector y A una matriz constante, resulta:

- 1) $V(A\mathbf{x}) = \text{cov}(\sum_k a_{ik}x_k, \sum_h a_{jh}x_h) = \sum_k \sum_h a_{ik} a_{jh} \text{cov}(x_k, x_h) = AV(\mathbf{x})A'$
- 2) $E(\mathbf{x}'A\mathbf{x}) = E(\sum_j \sum_i x_i a_{ij} x_j) = \sum_j \sum_i E(x_i a_{ij} x_j) = \sum_j \sum_i a_{ij} E(x_i x_j) = \sum_j \sum_i a_{ij} (E(x_i)E(x_j) + v_{ij}) = \sum_j \sum_i a_{ij} E(x_i)E(x_j) + \sum_j \sum_i a_{ij} v_{ij} = E(\mathbf{x})'A E(\mathbf{x}) + \text{tr}(AV)$, desde que V es simétrica.

$$\begin{aligned}
V(\hat{\beta}) &= V((X'X)^{-1}X'y) = (X'X)^{-1}X'V(y)X(X'X)^{-1} = \sigma^2(X'X)^{-1} \\
V(\hat{\eta}) &= V(X\hat{\beta}) = XV(\hat{\beta})X' = \sigma^2X(X'X)^{-1}X' = \sigma^2\varrho \\
V(e) &= V(y - X\hat{\beta}) = V(y - X(X'X)^{-1}X'y) = \sigma^2(I_n - \varrho) = \sigma^2\mathcal{E}
\end{aligned}$$

Hay que observar que los estimadores de los parámetros tienen una covarianza entre ellos, dependiendo de las relaciones lineales entre regresores.

Teorema (Gauss). Dados n conjuntos de observaciones, dispuestas en forma de matriz (X, y) , cuyos X son valores previamente elegidos y los y medidas correspondientes e independientes para las cuales $E(y|X) = X\beta$, $V(y|X) = \sigma^2I$; sea $\hat{\beta}$ la estimación de mínimos cuadrados de β . Si $\tau = a'\beta$, donde $a' = (a_1, a_2, \dots, a_n)$ es un vector de constantes, entonces entre todos los estimadores insesgados y lineal en y de τ , la estimación de mínimos cuadrados $\hat{\tau} = a'\hat{\beta}$ es la de varianza mínima.

Prueba.

Es evidente que $E(\hat{\tau}) = \tau$ y que $\hat{\tau}$ es lineal en y pues $\hat{\tau} = a'\hat{\beta} = a'(X'X)^{-1}X'y = c'y$. Entonces su varianza es $V(\hat{\tau}) = V(a'\hat{\beta}) = \sigma^2 a'(X'X)^{-1}a$. Supongamos exista otra estimación de τ , insesgada y lineal en y , digamos $t = d'y$, con $d \neq c$. Por la condición de insesgamiento resulta $E(t|X) = \tau$ y por tanto

$$E(t) = d'E(y|X) = d'X\beta = \tau = a'\beta$$

para cada β . Por tanto $d'X = a'$.

La covarianza entre t y $\hat{\tau}$ vale

$$\begin{aligned}
cov(\hat{\tau}, t) &= cov(c'y, d'y) = E(d'(y - \eta)(y - \eta)'c) = \sigma^2 d'c = \\
&= \sigma^2 d'X(X'X)^{-1}a = \sigma^2 a'(X'X)^{-1}a = V(\hat{\tau})
\end{aligned}$$

Se encuentra entonces que

$$\begin{aligned}
0 \leq V(t - \hat{\tau}) &= V(t) + V(\hat{\tau}) - 2cov(t, \hat{\tau}) \\
&= V(t) - V(\hat{\tau})
\end{aligned}$$

y por tanto $V(t) \geq V(\hat{\tau})$. ■

Consecuencias del teorema de Gauss son los siguientes:

Corolario 1. Si $a_j = 1$, $a_{h \neq j} = 0$, $h = 1 \dots k$, Para cada $j = 1, \dots, k$, el estimador de mínimos cuadrados de β_j , j -ésima componente de β , es, entre todos los estimadores insesgados y lineales en y, β_j , es el de varianza mínima.

Corolario 2. Si $a_i = x_i$, es para cada $i = 1, \dots, n$, el estimador de mínimos cuadrados de η_i , entre todos los estimadores insesgados y lineales en y, η_i , es el de varianza mínima.

5.5 El análisis de varianza del modelo

Una vez estimados los β resulta que el cuadrado de la distancia promedio entre y y S es

$$\begin{aligned} e'e &= (y - \hat{\eta})'(y - \hat{\eta}) = (y - X\hat{\beta})'(y - X\hat{\beta}) = \\ &= y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} = \\ &= y'y - \hat{\beta}'X'X\hat{\beta} = y'y - \hat{\eta}'\hat{\eta} \end{aligned}$$

pues, según las ecuaciones normales, $y'X\hat{\beta} = (\hat{\beta}'X'y)' = \hat{\beta}'X'X\hat{\beta}$, de donde

$$y'y = \hat{\eta}'\hat{\eta} + e'e = y'\rho y + y'\mathcal{E}y$$

Esto se puede escribir también

$$SS_t = SS_r + SS_e$$

donde se ha particionado la suma de cuadrados de las observaciones en dos, una parte SS_r , debido a la regresión de y sobre X y la otra, SS_e , debido al error. Por tanto, por cuanto $\hat{\eta}$ contiene la información sobre el modelo $\eta = X\beta$, e solo contiene la información contenida sobre el error y entonces $e'e$ debe informar sobre σ^2 .

Calculemos ahora SS_r y SS_e

$$E(SS_r) = E(y'\rho y) = \beta'X'\rho X\beta + tr(\rho \sigma^2 I) = \beta'X'X\beta + k\sigma^2$$

$$E(SS_e) = E(y'\mathcal{E}y) = \beta'X'\mathcal{E}X\beta + tr(\mathcal{E} \sigma^2 I) = (n - k)\sigma^2$$

resultando el producto $\mathcal{E}X = 0$ y por tanto los cuadrados promedios

$$MS_r = SS_r/k \quad E(MS_r) = \beta'X'X\beta/k + \sigma^2$$

$$MS_e = SS_e/(n-k) \quad E(MS_e) = \sigma^2$$

o sea MS_e es un estimador insesgado de σ^2 .

Los elementos de esta partición se representan en una *tabla de análisis de varianza* como sigue:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Regresión	k	SS_r	SS_r/k	$\sigma^2 + \beta'X'X\beta/k$
Error	$n - k$	SS_e	$SS_e/(n - k)$	σ^2
Total	n	SS_t		

Se sabe que los grados de libertad son las dimensiones de los espacios en los cuales se encuentran los vectores: efectivamente, el espacio de los estimadores S tiene k dimensiones, mientras el espacio de los residuos tiene dimensión $n - k$.

La tabla de análisis de varianza revela que las esperanzas de los cuadrados medios de la regresión y del error coinciden si $\beta = 0$. Esto se puede utilizar para testar la hipótesis nula $H_0: \beta = 0$.

Si se supone que el valor de los parámetros sea medianamente $\beta = \beta_0$, por ejemplo como resultado de otro experimento, se trata de transformar el problema en uno en el cual se tiene que examinar la nulidad de algunos parámetros. Exactamente, resulta (teorema de Pitágoras) que el cuadrado de la distancia entre y y $\eta_0 = X\beta_0 \in S$ es:

$$(y - X\beta_0)'(y - X\beta_0) = (y - X\beta)'(y - X\beta) + (\beta - \beta_0)'X'X(\beta - \beta_0)$$

Si entonces se hace una translación del origen de 0 a $\eta_0 = X\beta_0$ se obtiene la nueva variable criterio

$$z = y - X\beta_0 \tag{151}$$

a la cual se quiere aplicar el modelo

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\phi} + \boldsymbol{\varepsilon} \quad (152)$$

donde se escribe $\boldsymbol{\phi} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$. Formalmente, el sistema (152) es idéntico al (128), pero es fácil de mostrar que la traslación de la solución del sistema (128), o sea $\boldsymbol{\eta}_0 = \boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}\hat{\boldsymbol{\phi}}$ es solución de (152), mientras la hipótesis nula $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ es equivalente a la hipótesis $\boldsymbol{\phi} = \mathbf{0}$, y que el término residual sigue idéntico. Por esto hay la tabla de análisis de varianza correspondiente:

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>
Regresión	k	$SS_{r(z)} = \hat{\boldsymbol{\phi}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\phi}}$	$SS_{r(z)} / k$	$\sigma^2 + \hat{\boldsymbol{\phi}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\phi}} / k$
Error	$n - k$	$SS_{e(z)} = SS_e$	$SS_e / (n - k)$	σ^2
Total	n	$SS_{(z)}$		

6. Inferencia

6.1 Distribuciones normales multivariadas

En lo que sigue sera necesario imponer la hipótesis que la distribución de \mathbf{y} es multinormal con respecto a \mathbf{X} . Por tanto se analiza aquí los resultados más importantes relativos a las muestras de una distribución normal multivariada, que servirán a los test estadísticos. Para una lectura más detallada y referencias se puede ver Guttman (1982).

Definición. Se dice que el vector aleatorio $\mathbf{y} \in \mathbb{R}^n$, tiene una distribución *normal* si su función de densidad $p_{\mathbf{y}}(\mathbf{y})$ es

$$p_{\mathbf{y}}(\mathbf{y}) = \sqrt{\frac{|\Sigma^{-1}|}{(2\pi)^n}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

donde la matriz Σ^{-1} es simétrica definida positiva y $\boldsymbol{\mu}$ tiene componentes finitas. Entonces se escribe $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$, con $E(\mathbf{y}) = \boldsymbol{\mu}$ y $V(\mathbf{y}) = \Sigma$.

Teorema. Si $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$ y Σ es diagonal, su componentes y_i de \mathbf{y} son estadísticamente independientes.

Teorema. Sea un vector aleatorio $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$, con $\Sigma = P'P$ y Q y considere la forma cuadrática centrada

$$Q = (\mathbf{y}-\boldsymbol{\mu})'G(\mathbf{y}-\boldsymbol{\mu})$$

donde la matriz G es simétrica y real. Entonces la ley de distribución de Q es una combinación lineal de n variables aleatorias independientes de ley chi-cuadrado con 1 grado de libertad

$$Q = \sum_i \lambda_i \chi_1^2(i)$$

donde los λ_i son los autovalores de $P'GP$ (y también de ΣG y de $G\Sigma$).

