Regresión Lineal Múltiple - Semana 09

Raúl Alberto Pérez
raperez1@unal.edu.co
Profesor Asociado - Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Semestre 2021-02

Multicolinealidad

Comparación de efectos parciales de las variables predictoras

Considere el modelo de RLM:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \text{ con } \varepsilon_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right).$$

1. Si las variables predictoras no están en una misma escala de medida, no podemos determinar cual tiene mayor o menor efecto parcial sobre la respuesta promedio, en presencia de las demás, esto es, la magnitud de β_j refleja las unidades de la variable X_j.
2. Para hacer comparaciones en forma directa de los coeficientes de regresión se recurre al uso de variables escaladas, tanto en la respuesta como en las predictoras.

Escalamiento de longitud unitaria: cada variable es escalada restando su media muestral y dividiendo esta diferencia por la raíz cuadrada de la suma de cuadrados corregida de cada variable, es decir,

$$Y_i^* = \frac{Y_i - \bar{Y}}{\sqrt{\sum_{h=1}^{n} \left(Y_h - \bar{Y}\right)^2}}, \quad X_{ij}^* = \frac{X_{ij} - \bar{X}_j}{\sqrt{\sum_{h=1}^{n} \left(X_{hj} - \bar{X}_j\right)^2}},$$

$$\text{con} \begin{cases} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, k \end{cases}$$

Luego, se ajusta el modelo de RLM sin intercepto

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \cdots + \beta_k^* X_{ik}^* + \varepsilon_i, \text{ con } \varepsilon_i \overset{\text{i.i.d}}{\sim} N\left(0, \sigma^2\right)$$

Los coeficientes β_j^* son llamados *coeficientes de regresión estandarizados*, y estos pueden ser comparados directamente teniendo en cuenta que siguen siendo coeficientes de regresión parcial, es decir, β_j^* mide el efecto de X_j^* dado que las demás variables predictoras están en el modelo.

Además, los β_j^* pueden servir para determinar la importancia relativa de X_j^* en presencia de las demás variables, en la muestra o conjunto de datos particular considerado para el ajuste.

Multicolinealidad

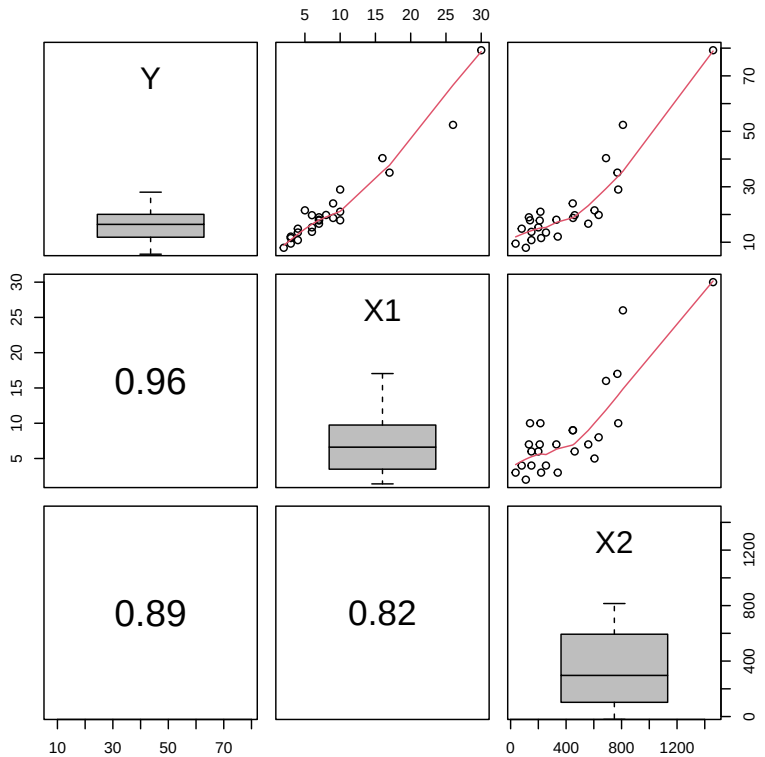
Multicolinealidad es la existencia de dependencia casi lineal entre variables predictoras en el modelo de RLM.

