Capítulo 4

Estimación con restricciones

4.1. Planteamiento del problema.

En ocasiones deseamos imponer a las estimaciones de los parámetros $\vec{\beta}$ ciertas condiciones, ya para hacer el modelo interpretable ya porque así lo imponen criterios extra-estadísticos.

Nótese que no nos estamos refiriendo exclusivamente a restricciones de identificación. Puede que el conjunto de restricciones que impongamos sea tal que, junto con las ecuaciones normales, determine un único vector de estimadores $\hat{\beta}$, en un problema que previamente admitía múltiples soluciones (como sucedía en el Ejemplo 4.2). En tal caso, todo se reduce a resolver el sistema (4.3). Las restricciones se han limitado a remover la indeterminación presente en las ecuaciones normales.

En otras ocasiones, sin embargo, partimos de un modelo ya identificable (con solución única para las ecuaciones normales), pero no obstante deseamos imponer una restricción que viene dictada al margen de los datos, como ilustra el ejemplo a continuación.

Ejemplo 4.1 Si quisiéramos estimar los parámetros de una función de producción Cobb-Douglas $Q = \alpha L^{\ell} K^{\gamma}$, podríamos desear que las estimaciones de los parámetros ℓ y γ verificaran la condición $\hat{\ell} + \hat{\gamma} = 1$ (rendimientos constantes a escala). Con tres o más observaciones es perfectamente posible estimar α, ℓ y γ; la restricción es innecesaria desde el punto de vista de la estimabilidad de los parámetros. No obstante, puede formar parte de la especificación que deseamos:

no queremos ajustar cualquier función de producción Cobb-Douglas a nuestros datos, sino una con rendimientos constantes a la escala.

FIN DEL EJEMPLO ■

De un modo general, nos planteamos el problema siguiente:

$$\text{mín} \parallel \vec{y} - X\hat{\beta} \parallel^2 \qquad \text{condicionado a}: \quad A\hat{\beta} = \vec{c} \tag{4.1}$$

Está claro que no podemos esperar obtener la solución de este problema resolviendo un sistema como (4.3), que en general será incompatible.

Hay al menos dos vías para resolver un problema como el indicado. Podemos recurrir a resolver el problema de optimización condicionada (5.1) escribiendo el lagrangiano,

$$\mathcal{L}(\beta_0, \ldots, \beta_{p-1}) = \sum_{i=1}^{N}(y_i - \beta_0 x_{i0} - \ldots - \beta_{p-1}x_{i,p-1})^2 - \vec{\lambda}\,'(A\hat{\beta} - \vec{c});$$

derivando respecto a $\beta_0, \ldots, \beta_{p-1}$ y a los multiplicadores de Lagrange en el vector $\vec{\lambda}$, e igualando las derivadas a cero, obtendríamos una solución que mediante las condiciones de segundo orden podríamos comprobar que corresponde a un mínimo.

Resolveremos el problema por un procedimiento diferente, análogo al seguido con el problema incondicionado: proyectando $\vec{y}$ sobre un subespacio adecuado. Para ello habremos de transformar el problema en otro equivalente, que nos permita utilizar la técnica de la proyección. Previamente precisamos algunos resultados instrumentales, de algunos de los cuales nos serviremos repetidamente en lo que sigue.

4.2. Lemas auxiliares.

Lema 4.1 *Si $K(C)$ designa el núcleo de la aplicación lineal representada por la matriz C, se tiene:*

$$K(C) = [R(C')]^{\perp}$$

DEMOSTRACIÓN:

$$\vec{x} \in K(C) \iff C\vec{x} = \vec{0} \iff \vec{x}\,'C' = \vec{0}\,' \iff \vec{x} \perp R(C')$$

■

Lema 4.2 *Si $h \subseteq M \subseteq H$, y P_h, P_M son las matrices de proyección sobre los subespacios respectivos, se verifica: $P_M P_h = P_h P_M = P_h$*

DEMOSTRACIÓN:

Para cualquier $\vec{v} \in H$,

$$\begin{aligned} P_h\vec{v} \in h \subseteq M \quad &\Rightarrow \quad P_M P_h \vec{v} = P_h \vec{v} \\ &\Rightarrow \quad P_M P_h = P_h \end{aligned}$$

La simetría de P_M y P_h (Lema 3.4) implica entonces que: $P_h = P_h' = P_h' P_M' = P_h P_M$.