Teorema. Una condición necesaria y suficiente por que Q tenga una ley de distribución de chi-cuadrado con $k < n$ grados de libertad es que $P'GP$ sea idempotente y de rango k . Si $\Sigma = \sigma^2 I$ la condición deviene en que G sea idempotente de rango k .

Teorema de Craig. Sea un vector aleatorio $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$ y las dos formas cuadráticas

$$Q_i = (\mathbf{y} - \boldsymbol{\mu})' G_i (\mathbf{y} - \boldsymbol{\mu}), \quad i = 1, 2$$

con G_i reales y simétricas. Entonces Q_1 y Q_2 son estadísticamente independientes si y solo si $G_1 \Sigma G_2 = 0$.

Teorema de Cochran. Sea $\mathbf{y} = N(0, I)$ una variable aleatoria y sean n observaciones de \mathbf{y} independientes, formando un vector aleatorio $\mathbf{y} = N(\mathbf{0}, I)$. Sea por otro lado

$$Q = \mathbf{y}' \mathbf{y} = Q_1 + Q_2 + \dots + Q_k$$

donde $Q_i = \mathbf{y}' A_i \mathbf{y}$ es una forma cuadrática de rango $rg(A_i) = n_i$, y A_i es una matriz simétrica $n \times n$, $i = 1, \dots, k$. Entonces las siguiente condiciones son equivalentes:

- 1) Q_1, Q_2, \dots, Q_n son estadísticamente independientes;
- 2) Q_1, Q_2, \dots, Q_n tienen individualmente distribuciones de chi-cuadrado;
- 3) $n_1 + n_2 + \dots + n_k = n$.

Teorema. Sea el vector aleatorio $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$ y considere su distribución de \mathbf{y} condicional $A(\mathbf{y} - \boldsymbol{\mu}) = 0$

donde A es una matriz $k \times n$, $k < n$, $rg(A) = k$. Entonces la distribución condicional de \mathbf{y} es tal que

$$Q_i = (\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \chi^2_{n-k}$$

6.2 Test del modelo y de los parámetros

Como se supone que el vector \mathbf{y} tiene en cada punto de medida una distribución normal centrada sobre su esperanza $\boldsymbol{\eta}$ y de varianza constante, $\mathbf{y} = N(\boldsymbol{\eta}, \sigma^2 I)$, bajo la hipótesis nula $H_0 : \boldsymbol{\beta} = \mathbf{0}$ se tiene $E(\mathbf{y}) = \mathbf{0}$, y por tanto $\mathbf{y} = N(\mathbf{0}, \sigma^2 I)$. Como resulta del teorema de Craig SS_r y SS_e son estadísticamente independientes, pues son formas cuadráticas de una distribución normal centrada y reducida, de matrices \mathcal{P} y \mathcal{E} tales que $\sigma^2 \mathcal{P} \mathcal{E} = 0$. En consecuencia resulta

$$SS_r = \sigma^2 \chi^2_k \quad SS_e = \sigma^2 \chi^2_{n-k}$$

con las leyes χ^2 independientes, donde

$$\text{bajo } H_0 : \boldsymbol{\beta} = \mathbf{0}, \quad F = \frac{MS_r}{MS_e} = F_{k, n-k}$$

Por tanto, fijado un nivel de probabilidad π el test resulta ser

$$\text{no aceptar } H_0 : \boldsymbol{\beta} = \mathbf{0}, \text{ si } \frac{MS_r}{MS_e} > F_{k,n-k;\pi}$$

aceptar en otro caso.

De otro lado, es fácil de ver que en

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad (162)$$

el primero término a la derecha vale $\mathbf{e}'\mathbf{e}$ y se puede escribir

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

Entonces resulta

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/k}{\mathbf{e}'\mathbf{e}/(n-k)} = \frac{\sigma^2 \chi^2_k/k}{\sigma^2 \chi^2_{n-k}/(n-k)} = F_{k,n-k} \quad (164)$$

donde se construye la región de confianza de $\boldsymbol{\beta}$ al nivel de $1 - \pi$

$$C_{1-\pi} = \left\{ \boldsymbol{\beta} \mid (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq k MS_e F_{k,n-k;\pi} \right\} \quad (165)$$

Como $\hat{\boldsymbol{\beta}}$ es una combinación lineal de las observaciones, es distribuido normalmente con promedio $\boldsymbol{\beta}$ y varianza $\sigma^2(X'X)^{-1}$, por tanto cada componente $\hat{\beta}_i$ tiene una distribución normal $\hat{\beta}_i = N(\beta_i, \sigma^2 c_{ii})$ donde c_{ii} es el elemento diagonal correspondiente a $(X'X)^{-1}$. Entonces se puede definir su *desvío estándar* (*SD*, standard deviation)

$$SD(\hat{\beta}_i) = \sqrt{MS_e c_{ii}}$$

y el test

$$\text{no aceptar } H_0 : \beta_i = \beta_i, \text{ si } \left| \frac{\hat{\beta}_i - \beta_i}{SD(\hat{\beta}_i)} \right| > t_{n-k;\pi/2}$$

aceptar en otro caso.

y el intervalo de confianza

$$C_{1-\pi} = \left\{ \beta_i \mid \hat{\beta}_i - SD(\hat{\beta}_i)t_{n-k;\pi/2} \leq \beta_i \leq \hat{\beta}_i + SD(\hat{\beta}_i)t_{n-k;\pi/2} \right\}$$

Hay que tener cuidado en la utilización de estos intervalos, porque en general son más grandes que los intervalos de confianza de cada región de confianza (47). Entonces tiene sentido hacer test individuales.

Para el tema de la varianza, su intervalo de confianza es dado por

$$C_{1-\pi} = \left\{ \sigma^2 \mid \frac{SS_e}{\chi^2_{n-k;\pi/2}} \leq \sigma^2 \leq \frac{SS_e}{\chi^2_{n-k;1-\pi/2}} \right\} \quad (168)$$

Como la varianza de $\hat{\eta}$ vale $V(\hat{\eta}) = V(X\beta) = XV(\beta)X' = \sigma^2 X(X'X)^{-1}X'$, por la estimación de $\hat{\eta}_0 = \mathbf{x}'_0\beta$, y el desvío estándar $SD(\hat{\eta}_0) = \sqrt{MS_e \mathbf{x}'_0(X'X)^{-1}\mathbf{x}_0}$ se deriva la

condición del test

$$\left| \frac{\hat{\eta} - \eta_0}{SD(\hat{\eta})} \right| > t_{n-k;\pi/2}$$

y su intervalo de confianza $\left\{ \eta \mid \hat{\eta} - SD(\hat{\eta})t_{n-k;\pi/2} \leq \eta \leq \hat{\eta} + SD(\hat{\eta})t_{n-k;\pi/2} \right\}$

De manera análoga resulta que el desvío estándar del predictor \hat{y}_0 vale

$SD(\hat{y}_0) = \sqrt{MS_e(1 + \mathbf{x}'_0(X'X)^{-1}\mathbf{x}_0)}$, y por tanto su intervalo de confianza resulta

$$\left\{ y \mid \hat{y}_0 - SD(\hat{y}_0)t_{n-k;\pi/2} \leq y \leq \hat{y}_0 + SD(\hat{y}_0)t_{n-k;\pi/2} \right\}$$

6.3 Partición de la regresión

En un estudio, puede resultar que el interés este concentrado solo sobre algunos parámetros. Esto significa que se consideran dos conjuntos de $k_1 + k_2 = k$ regresores separados, X_1 y X_2 , y por esto el modelo puede escribir como

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}$$

donde $\mathbf{X} = (\mathbf{X}_1 \mid \mathbf{X}_2)$, $\beta' = (\beta'_1 \mid \beta'_2)$. Bajo esta partición la matriz $X'X$ toma la forma

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \end{pmatrix} (\mathbf{X}_1|\mathbf{X}_2) = \begin{pmatrix} \mathbf{X}'_1\mathbf{X}_1 & \mathbf{X}'_1\mathbf{X}_2 \\ \mathbf{X}'_2\mathbf{X}_1 & \mathbf{X}'_2\mathbf{X}_2 \end{pmatrix}$$

Si solo se esta interesado en los β_2 , los β_1 se llaman parámetros de fastidio. Tiene que distinguir en el análisis dos casos: si los dos conjuntos de regresores son ortogonales o sea $\mathbf{X}'_1\mathbf{X}_2 = \mathbf{0}$, o no.

6.4 Dos conjuntos ortogonales

En este caso, bajo la condición $\mathbf{X}'_1\mathbf{X}_2 = \mathbf{0}$ la matriz $\mathbf{X}'\mathbf{X}$ toma la forma $\begin{pmatrix} \mathbf{X}'_1\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'_2\mathbf{X}_2 \end{pmatrix}$

y entonces

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} (\mathbf{X}'_1\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}'_2\mathbf{X}_2)^{-1} \end{pmatrix} \quad (172)$$

En consecuencia la estimación de β se puede dividir en dos partes independientes, por lo que resulta

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{y} \\ (\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{y} \end{pmatrix}$$

con varianza

$$V(\beta) = V\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} (\mathbf{X}'_1\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}'_2\mathbf{X}_2)^{-1} \end{pmatrix}$$

De aquí resulta que la covarianza entre β_1 y β_2 es cero. Por otro lado se tiene

$$\begin{aligned} \rho_S &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = (\mathbf{X}_1|\mathbf{X}_2) \begin{pmatrix} (\mathbf{X}'_1\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}'_2\mathbf{X}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \end{pmatrix} = \\ &= \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1 + \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2 = \rho_{S_1} + \rho_{S_2} \end{aligned}$$