Si existiera dependencia lineal exacta entre dos o más variables predictoras, la matriz $\mathbf{X}'\mathbf{X}$ sería singular y por tanto no podríamos hallar los estimadores de mínimos cuadrados!.

Causas de la multicolinealidad

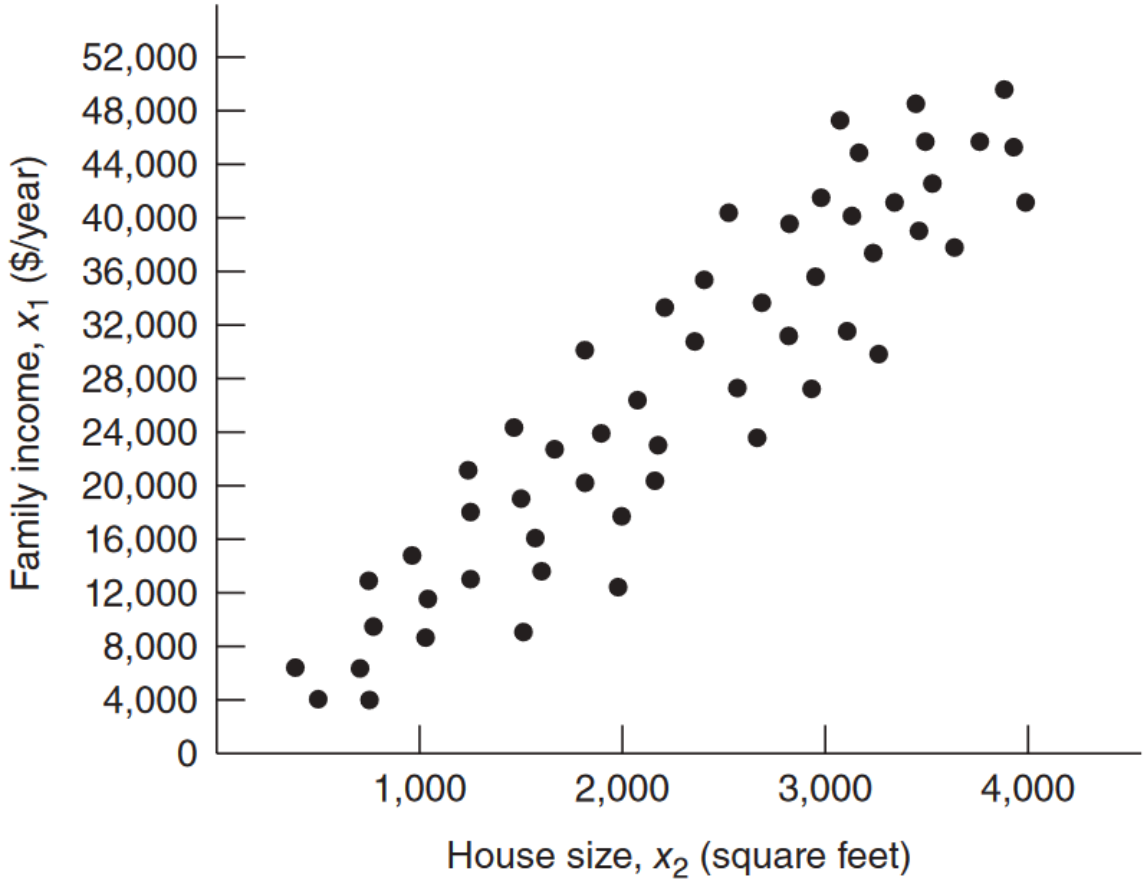
Algunas de las causas más comunes de la multicolinealidad son:

1. **El método de recolección de los datos.**



Datos de tiempos de entrega (*Fuente: Mongomery et al. (2002)*)

2. **Restricciones en el modelo o en la población.**



Datos de consumo de electricidad

(*Fuente: Mongomery et al. (2002)*)

3. **Especificación del modelo.** Por ejemplo, al agregar términos polinomiales a un modelo cuando el rango de la predictora es pequeño.