∎

Lema 4.3 *Si $h \subseteq M \subseteq H$, se tiene:*

$$P_M - P_h = P_{M \cap h^\perp}$$

DEMOSTRACIÓN:

Partimos de la identidad,

$$P_M\vec{v} = P_h\vec{v} + (P_M\vec{v} - P_h\vec{v})$$

en la que $P_h\vec{v} \in h \subseteq M$ mientras que $(P_M\vec{v} - P_h\vec{v}) \in M$. Por otra parte,

$$\begin{aligned} < P_h\vec{v}, (P_M\vec{v} - P_h\vec{v}) > \quad &= \quad \vec{v}\,' P_h (P_M\vec{v} - P_h\vec{v}) \\ &= \quad \vec{v}\,'(P_h P_M - P_h)\vec{v} \\ &= \quad 0, \end{aligned}$$

la última igualdad en virtud del Lema 5.2. Por consiguiente, $(P_M - P_h)$, que es simétrica idempotente, proyecta sobre un subespacio ortogonal a h e incluído en M; lo denotaremos mediante $M \cap h^\perp$.

∎

Lema 4.4 *Sea B una matriz cualquiera, y $K(B)$ el núcleo de la aplicación lineal que representa. Sea M un subespacio de H y $h = M \cap K(B)$. Entonces, $M \cap h^\perp = R(P_M B')$.*

La demostración puede hallarse en el Apéndice E.2, pág. 256.

4.3. Estimación condicionada.

Los Lemas anteriores proporcionan todos los elementos para obtener de forma rápida el estimador condicionado que buscamos. (Supondremos X y A de rango completo, pero es fácil generalizar el tratamiento reemplazando las inversas por inversas generalizadas.) Aunque el desarrollo formal es algo farragoso, la idea es muy simple. Vamos a transformar el modelo de modo que las restricciones $A\vec{\beta} = \vec{c}$ se conviertan en $A\vec{\beta} = \vec{0}$.

Lo haremos mediante la transformación

$$\tilde{y} = \vec{y} - X\vec{\delta} \tag{4.2}$$

$$\vec{\gamma} = \vec{\beta} - \vec{\delta}, \tag{4.3}$$

siendo $\vec{\delta}$ una solución cualquiera de $A\vec{\delta} = \vec{c}$ (de no existir tal solución, no tendría sentido el problema; estaríamos imponiendo condiciones a los parámetros imposibles de satisfacer). Se tiene entonces que:

$$\begin{aligned} \vec{y} &= X\vec{\beta} + \vec{\epsilon} \Longrightarrow \vec{y} - X\vec{\delta} = X\vec{\beta} - X\vec{\delta} + \vec{\epsilon} \Longrightarrow \tilde{y} = X\vec{\gamma} + \vec{\epsilon} \\ A\vec{\beta} &= \vec{c} \Longrightarrow A(\vec{\gamma} + \vec{\delta}) = \vec{c} \Longrightarrow A\vec{\gamma} = \vec{c} - A\vec{\delta} \Longrightarrow A\vec{\gamma} = \vec{0} \end{aligned}$$

y el problema original (5.1) puede ahora reescribirse así:

$$\text{mín} \parallel \tilde{y} - X\hat{\gamma} \parallel^2 \qquad \text{condicionado a} \quad A\hat{\gamma} = \vec{0},$$

o, alternativamente,

$$\text{mín} \parallel \tilde{y} - X\hat{\gamma} \parallel^2 \qquad \text{condicionado a :} \quad A(X'X)^{-1}X'(X\hat{\gamma}) = \vec{0}. \tag{4.4}$$

¿Qué ventajas presenta la expresión (5.4) del problema comparada con la original? Una importante: muestra que el $X\hat{\gamma}$ buscado no es sino la proyección de $\tilde{y}$ sobre un cierto subespacio: $h = M \cap K(A(X'X)^{-1}X')$. *Hay garantía de que h es un subespacio porque M y $K(A(X'X)^{-1}X')$ lo son.* Basta proyectar $\tilde{y}$ sobre h para obtener $X\hat{\gamma}$ y, si X es de rango completo, $\hat{\gamma}$; y esta proyección se puede obtener fácilmente con ayuda de los Lemas anteriores.