Entonces, la proyección sobre S es la suma de las dos proyecciones sobre los espacios S_1

y S_2 generados respectivamente para X_1 y X_2 . En consecuencia

$$\hat{\eta} = \rho_S \mathbf{y} = (\rho_{S_1} + \rho_{S_2}) \mathbf{y} = \rho_{S_1} \mathbf{y} + \rho_{S_2} \mathbf{y} = \rho_{S_1} \hat{\eta}_1 + \rho_{S_2} \hat{\eta}_2 = \hat{\eta}_1 + \hat{\eta}_2$$

y como las proyecciones son simétricas e idempotentes,

$$\hat{\eta}' \hat{\eta} = \mathbf{y}' \rho_S \mathbf{y} = \mathbf{y}' \rho_{S_1} \mathbf{y} + \mathbf{y}' \rho_{S_2} \mathbf{y} = \mathbf{y}' \rho_{S_1}' \rho_{S_1} \mathbf{y} + \mathbf{y}' \rho_{S_2}' \rho_{S_2} \mathbf{y} = \hat{\eta}_1' \hat{\eta}_1 + \hat{\eta}_2' \hat{\eta}_2$$

Por los residuos resulta

$$\mathbf{e}' \mathbf{e} = \mathbf{y}' \mathbf{y} - \hat{\eta}' \hat{\eta} = \mathbf{y}' \mathbf{y} - \hat{\eta}_1' \hat{\eta}_1 - \hat{\eta}_2' \hat{\eta}_2 = \mathbf{y}' (\mathbf{I} - \rho_{S_1} - \rho_{S_2}) \mathbf{y}$$

Así se pueden organizar los resultados en la tabla de análisis de varianza siguiente:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
X_1	k_1	$SS_{X_1} = \mathbf{y}' \rho_{S_1} \mathbf{y}$	SS_{X_1} / k_1	$\sigma^2 + \beta_1' X_1' X_1 \beta_1 / k_1$
X_2	k_2	$SS_{X_2} = \mathbf{y}' \rho_{S_2} \mathbf{y}$	SS_{X_2} / k_2	$\sigma^2 + \beta_2' X_2' X_2 \beta_2 / k_2$
Error	$n - k_1 - k_2$	$SS_e = \mathbf{y}' (\mathbf{I} - \rho_{S_1} - \rho_{S_2}) \mathbf{y}$	$SS_e / (n - k_1 - k_2)$	σ^2
Total	n	SS_t		

Si el interés solo está concentrado sobre β_2 se puede borrar la primera línea y cambiar la última, de manera que se obtiene:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
X_2	k_2	$SS_{X_2} = \mathbf{y}' \rho_{S_2} \mathbf{y}$	SS_{X_2} / k_2	$\sigma^2 + \beta_2' X_2' X_2 \beta_2 / k_2$
Error	$n - k_1 - k_2$	$SS_e = \mathbf{y}' (\mathbf{I} - \rho_{S_1} - \rho_{S_2}) \mathbf{y}$	$SS_e / (n - k_1 - k_2)$	σ^2
Total	$n - k_1$	$\mathbf{y}' (\mathbf{I} - \rho_{S_1}) \mathbf{y}$		

Esta tabla muestra que cuando el interés está limitado a β_2 , la suma de cuadrados total es la suma de cuadrados de los residuos de la regresión de \mathbf{y} sobre X_1 . Esto es, dado que $\mathbf{I} - \rho_{S_1}$ es simétrica e idempotente

$$\mathbf{y}' (\mathbf{I} - \rho_{S_1}) \mathbf{y} = (\mathbf{y} - \hat{\eta}_1)' (\mathbf{y} - \hat{\eta}_1) = \mathbf{e}_1' \mathbf{e}_1$$

Esto sugiere proceder en dos etapas:

- 1) Calcular la regresión *de y sobre* X_1 aunque como los cuadrados de los residuos valen $\mathbf{e}_1' \mathbf{e}_1 = \mathbf{y}'(I - \rho_{S_1})\mathbf{y}$, su esperanza no vale más σ^2 , ya que también X_2 es considerado como regresor, sino $E(\mathbf{e}_1' \mathbf{e}_1) = (n - k_1) \sigma^2 + \beta_2 X_2' X_2 \beta_2$.
- 2) Calcular la regresión *de* \mathbf{e}_1 *sobre* X_2 . Entonces \mathbf{e}_1 es una nueva variable criterio que se tiene que explicar. Desde el punto de vista de la estimación de los parámetros, se consigue lo mismo: efectivamente resulta

$$\begin{aligned} \beta_{\mathbf{e}_1} &= (X_2' X_2)^{-1} X_2' \mathbf{e}_1 = (X_2' X_2)^{-1} X_2' (I - \rho_{S_1}) \mathbf{y} = \\ &= (X_2' X_2)^{-1} (X_2' - X_2' \rho_{S_1}) \mathbf{y} = (X_2' X_2)^{-1} X_2' \mathbf{y} = \beta_2 \mathbf{y} \end{aligned}$$

porque $X_2' \rho_{S_1} = 0$. Las sumas de cuadrados de las regresiones son las mismas:

$$\mathbf{e}_1' \rho_{S_2} \mathbf{e}_1 = \mathbf{y}' \mathcal{E}_{S_1}' \rho_{S_2} \mathcal{E}_{S_1} \mathbf{y} = \mathbf{y}' \rho_{S_2} \mathbf{y}$$

pues las proyecciones son idempotentes y $\rho_{S_2} \mathcal{E}_{S_1} \mathbf{y} = \rho_{S_2} \mathbf{y}$ por la ortogonalidad. Por tanto

$$\mathbf{e}_1' \mathcal{E}_{S_2} \mathbf{e}_1 = \mathbf{y}' \mathcal{E}_{S_1}' \mathcal{E}_{S_2}' \mathcal{E}_{S_2} \mathcal{E}_{S_1} \mathbf{y} = \mathbf{y}' \mathcal{E}_{S_2} \mathbf{y}$$

y también las sumas de cuadrados de los residuos son las mismas.

6.5 Caso no ortogonal

En el caso que $X_1' X_2 \neq 0$, la matriz inversa $(X'X)^{-1}$ no tiene más la forma (172) así que no se pueden tratar los dos conjuntos de regresores de manera independiente, porque los espacios S_1 y S_2 no son ortogonales. No obstante se puede volver al caso de independencia a través de un proceso de *ortogonalización*. Se trata de escribir la matriz en dos partes, $X_2 = X_{21} + X_{2\cdot 1}$, donde cada columna de X_{21} es la regresión de la columna correspondiente sobre X_1 y la matriz $X_{2\cdot 1}$ es formada por los residuos de esta regresión, o sea

$$\begin{aligned} X_{21} &= \rho_{S_1} X_2 = X_1 (X_1' X_1)^{-1} X_1' X_2 = X_1 A \\ X_{2\cdot 1} &= \mathcal{E}_{S_1} X_2 = X_2 - X_1 (X_1' X_1)^{-1} X_1' X_2 = X_2 - X_1 A \end{aligned}$$

donde $A = (X_1' X_1)^{-1} X_1' X_2$ es de orden $k_1 \times k_2$. Resulta $X_{2\cdot 1} \perp X_1$, por cuanto se tiene que

$$X_1' X_{2\cdot 1} = X_1' X_2 - X_1' X_1 (X_1' X_1)^{-1} X_1' X_2 = 0$$

Así se ha ortogonalizado los regresores X_2 con respecto a X_1 . El espacio S siempre va ser el

mismo, porque solo su base cambió. Efectivamente, los vectores de X_2 , ya combinaciones lineales de vectores de S pero independientes de X_1 , se remplazaron con estos últimos, escritos $X_{2\cdot 1}$, ortogonales a X_1 . Con respecto al cambio de base, resulta que

$$\begin{aligned} E(\mathbf{y}) &= X_1\beta_1 + X_2\beta_2 = \\ &= X_1\beta_1 + X_1A\beta_2 - X_1A\beta_2 + X_2\beta_2 = \\ &= X_1(\beta_1 + A\beta_2) + (X_2 - X_1A)\beta_2 = \\ &= X_1\phi + X_{2\cdot 1}\beta_2 \end{aligned}$$

con $X_{2\cdot 1} \perp X_1$. Por tanto se puede escribir $S = S_1 \oplus S_{2\cdot 1}$ y se vuelve al caso anterior, con el modelo

$$\boldsymbol{\eta} = E(\mathbf{y}) = X_1\boldsymbol{\phi} + X_{2\cdot 1}\boldsymbol{\beta}_2 \quad (186)$$

que permite de estimar los parámetros mediante

$$\begin{pmatrix} \hat{\boldsymbol{\phi}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} (X_1'X_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (X_{2\cdot 1}'X_{2\cdot 1})^{-1} \end{pmatrix} \begin{pmatrix} X_1' \\ X_{2\cdot 1}' \end{pmatrix} \mathbf{y} = \begin{pmatrix} (X_1'X_1)^{-1}X_1'\mathbf{y} \\ (X_{2\cdot 1}'X_{2\cdot 1})^{-1}X_{2\cdot 1}'\mathbf{y} \end{pmatrix}$$

Entonces se obtiene como resultado

$$\begin{aligned} \hat{\boldsymbol{\eta}} &= \rho_S \mathbf{y} = (\rho_{S_1} + \rho_{S_{2\cdot 1}}) \mathbf{y} = \hat{\boldsymbol{\eta}}_1 + \hat{\boldsymbol{\eta}}_{2\cdot 1} \\ \hat{\boldsymbol{\eta}}' \hat{\boldsymbol{\eta}} &= \mathbf{y}' \rho_S \mathbf{y} = \mathbf{y}' \rho_{S_1} \mathbf{y} + \mathbf{y}' \rho_{S_{2\cdot 1}} \mathbf{y} = \hat{\boldsymbol{\eta}}_1' \hat{\boldsymbol{\eta}}_1 + \hat{\boldsymbol{\eta}}_{2\cdot 1}' \hat{\boldsymbol{\eta}}_{2\cdot 1} \\ \mathbf{e}' \mathbf{e} &= \mathbf{y}' \mathbf{y} - \hat{\boldsymbol{\eta}}_1' \hat{\boldsymbol{\eta}}_1 - \hat{\boldsymbol{\eta}}_{2\cdot 1}' \hat{\boldsymbol{\eta}}_{2\cdot 1} = \mathbf{y}' (I - \rho_{S_1} - \rho_{S_{2\cdot 1}}) \mathbf{y} \end{aligned}$$

Estos resultados pueden organizarse en la tabla de análisis de varianza siguiente:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
$X_{2\cdot 1}$	k_2	$SS_{X_{2\cdot 1}} = \mathbf{y}' \rho_{S_{2\cdot 1}} \mathbf{y}$	$SS_{X_{2\cdot 1}} / k_2$	$\sigma^2 + \boldsymbol{\beta}_2' X_{2\cdot 1}' X_{2\cdot 1} \boldsymbol{\beta}_2 / k_2$
Error	$n - k_1 - k_2$	$SS_e = \mathbf{y}' (I - \rho_{S_1} - \rho_{S_{2\cdot 1}}) \mathbf{y}$	$SS_e / (n - k_1 - k_2)$	σ^2
Total	$n - k_1$	$\mathbf{y}' (I - \rho_{S_1}) \mathbf{y}$		

Esto sugiere un proceso en tres etapas:

- 1) Calcular la regresión de X_2 sobre X_1 , formar la matriz de residuos $X_{2\cdot 1}$ y rescribir el modelo de la forma (186).
- 2) Calcular la regresión de \mathbf{y} sobre X_1 y buscar $\mathbf{e}_1 = (I - \rho_{S_1})\mathbf{y}$.
- 3) Calcular la regresión de \mathbf{e}_1 sobre $X_{2\cdot 1}$.