4. **Un modelo sobredefinido.** Por ejemplo, en investigación médica donde para algunos pocos pacientes se toma una gran cantidad de predictoras.

Efectos de la multicolinealidad

Algunos de los efectos más notorios de la multicolinealidad son:

1. **Inflación de las varianzas de los estimadores:** consiste en la inflación de los valores c_{jj}^* en las varianzas de los estimadores $Var(\widehat{\beta}_j^*) = \sigma^2 c_{jj}^*$, cuando se considera un modelo con variables escaladas de longitud unitaria, en cuyo caso se puede demostrar que:

$$c_{jj}^* = \frac{1}{1 - R_j^2},$$

donde R_j^2 es el coeficiente de determinación muestral obtenido de una regresión de X_j (como respuesta) en función de las otras variables predictoras consideradas en el modelo (actuando como predictoras de la primera).

2. $\widehat{\beta}_j$ **muy grandes en términos absolutos:** esto se manifiesta en una traza muy grande de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$, donde: $\text{traza}\,(\mathbf{X}'\mathbf{X})^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j}$, $\lambda_j > 0$ es el j-ésimo valor propio (usualmente ordenados de mayor a menor) de la matriz $\mathbf{X}'\mathbf{X}$.

   Si la $\text{traza}\,(\mathbf{X}'\mathbf{X})^{-1}$ es muy grande, mayor es la distancia entre el vector de parámetros estimados y el verdadero valor del vector de parámetros.

3. **Valores de los coeficientes con signo contrario a lo esperado:** esto puede ser causado por la presencia de multicolinealidad.

4. **Regresión significativa pero ninguna variable individualmente significativa:** otra de las maneras en que se puede manifestar la multicolinealidad grave es cuando el modelo de regresión ajustado es significativo (según la prueba F de la tabla ANOVA) pero individualmente, ninguno de los coeficientes asociados a las variables predictoras resulta significativo (según las pruebas T de significancia individual).

Diagnósticos de la multicolinealidad

Entre los diagnósticos más usados para detectar multicolinealidad en un modelo se tienen:

1. **Examinar la matriz de correlaciones entre las predictoras:** sea CORR($\mathbf{X}$) una matriz cuyo elemento (j, k) corresponde a la correlación entre las predictoras X_j y X_k, CORR(X_j, X_k).

   Esta matriz resulta útil para detectar multicolinealidad si en ésta no intervienen más de dos variables en una dependencia casi lineal.

   También valores de CORR(X_j, X_k) pequeños no necesariamente implican la ausencia de multicolinealidad.

2. **Factores de Inflación de Varianza:** denotados $\text{VIF}_j,\ j = 1, \ldots, k$ se calculan como: $\text{VIF}_j = c^*_{jj} = \frac{1}{1-R^2_j}$.

   A continuación se establece el criterio para detectar la multicolinealidad de acuerdo a esta medida.

   - Si $\text{VIF}_j \leq 5$ no hay problemas de multicolinealidad.
   - Si $5 < \text{VIF}_j \leq 10$ hay problemas de multicolinealidad moderada.
   - Si $\text{VIF}_j > 10$ hay problemas de multicolinealidad graves.

3. **Análisis de los valores propios de $\mathbf{X}'\mathbf{X}$:** se trata de evaluar si hay valores propios con valores cercanos a cero. Para ello se definen las medidas que se presentan a continuación:

   Número de condición: mide la dispersión en el espectro de los valores propios de la matriz $\mathbf{X}'\mathbf{X}$. Se calcula como: $\kappa = \lambda_{\text{max}}/\lambda_{\text{min}}$. En *R* se obtienen valores en raíz cuadrada, es decir, $\sqrt{\kappa}$, para el cual el criterio para detectar multicolinealidad es:

   - Si $\sqrt{\kappa} \leq 10$ no hay problemas de multicolinealidad.
   - Si $10 < \sqrt{\kappa} \leq 31.62$ hay problemas de multicolinealidad moderada.
   - Si $\sqrt{\kappa} > 31.62$ hay problemas de multicolinealidad graves.

