1. Modelo de Regresión Lineal

Estamos proponiendo estimar el siguiente modelo:

$$Y_i = X_i'\beta + u_i$$

$$= \underbrace{\sum_{j=1}^{k} \beta_j x_{ij}}_{k \text{ potenciales regresores}} + u_i, \qquad (1)$$

donde estos regresores son exógenos en el sentido de $E[X_{ij} u_i] = 0$, $\forall j = 1, \dots, k$. Además, por el momento supondremos que k es fijo y no depende de n. Noten que esto implica que $k/n \to 0$ cuando $n \to \infty$.

Denotamos $\mathcal{A}$ como la lista de regresores relevantes (coeficientes distintos de cero)

$$\mathcal{A} = \{ j : \beta_j \neq 0 \}.$$

Ejemplo: $\mathcal{A} = \{1, 4\}$ dice que solo $X_{i,1}$ y $X_{i,4}$ son relevantes.

Denotamos $\mathcal{A}_0$ como el verdadero set de regresores relevantes

¡Verdadero modelo! $\Rightarrow$ $$Y_i = \sum_{j \in \mathcal{A}_0} \beta_j X_{i,j} + u_i$$

Nuestra meta es estimar $\mathcal{A}_0$ con data $\{(Y_i, X_i')' : i = 1, \dots, n\}$. Vamos a estimar el set de regresores relevantes con un proceso de selección. Decimos que es un proceso de selección consistente si

$$(2) \quad P(\hat{\mathcal{A}}_n = \mathcal{A}_0) \longrightarrow 1 \quad \text{cuando } n \to \infty.$$

Asimismo $\beta_{\mathcal{A}}$ es el subvector de β que solo incluye los coeficientes de $\mathcal{A}$:

$$\beta_{\mathcal{A}} = (\beta_j : j \in \mathcal{A}).$$

$|\mathcal{A}|$: # de elementos en $\mathcal{A}$.

Con el procedimiento de selección $\hat{\mathcal{A}}_n$ que produce $\hat{\beta}_n$ donde $\hat{\beta}_{n,j} = 0$ para $j \notin \hat{\mathcal{A}}_n$. Decimos que el proceso es oráculo si

$$\sqrt{n}\,(\hat{\beta}_{\mathcal{A}_0} - \beta_{\mathcal{A}_0}) \xrightarrow{d} \mathcal{N}(0, V(\mathcal{A}_0)),$$

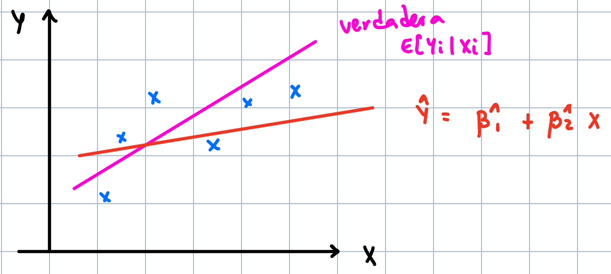
donde $V(\mathcal{A}_0)$ es la varianza asintótica cuando $\mathcal{A}_0$ es conocido.

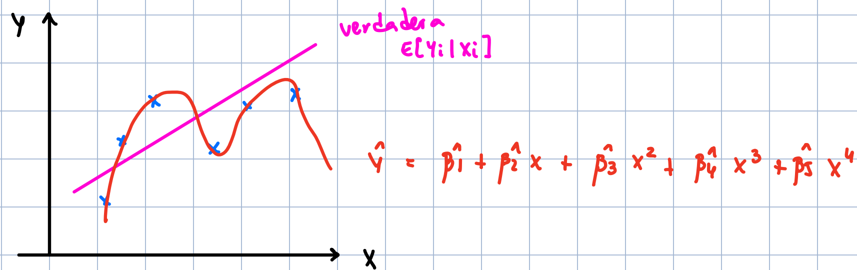
* Oráculo significa que es como si en muestras grandes hubiéramos usado el verdadero modelo $\mathcal{A}_0$.

1.1. Penalización

Recuerden que $\hat{\beta}_n$ resuelve $\min_b \; SSR := \sum_{i=1}^{n} (Y_i - X_i' b)^2$ y siempre que aumentamos regresores podemos disminuirlo.