6.6 Observación común

Se puede observar que, cuando se incluyó X_2 en el modelo la suma de cuadrados de la regresión aumentó y la suma de cuadrados de los residuos disminuyó en la misma cantidad, así que siempre se puede escribir

$$\begin{aligned} SS_r(\beta_2) &= SS_e(\beta_1) - SS_e(\beta_1, \beta_2) \\ SS_r(\beta_2) &= SS_r(\beta_1, \beta_2) - SS_r(\beta_1) \end{aligned}$$

6.7 La eliminación del promedio

Ocurre a menudo que el modelo más apropiado es de la forma

$$E(\mathbf{y}) = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \dots + \beta_{k-1} \mathbf{x}_{k-1}$$

donde $\mathbf{1} = (1, 1, \dots, 1)'$ tiene dimensión $n \times 1$, y donde el interés se centra sobre los parámetros $\beta_1, \beta_2, \beta_{k-1}$. Se puede observar que, si estos no son todos cero, $\beta_0 = \bar{y}$ el promedio de los y_i , que no es de interés, mientras que, si interesan los otros factores. Se vuelve así al modelo

$$E(\mathbf{y}) = (\mathbf{1}' | \mathbf{X}_1) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

donde $\beta_1 = (\beta_1, \beta_2, \dots, \beta_{k-1})$ es de dimensión $k_1 = k - 1$, donde queremos estimar los parámetros y la suma de cuadrados de los errores, conociendo la varianza de \mathbf{y} , o sea $\sigma^2 I$. Se debe entonces ortogonalizar previamente los X_1 por respecto al vector $\mathbf{1}$, obteniéndose

$$\mathbf{A} = (\mathbf{x}_0' \mathbf{x}_0)^{-1} \mathbf{x}_0' \mathbf{X}_1 = \frac{1}{n} \mathbf{1}' \mathbf{X}_1 = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{k-1}) = \bar{\mathbf{x}}'$$

En consecuencia resulta

$$\mathbf{X}_{1 \cdot 0} = \mathbf{X}_1 - \mathbf{1} \bar{\mathbf{x}}' = (\mathbf{x}_1 - \bar{x}_1 \mathbf{1} | \mathbf{x}_2 - \bar{x}_2 \mathbf{1} | \dots | \mathbf{x}_{k-1} - \bar{x}_{k-1} \mathbf{1})$$

lo que corresponde a haber centrado cada regresor alrededor de su promedio. Entonces es claro que

$$\mathbf{1}' \mathbf{X}_{1 \cdot 0} = (\mathbf{1}' (\mathbf{x}_i - \bar{x}_i \mathbf{1}))'_{i=1,2,\dots,n} = \mathbf{0}$$

como vector de desvíos al promedio. Entonces $\mathbf{1}$ y $\mathbf{X}_{1 \cdot 0}$ son ortogonales y se puede estimar

el modelo

$$\begin{cases} \mathbf{y} &= \mathbf{x}'_0 \phi + \mathbf{X}_{1,0} \beta_1 + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ V(\boldsymbol{\varepsilon}) &= \sigma^2 \mathbf{I} \end{cases} \quad (195)$$

Resulta que $\mathbf{1}'\mathbf{y} = \phi \mathbf{1}'\mathbf{1} + \mathbf{1}'\mathbf{X}_{1,0} \beta_1 + \mathbf{1}'\boldsymbol{\varepsilon} = n\phi + n\bar{\varepsilon}$, donde

$$\bar{y} = \phi + \bar{\varepsilon}$$

con

$$E(\bar{y}) = \phi$$

pero también

$$E(\hat{\phi}) = \phi = \bar{y}$$

Resulta que $\phi = \beta_0 + \bar{\mathbf{x}}'\beta_1$ y que

$$\begin{pmatrix} \hat{\phi} \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} (\mathbf{1}'\mathbf{1})^{-1} & \mathbf{0} \\ \mathbf{0} & |(\mathbf{X}'_{1,0}\mathbf{X}_{1,0})^{-1}| \end{pmatrix} \begin{pmatrix} \mathbf{1}' \\ \mathbf{X}_{1,0} \end{pmatrix} \mathbf{y} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}'_{1,0}\mathbf{X}_{1,0})^{-1}\mathbf{X}'_{1,0}\mathbf{y} \end{pmatrix}$$

Sobre la base del teorema de Gauss, $\hat{\phi}$ es un estimador insesgado de varianza mínima lineal en \mathbf{y} . Se verifica fácilmente que las varianzas de los parámetros valen

$$V \begin{pmatrix} \hat{\phi} \\ \hat{\beta}_1 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & |(\mathbf{X}'_{1,0}\mathbf{X}_{1,0})^{-1}| \end{pmatrix}$$

así que $\hat{\phi}$ y $\hat{\beta}_1$ son no correlacionados entre ellos.

Se sigue el proceso haciendo la regresión de \mathbf{y} sobre $\mathbf{1}$ para calcular los residuos. Esto nos deja

$$\mathbf{e}_0 = \mathcal{E}_1 \mathbf{y} = (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\mathbf{y} = \mathbf{y} - \bar{y}\mathbf{1}$$

correspondiendo a centrar \mathbf{y} , o sea *eliminar el promedio*. Por tanto se hace la regresión de \mathbf{e}_0 sobre $\mathbf{X}_{1,0}$ para obtener $\hat{\beta}_1$. Resulta entonces la tabla de análisis de varianza

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
$X_{1:0}$	$k - 1$	$SS_{S_{1:0}} = \beta_1' X_{1:0}' X_{1:0} \beta_1$	$SS_{X_{1:0}} / (k-1)$	$\sigma^2 + \beta_1' X_{1:0}' X_{1:0} \beta_1 / (k-1)$
Error	$n - k$	$SS_e = e_0' e_0 - y' \rho_{S_{1:0}} y$	$SS_e / (n - k)$	σ^2
Total	$n - 1$	$e_0' e_0 = \sum (y_i - \bar{y})^2$		

Seguendo esta tabla se puede definir el coeficiente de determinación de la misma manera que en la regresión simple:

$$R^2 = \frac{SS_{S_{1:0}}}{e_0' e_0} = 1 - \frac{SS_e}{e_0' e_0}$$

Se puede mostrar que su raíz cuadrada es el *coeficiente de correlación múltiple* entre y y X_1 , o sea entre y y $\hat{\eta}$:

$$R = \frac{\sum (y_i - \bar{y})(\hat{\eta}_i - \bar{\hat{\eta}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{\eta}_i - \bar{\hat{\eta}})^2}} \quad (203)$$

6.8 La falta de ajuste del modelo lineal

En las secciones precedentes se hizo la hipótesis de saber que la relación entre X y y era lineal o era una buena aproximación lineal. Sin embargo hay situaciones donde esto tiene que comprobarse. Esto significa que el ajuste lineal es exhaustivo por cuanto concierne a la estimación de los valores de la variable criterio basandose sobre el conocimiento de los valores de los estimadores x . Si esto no es verdadero, significa que hay otros efectos de los x sobre y , que no fueron considerados en el expresados modelo lineal y que se necesita un modelo diferente, donde una parte lineal solo sea una componente.

Como siempre se empieza con el hecho que $E(y|X) = \eta = X\beta$ y la varianza de los y es $V(y) = \sigma^2 I$. Se sabe que si el modelo ajusta bien a los datos, el cuadrado medio de los errores MS_e es un estimador insesgado de esta varianza.

Supongamos entonces que el modelo no ajuste bien, traduciendo esto en

$$E(y|X) = \gamma \neq \eta = X\beta$$

lo cual significa que $\boldsymbol{\gamma}$ no se encuentra en el sub-espacio S generado para las columnas de la matriz X . En este caso se puede proyectar $\boldsymbol{\gamma}$ sobre X y se consigue su proyección

$$\boldsymbol{\gamma}^\circ = \boldsymbol{\rho}\boldsymbol{\gamma}$$

y el vector

$$\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ = \boldsymbol{\mathcal{E}}\boldsymbol{\gamma}$$

que se puede llamar *vector residual del modelo*. Este informa sobre el desvío entre la esperanza verdadera y la hipotizada. Su cuadrado se indica con

$$\Lambda^2 = (\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ)'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ)$$

así que, si $\boldsymbol{\gamma} = \boldsymbol{\eta}$, entonces $\Lambda^2 = 0$.

Ahora, proyectando \boldsymbol{y} sobre S se consigue el vector de residuos, el cual se puede escribir como

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{\hat{\eta}} = \boldsymbol{y} - \boldsymbol{\rho}\boldsymbol{y} = \boldsymbol{\mathcal{E}}\boldsymbol{y}$$

Tomando las esperanzas, resulta

$$\boldsymbol{E}(\boldsymbol{e}) = \boldsymbol{E}(\boldsymbol{y}) - \boldsymbol{E}(\boldsymbol{\hat{\eta}}) = \boldsymbol{E}(\boldsymbol{y}) - \boldsymbol{\rho}\boldsymbol{E}(\boldsymbol{y}) = \boldsymbol{\gamma} - \boldsymbol{\rho}\boldsymbol{\gamma} = \boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ$$

Entonces \boldsymbol{e} informa sobre el vector residual del modelo, o sea sobre el desvío entre el modelo verdadero y el considerado. Al contrario, como $\boldsymbol{E}(\boldsymbol{\hat{\eta}}) = \boldsymbol{\gamma}^\circ$, $\boldsymbol{\hat{\eta}}$ no puede informar sobre $\boldsymbol{\gamma}$.