Si denotamos por $\hat{\gamma}_h$ las estimaciones mínimo cuadráticas condicionadas o restringidas por $A\hat{\gamma} = \vec{0}$, tenemos que:

$$X\hat{\gamma}_h = P_h\tilde{y} \tag{4.5}$$

$$= (P_M - P_{M\cap h^\perp})\tilde{y} \tag{4.6}$$

$$= [X(X'X)^{-1}X' - P_{M\cap h^\perp}]\tilde{y} \tag{4.7}$$

en que el paso de (5.5) a (5.6) ha hecho uso del Lema 5.3. Pero es que, de acuerdo con el Lema 5.4,

$$M \cap h^{\perp} = R[\underbrace{X(X'X)^{-1}X'}_{P_M}\underbrace{X(X'X)^{-1}A'}_{B'}] = R[\underbrace{X(X'X)^{-1}A'}_{Z}]$$

Por consiguiente, $P_{M\cap h^{\perp}}$ es, de acuerdo con el Lema 3.9, pág. 37,

$$P_{M\cap h^{\perp}} = Z(Z'Z)^{-1}Z', \tag{4.8}$$

ecuación que, llevada a (5.7), proporciona:

$$\begin{aligned} X\hat{\gamma}_h &= X(X'X)^{-1}X'\tilde{y} - X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}X'\tilde{y} \\ &= X\hat{\gamma} - X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A\hat{\gamma}, \end{aligned} \tag{4.9}$$

en que $\hat{\gamma}$ es el vector de estimadores mínimo-cuadráticos ordinarios al regresar $\tilde{y}$ sobre X. Si X es de rango total, como venimos suponiendo, de (5.9) se deduce:

$$\hat{\gamma}_h = \hat{\gamma} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A\hat{\gamma}. \tag{4.10}$$

(véase el Ejercicio 5.3.)

Hay algunas observaciones interesantes que hacer sobre las ecuaciones (5.9) y (5.10). En primer lugar, el lado izquierdo de (5.9) es una proyección. Ello garantiza de manera automática que $\| \tilde{y} - X\hat{\gamma}_h \|^2$ es mínimo[1]. Además, el tratamiento anterior se generaliza de modo inmediato al caso de modelos de rango no completo, sin más que reemplazar en los lugares procedentes matrices inversas por las correspondientes inversas generalizadas.

En segundo lugar, dado que los estimadores mínimo cuadráticos ordinarios estiman insesgadamente los correspondientes parámetros, tomando valor medio en (5.10) vemos que:

$$E[\hat{\gamma}_h] = \vec{\gamma} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A\vec{\gamma}$$

lo que muestra que $\hat{\gamma}_h$ es un estimador insesgado de $\vec{\gamma}$ si $A\vec{\gamma} = \vec{0}$. Es decir, la insesgadez se mantiene si los parámetros *realmente* verifican las condiciones impuestas sobre los estimadores.

[1] Si hubiéramos llegado al mismo resultado minimizando una suma de cuadrados por el procedimiento habitual (derivando un lagrangiano) tendríamos aún que mostrar que el punto estacionario encontrado es un mínimo y no un máximo.

En tercer lugar, si definimos: $G = (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A$ tenemos que: $\hat{\gamma}_h = (I - G)\hat{\gamma}$. Por consiguiente,

$$
\begin{aligned}
\Sigma_{\hat{\gamma}_h} &= (I - G)\Sigma_{\hat{\gamma}}(I - G') \\
&= (I - G)\sigma^2(X'X)^{-1}(I - G') \\
&= \sigma^2[(X'X)^{-1} - G(X'X)^{-1} - (X'X)^{-1}G' + G(X'X)^{-1}G'] \\
&= \sigma^2[(X'X)^{-1} - G(X'X)^{-1}G']
\end{aligned}
$$

que muestra, dado que el segundo sumando tiene claramente elementos no negativos en su diagonal principal (la matriz $(X'X)^{-1}$ es definida no negativa), que $\Sigma_{\hat{\gamma}_h}$ tiene en la diagonal principal varianzas no mayores que las correspondientes en $\Sigma_{\hat{\gamma}}$. Podemos concluir, pues, que *la imposición de restricciones lineales sobre el vector de estimadores nunca incrementa su varianza*, aunque eventualmente, si las restricciones impuestas no son verificadas por los parametros a estimar, *puede introducir algún sesgo.*

Hemos razonado en las líneas anteriores sobre el modelo transformado. Podemos sustituir sin embargo (5.3) en (5.10) y obtener la expresión equivalente en términos de los parámetros originales:

$$\hat{\beta}_h = \hat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c}) \tag{4.11}$$

R: Ejemplo 4.1 *(estimación condicionada)*

No hay en R una función de propósito general para realizar estimación condicionada. La extensibilidad del lenguaje hace sin embargo extraordinariamente fácil el definirla. El fragmento a continuación ilustra el modo de hacerlo y como utilizarla. No se ha buscado la eficiencia ni elegancia sino la correspondencia más directa con la teoría expuesta más arriba.