Sin embargo, caemos en el problema de que hacemos overfitting





Consideremos que el verdadero proceso generador de datos es:

$$Y_i = \sum_{j \in A_0} \beta_j X_{ij} + u_i$$

$$= X_{i,A_0}' \beta_{A_0} + u_i$$

Nosotros proponemos un modelo A tal que estimamos

$$\hat{\beta}_{n,A}(A) = \left(\sum_{i=1}^{n} X_{i,A} X_{i,A}' \right)^{-1} \sum_{i=1}^{n} X_{i,A} Y_i ,$$

$$\hat{\beta}_{n,A^c}(A) = 0$$

Ejemplo: $A = \{1, 3\}$ $\Rightarrow$ $A^c = \{2, 4\}$ $\Rightarrow$ $\hat{\beta_1} \neq 0$, $\hat{\beta_3} \neq 0$
$k = 4$ $\qquad\qquad\qquad\qquad$ $\hat{\beta_2} = 0$, $\hat{\beta_4} = 0$

Además, bajo A definimos

$$SSR_n(A) = \sum_{i=1}^{n} \left(Y_i - X_{i,A}' \hat{\beta}_{n,A}(A) \right)^2$$

Una forma razonable de penalizar es usando el número de regresores como un proxy de complejidad

$$\underbrace{BIC_n(A)}_{\approx \text{ lo que llamamos Bayesian Information Criterion}} := SSR_n(A) + \underbrace{|A| \log n}_{\text{penalización}}$$

Entonces

$$\hat{A}_n^{BIC} = \underset{A}{\operatorname{argmin}} \; BIC_n(A)$$

Teorema .- Si $E X_i X_i'$ tiene rango completo y $E u_i^2 X_i X_i' = O(1)$ y es p.d.

Entonces

$$P(\hat{A_n} = A_0) \to 1 \quad \text{cuando} \quad n \to \infty .$$

* Es decir, la selección es consistente.

proof:

Debemos mostrar que para todo $A \neq A_0$

$$P(\underbrace{SSR_n(A) + |A| \log n}_{BIC_n(A)} > \underbrace{SSR_n(A_0) + |A_0| \log n}_{BIC_n(A_0)}) \to 1$$

(Esto significa que A_0 en muestras grandes será lo que minimiza BIC_n con probabilidad alta)

Empezamos mostrando lo siguiente:

- $$\frac{SSR_n(A_0)}{n} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_{i,A_0}' \hat{\beta}_{n,A_0}(A_0))^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_{i,A_0}' \hat{\beta}_{n,A_0}(A_0) \pm X_{i,A_0}' \beta_{A_0})^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (u_i - X_{i,A_0}' (\hat{\beta}_{n,A_0}(A_0) - \beta_{A_0}))^2$$

$$= \underbrace{\frac{1}{n} \sum_{i=1}^{n} u_i^2}_{E u_i^2 + o_p(1)} - 2 \underbrace{\frac{1}{n} \sum_{i=1}^{n} u_i X_i' (\hat{\beta}_{n,A_0}(A_0) - \beta_{A_0})}_{\leq E \|u_i X_i\| \cdot \|\hat{\beta}_{n,A_0}(A_0) - \beta_{A_0}\| \; = 0 \cdot o_p(1)} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} [X_{i,A_0}' (\hat{\beta}_{n,A_0}(A_0) - \beta_{A_0})]^2}_{\leq E\|X_i X_i'\| \cdot \|\hat{\beta}_{n,A_0}(A_0) - \beta_{A_0}\| \; = O(1) \cdot o_p(1) = o_p(1)}$$

$$= E u_i^2 + o_p(1)$$

- $$\frac{1}{n} SSR_n(A) = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_{i,A}' \hat\beta_{n,A}(A))^2$$

s.a.

$(A \cap A_0) \neq A_0$

(omitimos regresores relevantes)

$$= \frac{1}{n}\sum_{i=1}^{n} u_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left[x_{i,A}'(\hat\beta_{n,A}(A) - \beta_A)\right]^2 - 2\,\frac{1}{n}\sum_{i=1}^{n} u_i x_i'(\hat\beta_{n,A}(A) - \beta_A)$$

* Excluir regresores relevantes implica $\hat\beta_n(A) - \beta \xrightarrow{p} \delta \neq 0$

$$= E u_i^2 + \delta' E x_i x_i' \delta + o_p(1).$$

- $$SSR_n(A) = \sum_{i=1}^{n} u_i^2 + \sum_{i=1}^{n}\left[x_{i,A}'(\hat\beta_{n,A}(A) - \beta_A)\right]^2 - 2\sum_{i=1}^{n} u_i x_i'(\hat\beta_{n,A}(A) - \beta_A)$$

s.a

$A_0 \subset A$

(contiene regresores irrelevantes)

Noten que no puse $1/n$

$$= \sum_{i=1}^{n} u_i^2 - 2\left(\frac{1}{\sqrt{n}}\sum x_{i,A} u_i\right)'\sqrt{n}(\hat\beta_{n,A} - \beta_A)$$

$$+ \sqrt{n}(\hat\beta_{n,A} - \beta_A)'\left(\sum_{i=1}^{n}\frac{x_{i,A} x_{i,A}'}{n}\right)\sqrt{n}(\hat\beta_{n,A} - \beta_A)$$

$$= \underbrace{\sum_{i=1}^{n} u_i^2}_{\geq 0} + O_p(1) + o_p(1)$$

Entonces

CASO 1 :

$$P(BIC_n(A) > BIC_n(A_0)) = P\left(\frac{SSR_n(A)}{n} + |A|\frac{\log n}{n} > \frac{SSR_n(A_0)}{n} + |A_0|\frac{\log n}{n}\right)$$

s.a.