Ahora calculamos las varianzas. Se tiene

$$V(\boldsymbol{\hat{\eta}}) = \sigma^2\boldsymbol{\rho} \quad \text{y} \quad V(\boldsymbol{e}) = \sigma^2\boldsymbol{\mathcal{E}}$$

y las esperanzas de los cuadrados medios son respectivamente

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{\hat{\eta}}'\boldsymbol{\hat{\eta}}) &= \boldsymbol{E}(\boldsymbol{y}'\boldsymbol{\rho}\boldsymbol{y}) = \boldsymbol{\gamma}^\circ'\boldsymbol{\gamma}^\circ + k\sigma^2 \\ \boldsymbol{E}(\boldsymbol{e}'\boldsymbol{e}) &= \boldsymbol{E}(\boldsymbol{y}'\boldsymbol{\mathcal{E}}\boldsymbol{y}) = \Lambda^2 + (n-k)\sigma^2 \end{aligned}$$

Bajo estas condiciones se puede decir que MS_e no informa más sobre la varianza de los errores, pero sí sobre la falta de ajuste. Esto permite decir que, si se conocía una estimación de la varianza *independiente del modelo*, se podría testar, bajo la hipótesis que la distribución de \boldsymbol{y} es normal, la igualdad de los dos estimadores, bajo la hipótesis $\Lambda^2 = 0$.

Si no se conoce directamente la varianza de \boldsymbol{y} se puede organizar la experimentación de manera de conseguir medidas repetidas de \boldsymbol{y} correspondientes a los mismos valores de \boldsymbol{x} . Esto se puede conseguir, a condición que sean $m > k$ valores diferentes del vector \boldsymbol{x} y que por lo menos un valor (pero mejor muchos) tenga al menos dos repeticiones (pero mejor muchas). Supongamos que se ordenaron las líneas de la matriz X y del vector \boldsymbol{y} , de modo que se agruparon las observaciones repetidas y que X sea de rango completo.

Con la estimación ordinaria de mínimos cuadrados se encuentra $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, lo que significa que la estimación de la observación y_{ij} es $\hat{\eta}_j = \mathbf{x}_j'\hat{\boldsymbol{\beta}}$. El desvío vale $e_{ij} = y_{ij} - \hat{\eta}_j$, así que la suma de cuadrados de los residuos vale

$$SS_e = \mathbf{e}'\mathbf{e} = \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \hat{\eta}_j)^2 \quad (212)$$

Por intuición se comprende que si la regresión no era lineal, los residuos tienen que contener alguna información sobre esto, porque los $y_{i,j}$ informan sobre los valores *verdaderos*, mientras $\hat{\boldsymbol{\eta}}$ solo informa sobre la linealidad. Entonces SS_e informa sobre σ^2 pero también sobre el desvío a la linealidad de la función verdadera y por tanto será más grande que σ^2 . Como se repitieron medidas por los mismos x_i , ya se puede medir σ^2 y averiguar su diferencia con SS_e .

Introduciendo en SS_e el promedio de los y en cada grupo se puede escribir

$$\begin{aligned} SS_e &= \mathbf{e}'\mathbf{e} = \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \hat{\eta}_j)^2 \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j + \bar{y}_j - \hat{\eta}_j)^2 = \\ &= \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^m n_j (\bar{y}_j - \hat{\eta}_j)^2 = \\ &= SS_W + SS_M \end{aligned} \quad (213)$$

Así, SS_e resulta compartido en dos partes, una, una varianza *intra* SS_W , que no informa sino que sobre σ^2 , mientras la otra, SS_M , informa sobre la falta de ajuste lineal. Resulta que

$$\begin{aligned} E(SS_W) &= E\left(\sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2\right) = \sum_{j=1}^m (n_j - 1)\sigma^2 = (n - m)\sigma^2 \\ E(SS_M) &= E\left(\sum_{j=1}^m n_j (\bar{y}_j - \hat{\eta}_j)^2\right) = \sum_{j=1}^m n_j E((\bar{y}_j - \hat{\eta}_j)^2) = (m - k)\sigma^2 + \Lambda^2 \end{aligned} \quad (214)$$

Se ve bien que $MS_W = SS_W / (n - m)$ es un estimador insesgado de σ^2 , *que no depende del modelo*.

Más aún SS_W y SS_M son independientes. Eso se puede ver escribiéndolos en forma matricial. Si se define

$$U = \begin{pmatrix} I_{n_1} - \frac{1}{n_1} \mathbf{1}_{n_1} \mathbf{1}_{n_1}' & \circ & \dots & \circ \\ \circ & I_{n_2} - \frac{1}{n_2} \mathbf{1}_{n_2} \mathbf{1}_{n_2}' & \dots & \circ \\ \vdots & & \ddots & \vdots \\ \circ & \dots & \circ & I_{n_m} - \frac{1}{n_m} \mathbf{1}_{n_m} \mathbf{1}_{n_m}' \end{pmatrix}$$

se puede escribir $SS_W = \mathbf{y}'U\mathbf{y}$

mientras

$$SS_L = \mathbf{y}'\mathbf{y} - \mathbf{y}'\phi\mathbf{y} - \mathbf{y}'U\mathbf{y} = \mathbf{y}'(I - \phi - U)\mathbf{y} = \mathbf{y}'Z\mathbf{y}$$

con $Z = I - X(X'X)^{-1}X' - U$. Ahora, debido a la estructura en bloques de Z se ve que $UZ = 0$. Entonces, en base al teorema de Craig, bajo $\Lambda^2 = 0$, las dos formas son independientes. Se puede resumir todos estos resultados en la tabla de análisis de varianza siguiente

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Inter	k	$\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \mathbf{y}'\phi\mathbf{y}$		$\sigma^2 + \boldsymbol{\gamma}'\boldsymbol{\gamma}$
Inter falta de ajuste	$m-k$	$SS_M = \sum_j n_j (\bar{y}_j - \hat{\eta}_j)^2$	$MS_M = SS_M / (m - k)$	$\sigma^2 + \Lambda^2 / (m - k)$
Intra	$n - m$	$SS_W = \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	$MS_W = SS_W / (n - m)$	σ^2
Total	n	$SS_T = \mathbf{y}'\mathbf{y}$		

Para testar el ajuste del modelo, se puede rechazar la hipótesis de linealidad a nivel de probabilidad π si

$$F_M = \frac{MS_M}{MS_W} > F_{m-k, n-m; \pi}$$

y aceptarla en caso contrario.

7. Selección de modelos

7.1 Las matrices de covarianza y de correlación

En lo que sigue será útil referirse al conjunto de relaciones lineales entre las variables que se estudian, y esto independientemente de lo hecho de deber explicar una variable criterio para variables explicativas, sino de manera general. Claro que el conjunto de estas relaciones será informativo también en un contexto asimétrico como es el caso de la regresión.

Por esto se introducen algunas matrices, que informan sobre este conjunto de relaciones. Supongamos dada una tabla de datos X con n líneas representando las n observaciones que se hicieron sobre p variables *cuantitativas*. Ya sabemos a través el estudio de la regresión de una variable y sobre x que el coeficiente de correlación $r(x,y)$, resultando del cálculo de los residuos de una regresión lineal, informa sobre la intensidad de la relación lineal entre las variables mismas. En particular, resulta cero en el caso que ninguna parte de SS_y sea explicada para una recta de regresión sobre x y ± 1 en el caso de relación funcional perfecta positiva o negativa. En realidad, junto al coeficiente de correlación, se encuentran también las sumas S_{xx} , S_{yy} y S_{xy} así que las mismas sumas pero centradas, o sea las sumas de los desvíos a los promedios $S_{\bar{x}\bar{x}}$, $S_{\bar{y}\bar{y}}$ y $S_{\bar{x}\bar{y}}$. Dividendo estas para n se consiguen respectivamente las varianzas de x y y y la covarianza entre ellas. Se conoce también la relación entre covarianza y correlación, o sea

$$r = \text{corr}(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{S_{\bar{x}\bar{y}}}{\sqrt{S_{\bar{x}\bar{x}}S_{\bar{y}\bar{y}}}} \quad (219)$$

En el caso de una tabla de datos X con p variables en columna, si hay interés a estudiar el conjunto de relaciones lineales entre variables, se usa sintetizar este conjunto a través de matrices (simétricas y semi-definidas positivas). Específicamente, se introducen la matriz de varianza-covarianza

$$V(X) = \begin{pmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \dots & \text{cov}(x_1, x_p) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \dots & \text{cov}(x_2, x_p) \\ \dots & \dots & \dots & \dots \\ \text{cov}(x_p, x_1) & \text{cov}(x_p, x_2) & \dots & \text{var}(x_p) \end{pmatrix} \quad (220)$$

y la matriz de correlación

$$C(X) = \begin{pmatrix} 1 & \text{corr}(x_1, x_2) & \dots & \text{corr}(x_1, x_p) \\ \text{corr}(x_2, x_1) & 1 & \dots & \text{corr}(x_2, x_p) \\ \dots & \dots & \dots & \dots \\ \text{corr}(x_p, x_1) & \text{corr}(x_p, x_2) & \dots & 1 \end{pmatrix} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix} \quad (221)$$

A veces puede ser útil saber como se pueden calcular de manera sintética dichas matrices. Sabiendo que el producto de la tabla de datos con su traspuesta vale

$$X'X = \begin{pmatrix} S_{x_1x_1} & S_{x_1x_2} & \dots & S_{x_1x_p} \\ S_{x_2x_1} & S_{x_2x_2} & \dots & S_{x_2x_p} \\ \dots & \dots & \dots & \dots \\ S_{x_px_1} & S_{x_px_2} & \dots & S_{x_px_p} \end{pmatrix} \quad (222)$$

la matriz de varianza-covarianza resulta del centrado de los datos alrededor de su promedio dividido por $n - 1$, o sea, definiendo el vector $S' = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ de los promedios

$$V(X) = \frac{1}{n-1} \begin{pmatrix} S_{\hat{x}_1\hat{x}_1} & S_{\hat{x}_1\hat{x}_2} & \dots & S_{\hat{x}_1\hat{x}_p} \\ S_{\hat{x}_2\hat{x}_1} & S_{\hat{x}_2\hat{x}_2} & \dots & S_{\hat{x}_2\hat{x}_p} \\ \dots & \dots & \dots & \dots \\ S_{\hat{x}_p\hat{x}_1} & S_{\hat{x}_p\hat{x}_2} & \dots & S_{\hat{x}_p\hat{x}_p} \end{pmatrix} = \frac{1}{n-1} (X'X - nSS') \quad (224)$$

Entonces, si se define la matriz diagonal de los desvíos estándar $\sigma(X) = \text{diag}(\sigma_{x_1}, \sigma_{x_2}, \dots, \sigma_{x_p})$, resulta

$$C(X) = \sigma(X)^{-1}V(X)\sigma(X)^{-1} \quad (226)$$

7.2 El coeficiente de correlación parcial

En la página 53, hemos estudiado el coeficiente de correlación múltiple (203) en el contexto de la regresión múltiple entre y y su estimadores X .