Definimos en primer lugar una función para uso posterior:

```
> lscond <- function(X, y, A, d, beta0 = TRUE) {
+     ajuste <- lsfit(X, y, intercept = beta0)
+     betas <- ajuste$coefficients
+     xxinv <- solve(t(X) %*% X)
+     axxa <- solve(A %*% xxinv %*% t(A))
+     betas.h <- betas - xxinv %*% t(A) %*%
+         axxa %*% (A %*% betas - d)
+     betas.h <- as.vector(betas.h)
+     names(betas.h) <- names(ajuste$coefficients)
+     return(list(betas = betas, betas.h = betas.h,
+         ajuste.inc = ajuste))
+ }
```

Generamos a continuación los datos y realizamos la estimación ciñéndonos a la teoría del modo más directo. `X` es la matriz de diseño, `beta` contiene los parámetros e `y` la variable respuesta:

```
> X <- matrix(c(1, 1, 1, 1, 1, 1, 1, 4,
+     12, 1, 4, 13, 0, 6, 7, 0, 2, 2), 6,
+     3)
> X

     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    4    6
[3,]    1   12    7
[4,]    1    1    0
[5,]    1    4    2
[6,]    1   13    2

> beta <- c(2, 3, 4)
> y <- X %*% beta + rnorm(6)
```

Especificamos la restricción lineal $\beta_1 = \beta_2$ tomando la matriz A y vector d siguientes:

```
> A <- matrix(c(0, 1, -1), 1, 3, byrow = TRUE)
> d <- 0
```

y a continuación realizamos la estimación condicionada:

```
> resultado <- lscond(X, y, A = A, d = d,
+     beta0 = FALSE)
> resultado$betas.h

    X1     X2     X3
2.8392 3.2647 3.2647

> resultado$betas

    X1     X2     X3
2.8037 3.0526 3.7138
```

Fin del ejemplo ■

Complementos y ejercicios

4.1 Sea un espacio vectorial M cualquiera, de dimensión finita. Compruébese que *siempre* existe una matriz C tal que $M = K(C)$. (Ayuda: considérese una matriz cuyas filas fueran una base de $M^{\perp}$).

4.2 ($\uparrow$ 5.1) Pruébese la igualdad (E.15), pág. 256.

4.3 Justifíquese el paso de (5.9) a (5.10).

4.4 El Ejemplo 5.1 *se sale* del marco conceptual en el que nos movemos. Los regresores (K y L, ó $\log(K)$ y $\log(L)$ al linealizar la función de producción) no pueden ser fijados por el experimentador: dependen de los agentes económicos. Estamos ante *datos observados* en oposición a *datos experimentales*. Faraway (2005), Sec. 3.8, contiene una diáfana discusión de los problemas que ello conlleva. Es también interesante, aunque de más difícil lectura, Wang (1993).

4.5 Las restricciones que hemos discutido en la Sección 5.3 son exactas. Los parámetros las verifican de modo exacto. En ocasiones se recurre a restricciones estocásticas, llevando a los parámetros a verificarlas de forma *aproximada*. Es muy fácil introducirlas. Recordemos que, al hacer estimación mínimo-cuadrática, los parámetros se fijan de modo que la suma de cuadrados de los residuos sea la mínima posible. Si tenemos restricciones $A\vec{\beta} = \vec{c}$ que queremos imponer de modo aproximado basta que añadamos las filas de A a la matriz X y los elementos correspondientes de $\vec{c}$ al vector $\vec{y}$ para obtener:

$$\begin{pmatrix} \vec{y} \\ \vec{c} \end{pmatrix} = \begin{pmatrix} X \\ A \end{pmatrix} \vec{\beta} + \vec{\epsilon}$$