$A \cap A_0 \neq A_0$

$$= P\left(E u_i^2 + \delta' E x_i x_i' \delta + o_p(1) + |A|\frac{\log n}{n} > E u_i^2 + |A_0|\frac{\log n}{n} + o_p(1)\right)$$

$$= P\left(\underbrace{(|A| - |A_0|)}_{\geq 0}\underbrace{\frac{\log n}{n}}_{=o(1)} + o_p(1) + \underbrace{\delta' E x_i x_i' \delta}_{>0} > 0\right)$$

$$\to 1.$$

CASO 2 :

$$P(BIC_n(A) > BIC_n(A_0)) = P(SSR_n(A) + |A|\log n > SSR_n(A_0) + |A_0|\log n)$$

s.a.

$A_0 \subset A$

$$= P\left(\sum_{i=1}^{n} u_i^2 + O_p(1) + o_p(1) + |A|\log n > \sum_{i=1}^{n} u_i^2 + O_p(1) + o_p(1) + |A_0|\log n\right)$$

$$= P\left(O_p(1) + o_p(1) > \underbrace{(|A_0| - |A|)}_{<0}\log n\right) \to 1.$$

¿Qué pasa si la penalización fuese algo que no crece con n?

$$AIC_n(A) := SSR_n(A) + 2|A|$$

Entonces

CASO 1: $P(\ AIC_n(A) > AIC_n(A_0)\ ) = P\left(\frac{SSR_n(A)}{n} + |A|\frac{2}{n} > \frac{SSR_n(A_0)}{n} + |A_0|\frac{2}{n}\right)$

s.a.
$A \cap A_0 \neq A_0$

$$= P\left(E u_i^2 + \delta' E x_i x_i' \delta + o_p(1) + |A|\frac{2}{n} > E u_i^2 + |A_0|\frac{2}{n} + o_p(1)\right)$$

$$= P\left(\underbrace{(|A| - |A_0|)}_{\geq 0}\underbrace{\frac{2}{n}}_{=o(1)} + o_p(1) + \underbrace{\delta' E x_i x_i' \delta}_{>0} > 0 \right)$$

$$\to 1.$$

CASO 2: $P(\ BIC_n(A) > BIC_n(A_0)\ ) = P\left(SSR_n(A) + |A|\log n > SSR_n(A_0) + |A_0|\log n\right)$

s.a.
$A_0 \subset A$

$$= P\left(\sum_{i=1}^{n} u_i^2 + O_p(1) + o_p(1) + |A|\,2 > \sum_{i=1}^{n} u_i^2 + O_p(1) + o_p(1) + |A_0|\,2 \right)$$

$$= P\left(O_p(1) + o_p(1) > \underbrace{(|A_0| - |A|)}_{<0}\,2 \right) \not\to 1.$$

En conclusión AIC puede seleccionar correctamente los regresores relevantes (caso 1) pero en muestras grandes puede seguir seleccionando regresores irrelevantes (más conservador).

1.2. Inferencia Post Selección

Si tenemos un proceso de selección consistente

$$P(\hat{A}_n = A_0) \to 1 \quad , \quad \text{(por ejemplo con BIC)}$$

entonces

$$P\left(\sqrt{n}\left(\hat{\beta}_{n_j}(\hat{A}_n) - \beta_j\right) \leq u\right) = P\left(\sqrt{n}\left(\hat{\beta}_{n_j}(A_0) - \beta_j\right) \leq u\right) + o(1)$$

proof:

$$P(\sqrt{n}\,(\hat{\beta}_{nj}(\hat{A_n}) - \beta_j) \le u) = P(\sqrt{n}\,(\hat{\beta}_{nj}(\hat{A_n}) - \beta_j) \le u,\ \hat{A_n} = A_0)$$

$$+$$

$$P(\sqrt{n}\,(\hat{\beta}_{nj}(\hat{A_n}) - \beta_j) \le u,\ \hat{A_n} \neq A_0)$$

$$\le P(\sqrt{n}\,(\hat{\beta}_{nj}(\hat{A_n}) - \beta_j) \le u,\ \hat{A_n} = A_0)$$