Supongamos ahora que dos variables aleatorias x_1 y x_2 dependen linealmente de una misma variable aleatoria z . Es posible de medir directamente el coeficiente de correlación $r = \text{corr}(x_1, x_2)$ sobre una muestra de tamaño n representada en \mathbb{R}^n para los vectores con n componentes x_1

y x_2 , pero queremos conocer el vínculo entre x_1 y x_2 *eliminando el efecto de la variable z* cuyas n observaciones son las componentes del vector z .

Existe una estadística que permite calcular este vínculo entre x_1 y x_2 *bajo z constante* de manera sencilla, también con muestras pequeñas, bajo una hipótesis de linealidad de los vínculos. Se trata del *coeficiente de correlación parcial* entre x_1 y x_2 , que se escribe como:

$$r_{x_1x_2.z} = \rho(x_1, x_2 | z)$$

Suponiendo a todas las variables centradas, para no considerar el parámetro α , su cálculo se basa sobre la hipótesis que el efecto de z sobre x_1 y x_2 se manifiesta para relaciones del tipo:

$$\begin{cases} x_1 = \beta_1 z + \varepsilon_1 \\ x_2 = \beta_2 z + \varepsilon_2 \end{cases}$$

donde ε_1 y ε_2 son como siempre los residuos. Si se escribe el sistema en función de los residuos mismos,

$$\begin{cases} \varepsilon_1 = x_1 - \beta_1 z \\ \varepsilon_2 = x_2 - \beta_2 z \end{cases}$$

se pueden estudiar las relaciones entre ε_1 y ε_2 y en particular su correlación. Entonces, se define coeficiente de *correlación parcial* teórico entre y_1 y y_2 como el coeficiente de correlación entre ε_1 y ε_2

$$r_{x_1x_2.z} = \text{corr}(\varepsilon_1, \varepsilon_2)$$

7.3 Caso de tres variables

Para n observaciones de 3 variables x_1 , x_2 , z estas relaciones de ajuste se escriben

$$\begin{cases} x_1 = \beta_1 z + e_1 \\ x_2 = \beta_2 z + e_2 \end{cases}$$

donde los $\hat{\beta}_i$ son las estimaciones de mínimos cuadrados de los β_i y los e_i son los residuos

observados.

La *covarianza parcial* se escribe como:

$$\text{cov}(x_1, x_2 | z) = \frac{1}{n} e_1' e_2 = \frac{1}{n} (x_1 - \beta_1 z)' (x_2 - \beta_2 z)$$

o sea

$$\text{cov}(x_1, x_2 | z) = \frac{1}{n} (x_1' x_2 - \beta_1 z' x_2 - \beta_2 x_1' z + \beta_1 \beta_2 z' z)$$

Si se substituyen los coeficientes de regresión para su valores estimados

$\hat{\beta}_1 = \frac{x_1' z}{z' z}$ y $\hat{\beta}_2 = \frac{x_2' z}{z' z}$ se obtiene luego de una simplificación:

$$\text{cov}(x_1, x_2 | z) = \frac{1}{n} \left(x_1' x_2 - \frac{(x_1' z)(x_2' z)}{z' z} \right)$$

que se puede escribir también como:

$$\text{cov}(x_1, x_2 | z) = \text{cov}(x_1, x_2) - \frac{\text{cov}(x_1, z) \text{cov}(x_2, z)}{\text{var } z}$$

Las varianzas de los residuos se calculan de manera parecida y se encuentra, para e_i , $i = 1, 2$:

$$\frac{1}{n} e_i' e_i = \text{var } x_i - \frac{\text{cov}^2(x_i, z)}{\text{var } z} = \text{var } x_i (1 - r^2(x_i, z))$$

El coeficiente de correlación parcial $r_{x_1, x_2 \cdot z}$ se puede escribir también haciendo aparecer los coeficientes de correlación usual desde que

$$r_{x_1, x_2} = \frac{e_1' e_2}{\sqrt{(e_1' e_1)(e_2' e_2)}}$$

se obtiene

$$x_i = \beta_{i1}z_1 + \beta_{i2}z_2 + \dots + \beta_{iq}z_q + e_i, \quad i = 1, \dots, p$$

con e_i el vector de los residuos. Para todos los p ajustes, se puede construir la matriz $p \times q$ \hat{B} de los parámetros, y la matriz $n \times p$ E de los residuos, así que el sistema se puede escribir compactamente por

$$\underset{n \times p}{X} = \underset{n \times q}{Z} \underset{q \times p}{\hat{B}} + \underset{n \times p}{E}$$

En la matriz \hat{B} la i -ésima columna es

$$\beta_i = (Z'Z)^{-1}Z'x_i$$

y por tanto se puede escribir \hat{B} bajo la forma

$$\hat{B} = (Z'Z)^{-1}Z'X \tag{255}$$

Con esta notación, la matriz $p \times p$ de varianza-covarianza parcial entre X con Z constante se puede escribir:

$$\begin{aligned} V(X|Z) &= \frac{1}{n}E'E = \frac{1}{n}(X - Z\hat{B})'(X - Z\hat{B}) = \\ &= \frac{1}{n}(X'X - \hat{B}'Z'X - X'Z\hat{B} + \hat{B}'Z'Z\hat{B}) \end{aligned}$$

y substituyendo \hat{B} para su expresión (255) se obtiene

$$V(X|Z) = \frac{1}{n}(X'X - X'Z'(Z'Z)^{-1}Z'X) \tag{259}$$

Ahora imaginemos que las tablas centradas X y Z se juntan para constituir una tabla T con n líneas y $p+q$ columnas $T = [XZ]$. En este caso, la matriz de varianza-covarianza se puede compartir en 4 sub-matrices de covarianza

$$V(T) = \begin{pmatrix} V_{XX} & V_{XZ} \\ V_{ZX} & V_{ZZ} \end{pmatrix}$$

con $V_{XX} = \frac{1}{n}X'X$; $V_{XZ} = V_{ZX} = \frac{1}{n}Z'X$; $V_{ZZ} = \frac{1}{n}Z'Z$. Entonces, empleando (259) resulta

$$V(X|Z) = V_{XX} - V_{ZX}V_{ZZ}^{-1}V_{XZ}$$

La matriz de correlaciones parciales se calcula fácilmente empezando con la de $V(X|Z)$ como se calculó la matriz de correlación ordinaria: tomando $\sigma(x|Z) = \sqrt{\text{var}(x|Z)}$ análogamente a (226) se obtiene

$$C(X|Z) = \sigma(X|Z)V(X|Z)\sigma(X|Z) \quad (263)$$

7.5 Técnicas de regresión

Supongamos que para modelar y , *variable que explicar o criterio*, se dispone de p predictores x_1, x_2, \dots, x_p . En vez de explicar y con todas las p variables explicativas, se puede intentar de explicar y solo con un sub-conjunto de q variables extraídas de las p disponibles, de manera de conseguir una explicación casi igualmente satisfactoria de y .

Existen muchas razones para esta operación: reducir el numero de predictores, seleccionarlos entre un numero demasiado grande, obtener fórmulas más estables pero con buena capacidad de predicción. A parte tiene que haber cuidado en dos asuntos principales: 1) aumentando los predictores el coeficiente de correlación múltipla siempre aumenta, pero también aumenta la varianza de los estimadores; 2) aumentando los predictores aumenta el riesgo de collinealidad y por consecuencia la inestabilidad de los parámetros.

7.6 El registro exhaustivo (*all possible regression*)

Es un método que consiste en estudiar todas las posibles regresiones. Considerando que el término de intersección β_0 se encuentra en todas las ecuaciones, resultan 2^p regresiones ha comparar. Para esto se necesitan métodos que permiten una comparación rápida

7.7 Método paso a paso: selección para adelante

En lugar de seleccionar entre todas las regresiones posibles, hay métodos que automáticamente construyen un modelo *paso a paso*, claramente sub-optimal, a través de una secuencia de regresiones ligadas, obtenidas ajuntando o quitando un predictor a cada paso.

El método de selección para adelante consiste en juntar a cada paso una variable en el

modelo, en base a su capacidad predictiva. Para esto se empieza considerando una regresión limitada al término constante β_0 . Luego se selecciona la variable cuyo coeficiente de correlación con y sea más grande en valor absoluto, digamos x_1 . En este método, se selecciona x_1 como la variable cuya F testando la significatividad de la regresión es mas grande, pero siempre considerando que sea significativa al nivel de probabilidad π fijado previamente.