Índice de condición: es una medida útil para determinar el número de dependencias casi lineales en $\mathbf{X}'\mathbf{X}$. Se calcula como: $\kappa_j = \lambda_{\max}/\lambda_j,\ j = 1, \ldots, p$ (en *R* se obtienen los valores $\sqrt{\kappa_j}$). El criterio para detectar multicolinealidad es:

- Si $\sqrt{\kappa_j} \leq 10\ \forall j$, no hay problemas de multicolinealidad.
- Si al menos para un j, $10 < \sqrt{\kappa_j} \leq 31.62$, entonces hay problemas de multicolinealidad moderada.
- Si al menos para un j, $\sqrt{\kappa_j} > 31.62$, entonces hay problemas de multicolinealidad graves (por lo menos hay una asociación casi lineal entre dos o más predictoras).

Proporciones de descomposición de varianza: denotados π_{ij} representan la proporción de la varianza de cada $\widehat{\beta}_j$ (o de cada factor de inflación de varianza) debida al i-ésimo valor propio de la matriz $\mathbf{X}'\mathbf{X}$.

Proporciones altas ($\pi_{ij} > 0.5$) para dos o más coeficientes de regresión asociados con un mismo valor propio pequeño es evidencia de multicolinealidad entre las variables correspondientes a tales coeficientes.

NOTA: El análisis de valores propios se puede realizar usando datos centrados o con los datos originales.

Selección de variables (construcción de un modelo)

En algunos estudios observacionales o exploratorios se parte de un modelo de regresión en el que se considera un conjunto grande de variables predictoras potenciales para luego identificar un subconjunto entre tales variables, que resulte potencialmente útil para construir el modelo de regresión final.

Dependiendo de los usos que se deseen dar a un modelo de regresión variará el subconjunto de variables seleccionadas. Por ejemplo, desde el punto de vista del ajuste, cierto subgrupo de variables serán útiles, en tanto que desde el punto de vista del pronóstico, otro subconjunto podría resultar ser mejor. Es necesario pues fijar un criterio de selección del mejor subconjunto de variables.

Método de todas las regresiones posibles

Este procedimiento consiste en correr todos los $2^k - 1$ modelos posibles (con intercepto) de la variable respuesta vs. los posibles subconjuntos de variables predictoras,

$$\begin{array}{rl}
\binom{k}{1} = k & \text{modelos de una predictora} \\
\binom{k}{2} & \text{modelos de dos predictoras} \\
\vdots & \\
\binom{k}{k} = 1 & \text{modelo de } k \text{ predictoras} \\
\hline
2^k - 1 & \text{modelos posibles,}
\end{array}$$