$$+ \quad P(\hat{A_n} \neq A_0)$$

$$= P(\sqrt{n}\,(\hat{\beta}_{nj}(\hat{A_n}) - \beta_j) \le u,\ \hat{A_n} = A_0) \quad + o(1)$$

$$= P(\sqrt{n}\,(\hat{\beta}_{nj}(\hat{A_n}) - \beta_j) \le u \mid \hat{A_n} = A_0)\, P(\hat{A_n} = A_0) + o(1)$$

$$= P(\sqrt{n}\,(\hat{\beta}_{nj}(A_0) - \beta_j) \le u)\ (1 + o(1)) + o(1)$$

$$= P(\sqrt{n}\,(\hat{\beta}_{nj}(A_0) - \beta_j) \le u) + o(1) \quad . \quad \blacksquare$$

Por lo tanto, todo procedimiento que seleccione consistentemente los regresores es también uno que cumple la propiedad de ser oráculo.

---

* Paréntesis : Delta Method

La prueba anterior ha mostrado una herramienta útil de como probar distintos teoremas cuando hay consistencia.

Teorema.- (Método Delta) Sea $\hat{\beta_n}$ consistente y $h(\cdot)$ una función continua en β_0 (verdadero valor). Entonces, si $Var(x_i u_i) = O(1)$ tenemos

$$\sqrt{n}\,(\hat{\beta_n} - \beta_0) \xrightarrow{d} N(0, \underbrace{(E x_i x_i')^{-1} E[u_i^2 x_i x_i'](E x_i x_i')'}_{V})$$

$$\sqrt{n}\,(h(\hat{\beta_n}) - h(\beta_0)) \xrightarrow{d} N(0,\ \underset{1\times k}{\nabla h(\beta_0)^T}\ \underset{k\times k}{V}\ \underset{k\times 1}{\nabla h(\beta_0)})$$

proof:

$$h(\hat{\beta_n}) = h(\beta_0) + \nabla h(\beta_n^*)\,(\hat{\beta_n} - \beta_0)$$

* Mean value expansion
$\beta_0 \le \beta_n^* \le \hat{\beta}_n$

Tenemos que

$$P\left(\|\nabla h(\beta_n^*) - \nabla h(\beta_0)\| > \varepsilon\right) = P\left(\|\nabla h(\beta_n^*) - \nabla h(\beta_0)\| > \varepsilon,\ \|\beta_n^* - \beta_0\| > \delta\right)$$

$$+ P\left(\|\nabla h(\beta_n^*) - \nabla h(\beta_0)\| > \varepsilon,\ \|\beta_n^* - \beta_0\| \leq \delta\right)$$

$$\leq P\left(\|\beta_n^* - \beta_0\| > \delta\right)$$

$$+ P\left(\|\nabla h(\beta_n^*) - \nabla h(\beta_0)\| > \varepsilon,\ \|\beta_n^* - \beta_0\| \leq \delta\right)$$

$$\leq \underbrace{P\left(\|\hat{\beta}_n - \beta_0\| > \delta\right)}_{o(1)}$$

$$+ P\left(\|\nabla h(\beta_n^*) - \nabla h(\beta_0)\| > \varepsilon,\ \|\beta_n^* - \beta_0\| \leq \delta\right)$$

$$\underset{\substack{\text{def}\\ \text{de}\\ \text{continuidad}}}{=} P\left(\frac{\varepsilon}{2} > \varepsilon\right) + o(1)$$ si elegimos un δ suficientemente pequeño.

$$= o(1).$$

Entonces,

$$h(\hat{\beta}_n) = h(\beta_0) + \nabla h(\beta_n^*)(\hat{\beta}_n - \beta_0)$$

$$= h(\beta_0) + \left[\nabla h(\beta_0) + o_p(1)\right](\hat{\beta}_n - \beta_0)$$

Reordenando y multiplicando por $\sqrt{n}$

$$\sqrt{n}\left(h(\hat{\beta}_n) - h(\beta_0)\right) = \left[\nabla h(\beta_0) + o_p(1)\right]\sqrt{n}(\hat{\beta}_n - \beta_0)$$

$$= \nabla h(\beta_0)\sqrt{n}(\hat{\beta}_n - \beta_0) + o_p(1)\,O_p(1)$$

$$= \nabla h(\beta_0)\sqrt{n}(\hat{\beta}_n - \beta_0) + o_p(1)$$

$$\xrightarrow{d} \mathcal{N}\Big(0,\ \underbrace{\nabla h(\beta_0)^T}_{1\times k}\ \underbrace{V}_{k\times k}\ \underbrace{\nabla h(\beta_0)}_{k\times 1}\Big). \ \blacksquare$$