Como segundo se elije la variable x_2 cuya correlación con los residuos de la regresión hecha en le paso precedente sea más grande o sea, o lo que es lo mismo, cuya correlación parcial con y , bajo los efectos conocidos de los predictores ya en el modelo, es más grande. Como en cada caso, el coeficiente de determinación aumenta, interesa un método parecido si, a cada paso, hay un test para elegir si se consigue un modelo realmente mayor o no del precedente. Para este asunto se hace uso del test F de entrada F_E , o sea testando si la estadística

$$F_E = \frac{SS_R(x_2 | x_1)}{MS_E(x_1, x_2)} > F_{\pi, 1, n-1}$$

es significativa o no a nivel de probabilidad π . Si el test es positivo, la variable x_2 cuya F_E es más significativa se adjunta al modelo. En general, se itera esto proceso, en el paso s eligiendo la variable x_s cuya correlación con los residuos de la regresión hecha al paso $s - 1$ sea más grande o sea cuya correlación parcial con y , bajo los efectos conocidos de los predictores ya en el modelo, es mas grande. Claro que la F de entrada se modifica a cada paso teniendo en cuenta los $s - 1$ predictores

$$F_E = \frac{SS_R(x_s | x_1, \dots, x_{s-1})}{MS_E(x_1, \dots, x_s)} > F_{\pi, 1, n-s+1}$$

La construcción del modelo culmina si no hay mas variables o si por ninguna de las variables fuera del modelo se encuentra una F de entrada significativa.

7.8 La selección para atrás

La selección para adelante tiene el defecto que alguna variable, aunque importante para describir el fenómeno del punto de vista causal, podría no resultar incluida en el modelo final, debido a la presencia de otros predictores que *cobren* influencia, así que resulta una F_E no significativa por esta. Por esta razón algunos consideran mejor la *selección para atrás*, porque con esto criterio se puede averiguar que ningún predictor importante sea olvidado. La selección *para atrás* busca un buen modelo empezando con todas las p variables y procediendo al revés, o sea quitando un predictor paso a paso. Para elegir el predictor a quitar, se utiliza igualmente la estadística F , pero esta vez en la forma de F de salida, F_S , o sea

$$F_S = \frac{SS_R(x_s | x_1, \dots, x_{s-1})}{MS_E(x_1, \dots, x_s)} < F_{\pi, 1, n-s+1}$$

quedando la variable x_s cuya F_S sea mas baja. El proceso culmina cuando no hay más predictores o cuando ningún predictor tiene una F_S no significativa.

7.9 La regresión paso a paso (*stepwise*)

Los dos métodos, para adelante y para atrás sugieren muchas combinaciones posibles. La más conocida consiste en el método *stepwise*. Supongamos nos encontramos en un paso r en el cual s predictores ya entraron al modelo: en este paso, en base a las correlaciones parciales, se testan todas las variables sobre la base de sus F parciales. Una variable junta en un paso precedente puede ser devenida redundante por causa de su relaciones con variables adjuntas luego en el modelo. Así se examinan las F_S de las variables que ya se encuentran en el modelo y se queda la que resulta con la F_S no significativa y mínima, iterando el proceso si necesario; luego se testan todas las F_E de las variables que no entraron en el modelo y se incluye la que resulta con la F_E más significativa y máxima. Claro que para que el método funcione, se necesita que las estadísticas de referencia sean diferentes. Por esto se elige el nivel de la F de entrada un poco más grande que el nivel de la F de salida, así que la entrada de una variable es más difícil que su salida.

Debe considerarse siempre que el orden de entrada o de salida de una variable en el modelo no implica su importancia relativa o absoluta con respecto de las otras. Puede ocurrir que una variable entre como primera en el modelo y luego quede eliminada por causa de la entrada de otras, cuya cualidad explicativa global resulte mejor. Esto es un problema en la selección para adelante. Resulta igualmente que los tres métodos no producen el mismo modelo de regresión y por esto se sugiere de emplearlos todos juntos. En fin, no hay razón porque exista un acuerdo entre los métodos paso a paso y lo de todas las regresiones, porque el vínculo dado para la iteratividad de los procesos no implica un mejor modelo posible.

Por último hay que precisar que en lugar de los testes F_E y F_S , se puede recurrir respectivamente a elegir en cada paso la variable que entrada según el aumento de R^2 sea máximo o la variable que quede según la disminución de R^2 sea mínimo.

8. Ejemplos

8.1 Una regresión simple

En la tabla siguiente se tomaron seis observaciones en las cuales se midieron dos variables, denotadas por x y y . Los datos se muestran en la tabla siguiente:

Unidad	x	y	xx	Sxy	$\hat{\eta}$	e	ee
P1	1	2	1,00	2,00	1,50	0,50	0,25
P2	2	3	4,00	6,00	2,50	0,50	0,25
P3	3	4	9,00	12,00	3,50	0,50	0,25
P4	2	1,5	4,00	3,00	2,50	-1,00	1,00
P5	1	0,5	1,00	0,50	1,50	-1,00	1,00
P6	0	1	0,00	0,00	0,50	0,50	0,25
Sumas	9,00	12,00	19,00	23,50	12,00	0,00	3,00

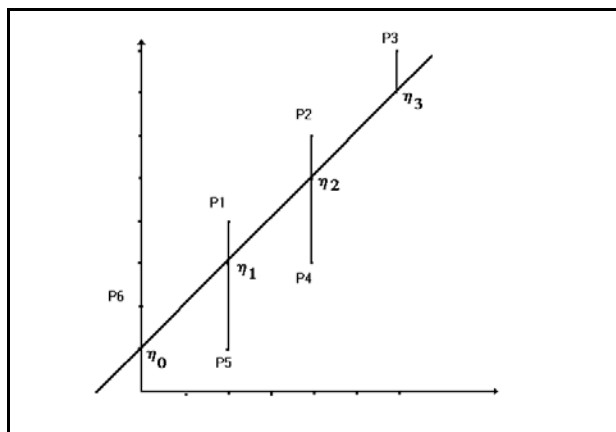


Figure 4 - Recta de regresión sobre 6 puntos y residuos.

Luego de los cálculos se consigue

$$\hat{\alpha} = 0,5 \text{ y}$$

$$\hat{\beta} = 1,0, \text{ resultando}$$

$$\hat{\eta}_{xi} = 0,5 + x_i$$

cuyos valores se encuentran en la sexta columna. Los puntos y la recta de regresión son representados en el gráfico en Fig. 4.

8.2 Una aplicación de la regresión simple

Para 13 países de América Latina, se supone una relación lineal entre el nivel de población urbana (x) y un indicador de desarrollo humano (IDH), creado para un economista pakistaní sobre la base del producto bruto, la esperanza de vida, la alfabetización y la escolarización (y). Los datos se encuentran en la tabla siguiente

<i>País</i>	<i>Tajo urbano</i> <i>x</i>	<i>IDH</i> <i>y</i>
<i>Argentina</i>	86	0,833
<i>Bolivia</i>	51	0,394
<i>Brasil</i>	75	0,739
<i>Chile</i>	86	0,863
<i>Colombia</i>	70	0,758
<i>Ecuador</i>	56	0,641
<i>Guyana</i>	35	0,539
<i>Panamá</i>	54	0,731
<i>Paraguay</i>	48	0,637
<i>Perú</i>	70	0,6
<i>Suriname</i>	48	0,749
<i>Uruguay</i>	86	0,88
<i>Venezuela</i>	84	0,00004
<i>Suma</i>	849	9,188
<i>Promedio</i>	65,308	0,70677

Por las variables se tienen las estadísticas siguientes:

<i>Estadística</i>	<i>x</i>	<i>y</i>	<i>xy</i>
<i>Mínimo</i>	35	0	
<i>Máximo</i>	86	0,88	
<i>Moyenne</i>	65,308	0,70677	
<i>Total</i>	849	9,188	
<i>Suma de cuadrados</i>	59155	6,730488	622,094
<i>Cuadrados centrados</i>	3708,767	0,2366924	22,0469
<i>Varianza</i>	285,2898	0,0182071	1,6959
<i>Desvío estándar</i>	16,89052	0,1349337	0,74412

donde se construye la tabla de análisis de varianza del modelo dada por:

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>	<i>Probabilidad</i>
<i>Regresión</i>	2	6,6248540	3,3124270	344,9337049	0,0000000
<i>Error</i>	11	,1056339	,0096031		
<i>Total</i>	13	6,7304880			

El valor de F obtenido tiene que compararse con el valor de referencia $F_{2,11;.05} = 3,9822$: entonces, la regresión tiene sentido. La estimación de α y β es la siguiente:

	<i>Estimación</i>	<i>Desvío estándar</i>	<i>t</i>
<i>beta</i>	0,005945	0,00160912	3,69426936
<i>alpha</i>	0,31854511	0,10854596	2,934656

Los valores t encontrados deben ser comparados con los valores de la función t de student con 11 grados de libertad, correspondientes a la probabilidad de 0,025, que vale 2,201. Entonces, ambas estimaciones son significativas a nivel de 5%. La covarianza entre α y β vale -.0001691. La tabla de análisis de varianza con interés sobre β es la siguiente:

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>	<i>Probabilidad</i>
β	1	,1310587	,1310587	13,6475641	0,0035366
<i>Error</i>	11	,1056339	,0096031		
<i>Total</i>	13	,2366924			

Se observa que $F_{1,11;.05} = 4,8442$. El coeficiente de determinación R^2 vale .5537090, así que la correlación entre x y y es .7441162.

Como hay repeticiones de valores de y con respecto al mismo valor de x , se puede comprobar si hay falta de ajuste lineal. Precisamente, hay tres observaciones para $x = 86$, cuya suma de cuadrados vale 0,00113267, dos para $x = 70$, con $SS = 0,012482$ y dos para $x = 48$, con $SS = 0,006272$. Entonces, resulta $SSw = 0,01988667$, donde la tabla de análisis de varianza por la falta de ajustes:

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>	<i>Probabilidad</i>
SSb	8	0,21681	0,02710071	5,4510302	0,0595
SSr(β)	1	0,13105879	0,13105879		
SSm	7	0,08574685	0,01224955	2,4638719	0,2006
SSw	4	0,01989	0,00497167		
SS _t	12	0,23669			

El segundo valor de F es lo que nos interesa (el primero no tiene interés en este caso) y tiene que ser comparado con $F_{7,4;0,5} = 6,09$. Se puede concluir aceptando la hipótesis nula que existe ajuste lineal. En la tabla siguiente se encuentran por cada país los cuadrados, la estimación, su cuadrado y el desvío estándar del promedio y de la previsión, el residuo y su cuadrado.