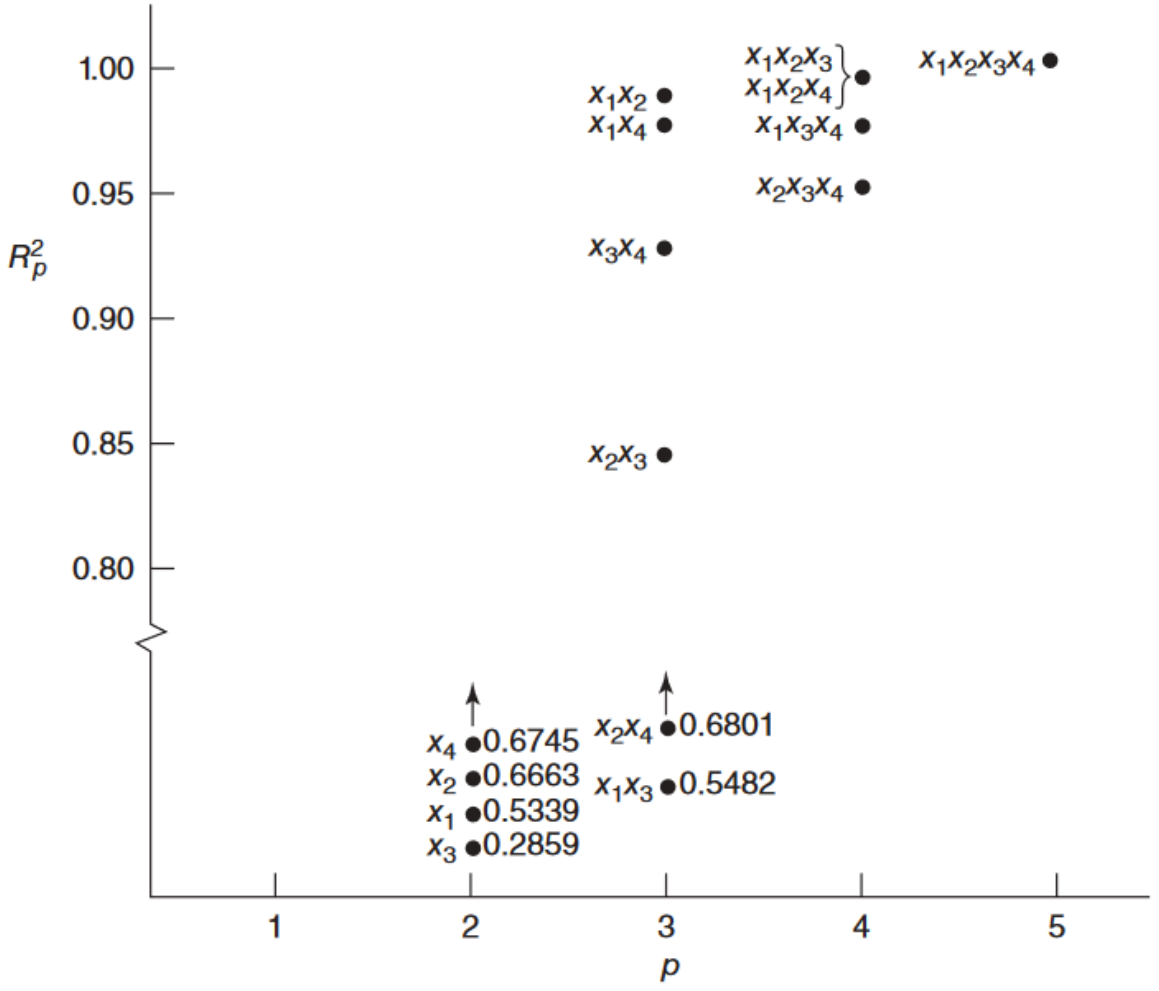
y comparar tales modelos con base en criterios estadísticos.

Criterios estadísticos en la comparación de modelos

- $\boldsymbol{R_p^2}$: el mejor modelo es aquel con el mayor valor en este estadístico, sin embargo, al ser una función no decreciente del número de predictoras, tiende a señalar al modelo con todas las predictoras.