y hagamos mínimos cuadrados ordinarios con la muestra ampliada (las filas añadidas se denominan en ocasiones *pseudo-observaciones*). La idea es que las filas añadidas funcionan como observaciones y, por tanto, el procedimiento de estimación tenderá a hacer $A\hat{\beta} \approx \vec{c}$ (para que los residuos correspondientes $\vec{c} - A\hat{\beta}$ sean "pequeños"). Aún más: podemos graduar la importancia que damos a las pseudo-observaciones (y por tanto el nivel de aproximación con que deseamos imponer las restricciones estocásticas): basta que las multipliquemos por una constante adecuada k para estimar

$$\begin{pmatrix} \vec{y} \\ k\vec{c} \end{pmatrix} = \begin{pmatrix} X \\ kA \end{pmatrix} \vec{\beta} + \vec{\epsilon}. \tag{4.12}$$

Obsérvese que ahora los residuos de las pseudo-observaciones serán $k(\vec{c} - A\hat{\beta})$ y si tomamos k elevado el método mínimo cuadrático tendrá que prestar atención preferente a que $A\hat{\beta} \approx \vec{c}$ se verifique con gran aproximación (porque los cuadrados de los residuos correspondientes entran en SSE afectados de un coeficiente k^2). Cuando $k \to \infty$ nos acercamos al efecto de restricciones exactas.

4.6 ($\uparrow$ 5.5) Un caso particular de interés se presenta cuando en el problema anterior se toma $A = I$ y $\vec{c} = \vec{0}$. Se dice entonces que estamos ante el estimador *ridge* de parámetro k. En 11.3, pág. 144, abordamos su estudio y justificación con detalle.

4.7 ($\uparrow$ 5.5) La estimación de (5.12) haciendo uso de las ecuaciones normales proporciona

$$\hat{\beta} = (X'X + k^2A'A)^{-1}(X'\vec{y} + k^2A'\vec{c}), \tag{4.13}$$

que admite una interpretación bayesiana. Supongamos que *a priori* $\vec{\beta} \sim N(\vec{\beta}_0, \Sigma_0)$. Dado $\vec{\beta}$, $\vec{Y}$ se distribuye como $N(X\vec{\beta}, \sigma^2 I)$. La densidad *a posteriori* de $\vec{\beta}$ es entonces

$$\begin{aligned} f(\vec{\beta}|\vec{y}, \sigma^2, \vec{\beta}_0, \Sigma_0) &\propto \exp\left\{-\frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})'(\vec{y} - X\vec{\beta})\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}(\vec{\beta} - \vec{\beta}_0)'\Sigma_0^{-1}(\vec{\beta} - \vec{\beta}_0)\right\} \\ &= \exp\Bigg\{ - \frac{1}{2\sigma^2}\Big[(\vec{y} - X\vec{\beta})'(\vec{y} - X\vec{\beta}) \\ &\qquad + \sigma^2(\vec{\beta} - \vec{\beta}_0)'\Sigma_0^{-1}(\vec{\beta} - \vec{\beta}_0)\Big]\Bigg\} \end{aligned}$$

Tomando el logaritmo neperiano e igualando a cero su derivada respecto a $\vec{\beta}$ tenemos entonces

$$-\frac{1}{2\sigma^2}\left[(-2X'(\vec{y} - \vec{X}\vec{\beta}) + 2\sigma^2\Sigma_0^{-1}(\vec{\beta} - \vec{\beta}_0)\right] = \vec{0},$$

que proporciona

$$(X'X + \sigma^2\Sigma_0^{-1})\vec{\beta} - X'\vec{y} - \sigma^2\Sigma_0^{-1}\vec{\beta}_0 = \vec{0},$$

y por tanto la moda de la distribución *a posteriori* (que fácilmente se comprueba es normal multivariante) es:

$$\hat{\beta} = (X'X + \sigma^2\Sigma_0^{-1})^{-1}(X'\vec{y} + \sigma^2\Sigma_0^{-1}\vec{\beta}_0). \tag{4.14}$$

Comparando (5.14) con (5.13) vemos que son idénticas cuando $kA = \sigma\Sigma_0^{-\frac{1}{2}}$ y $k\vec{c} = \sigma\Sigma_0^{-\frac{1}{2}}\vec{\beta}_0$: para obtener el estimador bayesiano con información *a priori* como la indicada, basta por tanto con obtener el estimador MCO en una muestra ampliada con pseudo-observaciones.