<i>País</i>	<i>xx</i>	<i>xy</i>	<i>yy</i>	<i>eta</i>	<i>eta ^ 2</i>	<i>SD eta</i>	<i>SD prev</i>	<i>e</i>	<i>ee</i>
<i>Argentina</i>	7396	71,638	0,69389	0,830	0,68853	0,04298	0,10701	0,00322	0
<i>Bolivia</i>	2601	20,094	0,15524	0,622	0,38653	0,03562	0,10427	-0,22772	0,05185
<i>Brasil</i>	5625	55,425	0,54612	0,764	0,58429	0,03134	0,10288	-0,02539	0,0006
<i>Chile</i>	7396	74,218	0,74477	0,830	0,68853	0,04298	0,10701	0,03322	0,0011
<i>Colombia</i>	4900	53,060	0,57456	0,735	0,53973	0,02821	0,10197	0,02334	0,0005
<i>Ecuador</i>	3136	35,896	0,41088	0,651	0,42437	0,03103	0,10279	-0,01044	0,0001
<i>Guyana</i>	1225	18,865	0,29052	0,527	0,27731	0,05583	0,11278	0,0124	0,0002
<i>Panamá</i>	2916	39,474	0,53436	0,640	0,40902	0,03271	0,10331	0,09145	0,00836
<i>Paraguay</i>	2304	30,576	0,40577	0,604	0,36467	0,03891	0,10544	0,03312	0,0011
<i>Perú</i>	4900	42,000	0,36000	0,735	0,53973	0,02821	0,10197	-0,13466	0,01813
<i>Suriname</i>	2304	35,952	0,56100	0,604	0,36467	0,03891	0,10544	0,14512	0,02106
<i>Uruguay</i>	7396	75,680	0,77440	0,830	0,68853	0,04298	0,10701	0,05022	0,00252
<i>Venezuela</i>	7056	69,216	0,67898	0,818	0,66894	0,04054	0,10605	0,00611	0
<i>Sumas</i>	59155	622,094	6,73049	9,188	6,62485			0,00000	0,10563

8.3 Un ejemplo de regresión múltiple

Se tienen 20 observaciones de 5 variables. Se pregunta si hay una relación lineal entre la y y las otras, indicadas x_1, x_2, x_3, x_4 .

N.	y	x_1	x_2	x_3	x_4
1	40	8	32	51	104
2	50	12	40	11	74
3	50	11	38	18	96
4	70	12	60	99	97
5	90	14	70	50	89
6	95	15	70	64	86
7	100	18	85	68	73
8	105	17	90	24	64
9	110	20	90	96	64
10	105	21	80	97	74
11	120	21	100	65	59
12	125	22	110	97	57
13	130	23	105	23	41
14	140	23	120	73	44
15	155	24	130	94	38
16	160	25	135	90	22
17	175	25	130	93	31
18	180	26	160	96	24
19	195	29	170	99	11
20	205	30	175	105	18

Considerando como quinto regresor una columna de 1 para la intersección, se calculan las matrices siguientes:

<i>Matriz (X'X)</i>				
20	396	1990	1413	1166
396	8574	44224	30230	19880
1990	44224	231968	156386	94206
1413	30230	156386	118387	74462
1166	19880	94206	74462	83488
<i>Matriz (X'X)⁻¹</i>				
22,2501	-0,491765	-0,051707	0,030546	-0,162551
-0,491765	0,027237	-0,001741	-0,000597	0,002879
-0,051704	-0,001741	0,000668	-0,000141	0,000508
0,030546	-0,000597	-0,000141	0,000139	-0,00025
-0,162551	0,002879	0,000508	-0,00025	0,001246
<i>Matriz X'y</i>				
2400	53045	277105	187555	115145

Resulta que $\bar{y} = 120$, $SSy = 332000$ y $SS\hat{y} = 44000$, lo cual nos permiten de estimar los parámetros $\hat{\beta} = (-.62871775, 1.74190632, .86220081, .01784631, -.01562349)$. La tabla de análisis de varianza queda

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>	<i>Probabilidad</i>
<i>Regresión</i>	5	331358,85	22090,59	516,859	0,0000000
<i>Error</i>	15	641,15	42,74		
<i>Total</i>	20	332000			

Puesto que $F_{5,15;.05} = 2.9013$ se puede rechazar la hipótesis nula $\beta = 0$. Para esta regresión se encuentra un coeficiente de determinación $r^2 = .98543$.

La matriz de varianza-covarianza de los betas vale

951,042621	-21,01964845	-2,2099835	1,30564024	-6,94796813
-21,01964845	1,16419844	-0,07440723	-0,02550462	0,12305215
-2,2099835	-0,07440723	0,02856389	-0,0060131	0,02171467
1,30564024	-0,02550462	-0,0060131	0,00595647	-0,01068903
-6,94796813	0,12305215	0,02171467	-0,01068903	0,053278

Ahora se quiere saber si todas las variables son realmente necesarias en el modelo, o sea se tiene interés sobre x_3 y x_4 . Entonces, se necesita hacer antes la regresión de y sobre (x_1, x_2) , luego ortogonalizar x_3 y x_4 con respecto a la constante y a x_1 y x_2 y finalmente hacer la regresión de los residuos de la regresión de y sobre (x_1, x_2) para las variables x_3 y x_4 ortogonalizadas.

La regresión sobre las primeras dos variables deja estos resultados

<i>Matriz (X'X)</i>		
20	396	1990
396	8574	44224
990	44224	231968
<i>Matriz (X'X)⁻¹</i>		
1,00399106	-0,11677567	0,01364992
-0,11677567	0,02058379	-0,00292245
0,01364992	-0,00292245	0,00044437
<i>Matriz X'y</i>		
2400	53045	277105
β		
-2,32533687	1,78114369	0,87496173

Se consigue la tabla de análisis de varianza siguiente:

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>	<i>Probabilidad</i>
<i>Regresión</i>	3	331356,23	19491,53	514,71	0,0000000
<i>Error</i>	17	643,77	37,87		
<i>Total</i>	20	332000			

El valor $F_{3,17;0,05} = 3.1967$ y el coeficiente $R^2 = .98537$ hacen pensar que estas dos variables son suficientes. Efectivamente, después de la ortogonalización de x_3 y x_4 se vuelve a la regresión siguiente, formada para los residuos de y , x_3 , x_4 , o sea sus componentes ortogonales al espacio generado para $\mathbf{1}$, x_1 , x_2 :

<i>e</i>	<i>ex3</i>	<i>ex4</i>
0,0774121	12,21194147	-2,67236263
-4,04685647	-32,12182782	-21,04275378
-0,51578933	-24,0383855	-1,95015599
-1,54609101	47,18690928	8,36499862
6,14200434	-6,58735423	8,10212877
9,36086065	7,19832973	7,36875574
-4,10699632	4,03693446	5,97445096
-1,70066127	-41,92156523	-3,69023791
-2,04409233	29,43548667	3,10964301
-0,07561875	34,56680208	12,17239378
-2,57485329	-6,12446081	3,58014618
-8,10561425	21,31559171	7,05064935
-0,5119493	-50,72590861	-8,28466178
-3,63637521	-7,24435578	-0,47884748
0,83286383	9,19569674	-1,00834431
-0,32308849	2,80856499	-13,13977923
19,05172015	7,98138071	-5,74171733
-3,97827535	-2,26982966	-0,86346176
-3,07132369	-4,25840921	-3,85970464
0,77272399	-0,64554097	7,00886043

donde resulta

$$\begin{array}{cc}
 \begin{array}{cc}
 \textit{Matriz (X'X)} \\
 11212.94324 & 2249.62541 \\
 2249.62541 & 1253.60624
 \end{array} &
 \begin{array}{cc}
 \textit{Matriz (X'X)}^{-1} \\
 .00013935 & -.00025008 \\
 -.00025008 & .00124646
 \end{array} \\
 \\
 \begin{array}{cc}
 \textit{Matriz X'y} \\
 164.96265 & 20.56181
 \end{array} &
 \begin{array}{c}
 \beta \\
 .01784631 \quad -.01562349
 \end{array}
 \end{array}$$

lo cual se organiza en la tabla de varianza siguiente:

<i>Fuente</i>	<i>Grados de libertad (DF)</i>	<i>Sumas de cuadrados (SS)</i>	<i>Cuadrados medios (MS)</i>	<i>Esperanza de los cuadrados medios E(MS)</i>	<i>Probabilidad</i>
<i>Regresión</i>	2	2,62	0,175	0,004	0,9960000
<i>Error</i>	15	641,15	42,74		
<i>Total</i>	17	643,77			

El valor de F muestra que no hay ninguna regresión, lo que se confirma por el coeficiente $R^2 = .00407$. Se puede concluir que los últimos dos regresores no tiene ningún interés, una vez que ya se adquirieron los primeros dos en la explicación de y .

Bibliografía

- Benzécri, J.P. *et coll.*, 1982. *L'Analyse des Données*. 2 voll., Paris, Dunod.
- Daniel, C. y F.S. Wood, 1980. *Fitting Equations to Data*. 2nd ed., New York, J. Wiley and Sons.
- Golub, G., 1969. «Matrix Decompositions and Statistical Calculations». *Statistical Computation*. New York, Academic Press: pp. 365-385.
- Guttman, I., 1982. *Linear Models: An Introduction*. New York, J. Wiley and Sons, 358 pp.
- Lehmann, E.L., 1983. *Theory of Point Estimation*. New York, J. Wiley and Sons, 506 pp.
- Monfort, A., 1982. *Cours de Statistique Mathématique*. Paris, Economica, 317 pp.
- Montgomery, D.C. et E.A. Peck, 1982. *Introduction to Linear Regression Analysis*. New York, John Wiley and Sons, 504 pp.
- Mood, A.M., F.A. Graybill et D.C. Boes, 1974. *Introduction to the Theory of Statistics*. New York, McGrah-Hill.
- Saporta, G., 1990. *Probabilités, Analyse des données et Statistique*. Paris, Technip.
- Searle, S.R., 1971. *Linear Models*. New York, John Wiley and Sons.
- Tomassone, R., S. Audrain, E. Lesquoy-de Turkheim et C. Millier, 1992. *La Régression - nouveaux regards sur une ancienne méthode statistique*. Paris, Masson, Actualités scientifiques et agronomiques de l'INRA, 188 pp.