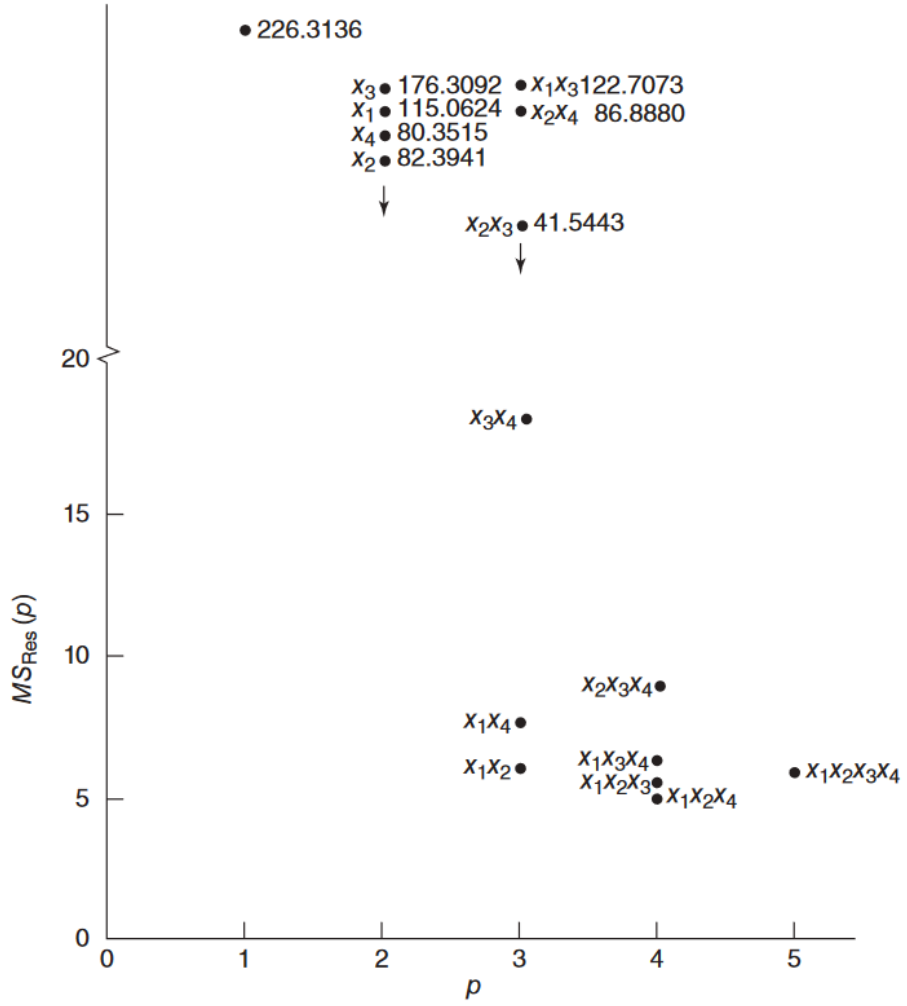
  Con el fin de resolver esta dificultad, se busca un modelo con un menor número de variables cuyo R^2 no cambie significativamente al aumentar el número de predictoras.

Ilustración criterio R_p^2



(*Fuente: Mongomery et al. (2002)*)

- $\boldsymbol{R^2_{adj,p}}$ (o $\mathbf{MSE}_p$): el mejor modelo es aquel con mayor (menor) valor en este estadístico.



(*Fuente: Mongomery et al. (2002)*)

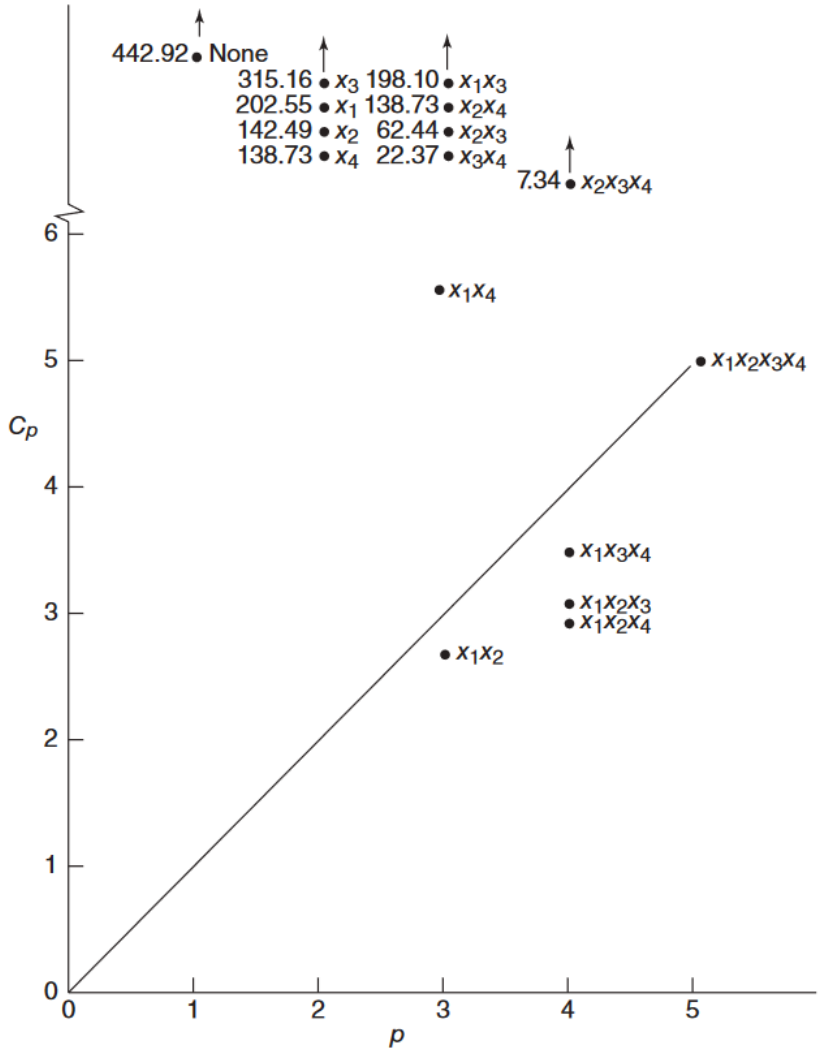
- **C_p de Mallows**: el mejor modelo es aquél para el cual C_p es el más pequeño posible (es decir, el modelo con el menor número de variables predictoras posible) y tal que la diferencia $|C_p - p|$ es mínima, con p igual al número de parámetros del modelo considerado, incluyendo el intercepto.

  Este estadístico es una medida del sesgo en el modelo de regresión, es decir, de $E\left(\widehat{Y}_i\right) - \mu_i$, y es tal que a mayor sesgo, mayor C_p. Este estadístico se calcula como:

$$C_p = \frac{\text{SSE}_p}{\text{MSE}\left(\beta_0, \beta_1, \ldots, \beta_k\right)} - (n - 2p)$$

  donde SSE_p es la suma de cuadrados del error del modelo considerado y $\text{MSE}\left(\beta_0, \beta_1, \ldots, \beta_k\right)$ es el cuadrado medio del error para el modelo de regresión con todas las k variables.

Ilustración criterio C_p



(*Fuente: Mongomery et al. (2002)*)

- **PRESS_p o suma de cuadrados de predicción**: mide qué también el uso de los valores ajustados por un submodelo puede predecir las respuestas observadas. Mientras menor sea esta medida, mejor se considera el modelo.

  El PRESS es como un SSE, pero en el cual el valor ajustado para cada observación Y_i se halla estimando el submodelo sin considerar dicha observación, tales valores ajustados se denotan por $\widehat{Y}_{(i)}$; así el PRESS es la suma de cuadrados de los residuales de predicción $e_{(i)} = Y_i - \widehat{Y}_{(i)}$, es decir, $\text{PRESS}_p = \sum_{i=1}^{n} e_{(i)}^2$.

Para cada submodelo, la definición del error de predicción implica correr n regresiones separadas (cada una con $n-1$ datos) con cada observación eliminada en cada caso, sin embargo, basta con correr una vez el modelo con todas las observaciones, hallar sus residuales ordinarios, y los elementos de la diagonal principal de la matriz hat y calcular $e_{(i)} = e_i / (1 - h_{ii})$.

Si se usan dos o más criterios de selección de modelos, es posible que cada criterio lleve a modelos distintos. La decisión final debe basarse en el análisis de residuales y otros diagnósticos, además de complementar con el conocimiento y la experiencia de personas expertas en el ámbito en el cual está inmerso el problema.