REGRESIÓN LINEAL MÚLTIPLE

SELECCIÓN DE VARIABLES Y CONSTRUCCION DEL MODELO

El problema de la construcción del modelo

La construcción de un modelo de regresión que solo incluya un subconjunto de las variables regresoras disponibles implica dos objetivos contradictorios:

1. Que el modelo incluya tantas regresoras como sea posible de tal manera que la información contenida en dichos factores pueda influenciar sobre el valor estimado de y.

2. Que el modelo incluya el menor número de regresores posibles debido a que la varianza de la predicción ^y se incrementa cuando el número de regresores se incrementa.

   El proceso de encontrar un modelo que sea un término medio o que combine entre los dos objetivos se llama ***selección de la "mejor" ecuación de regresión***.

Consecuencias de la mala especificación del modelo

Supóngase que hay K regresores candidatos, x_1, x_2 ... , X_K Y que hay n ≥ K+1 observaciones de esos regresores y de la respuesta. El modelo completo que contiene a todos los K regresores es

$$y_i = \beta o + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i, \quad i=1,2,\ldots,n$$

o lo que es igual $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

El modelo particionado para r variables omitidas se puede escribir en la siguiente forma

$$Y = X_p\beta_p + X_r\beta_r + \varepsilon$$

en donde la matriz X se ha dividido en

Xp, una matriz de n X p cuyas columnas representan la ordenada al origen y los p - 1 regresores que quedan en el modelo de subconjunto, y

Xr, una matriz de n X r cuyas columnas representan los regresores que se van a omitir del modelo completo, para el modelo particionado en β.

Para el modelo completo, el estimador de β por mínimos cuadrados es:

$\hat{\beta}^* = (X'X)^{-1} X'Y$

$$\hat{\sigma}_*^2 = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}^{*\prime}\mathbf{X}'\mathbf{y}}{n - K - 1} = \frac{\mathbf{y}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}}{n - K - 1}$$

Los componentes de $\hat{\beta}^*$ se representan por β_p^* y β_r^* y y_i^*, representa los valores ajustados.

$Y = X_p\beta_p + \varepsilon$

$\hat{\beta}_p = (X'_pX_p)^{-1} X'_pY$

$$\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'_P\mathbf{X}'_P\mathbf{y}}{n - p} = \frac{\mathbf{y}'\left[\mathbf{I} - \mathbf{X}_P(\mathbf{X}'_P\mathbf{X}_P)^{-1}\mathbf{X}'_P\right]\mathbf{y}}{n - p}$$

Propiedades de los estimadores ^βp y σ²,

1. El valor esperado de $\hat{\beta}_p$ es

$$E(\hat{\beta}_p) = \beta_p + (\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{X}_r\beta_r$$
$$= \beta_p + \mathbf{A}\mathbf{B}_r$$

en donde $A = (X'_pX_p)^{-1} X'_pX_r$,

$\beta\hat{}_p$ es un estimador sesgado de β_p,

2. Las varianzas de los estimadores por mínimos cuadrados, de los parámetros del modelo completo son mayor igual que las varianzas de sus correspondientes parámetros en el modelo reducido; en consecuencia, al eliminar las variables regresoras nunca se aumentan las varianzas de los estimados de los parámetros restantes.

3.Los estimadores por mínimos cuadrados de los parámetros en el modelo de subconjunto tienen menor error cuadrático medio que los correspondientes del modelo completo, cuando las variables eliminadas tienen coeficientes de regresión que son menores que los errores estándar de sus estimados en el modelo completo.

4. Sea $\hat{\sigma}^2$ un estimador insesgado de σ^2 para el modelo completo, en el modelo reducido $\hat{\sigma}^{2*}$ esto es
$E(\hat{\sigma}^{2*})$, es un estimado sesgado de σ^2.

CRITERIOS PARA EVALUAR UN SUBCONJUNTOS DE REGRESORES EN MODELOS DE REGRESIÓN

COEFICIENTE DE DETERMINACIÓN R^2

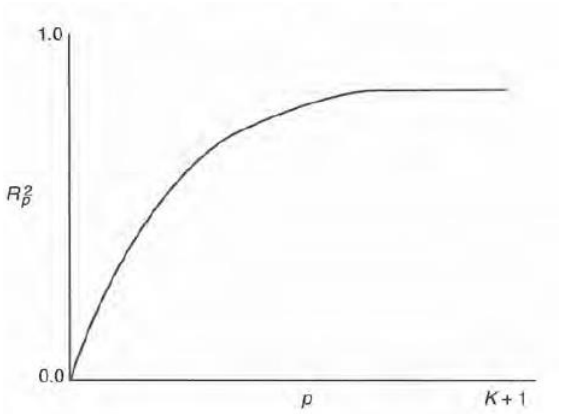
Este coeficiente se define como sigue:

R^2_p = SCR(p)/SCT = 1- (SCRes(p)/SCT)

- Es una medida descriptiva de la relación entre la variable respuesta y el conjunto de variables regresoras consideradas, es decir mide solo la bondad de ajuste de los datos muestrales, por lo tanto aunque $R^2_p \rightarrow 1$ no hay garantía que el modelo sea apropiado para realizar pronósticos de la variable respuesta.

- Si se va a comparar varios modelos propuestos, todos estos modelos deben tener la misma variable respuesta.

- **R^2_p** se incrementa cuando se agregan variables regresoras al modelo



Dado que no existe un valor óptimo para R^2_p para un subconjunto de modelos de regresión siempre se debe buscar un valor "satisfactorio". La solución para este problema es utilizar una prueba mediante el cual todos los subconjuntos de modelos de regresión que tienen una R^2p no muy distinta de la R^2 para el modelo completo.

Se define la prueba como:

$$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{a,\,n,\,K})$$

En donde

$$d_{a,\,n,\,K} = \frac{KF_{a,\,n,\,n-K-1}}{n - K - 1}$$

Todo subconjunto de variables regresoras que produce un R^2 mayor que R^2_0 se considera adecuado.

COEFICIENTE DE DETERMINACIÓN MULTIPLE R^2 AJUSTADO

Para evitar las dificultades de interpretar a R^2, es preferible utilizar el coeficiente de determinación múltiple R^2 ajustado. Algunos analistas prefieren usar la estadística R^2 ajustada, definida para una ecuación de *p* términos como sigue:

$$R^2_{aj,p} = 1 - \left[(n-1)/(n-p)\right](1 - R^2p)$$

La estadística $R^2_{aj,p}$ no necesariamente incrementa su valor al incluir nuevas variables regresores al modelo. Por lo que se tiene que:

$R^2_{aj,p+s} > R^2_{aj,p}$ si, y sólo si la estadística F parcial para probar el significado de los s regresores adicionales es mayor que 1, en consecuencia, un criterio para seleccionar un modelo con subconjunto óptimo es elegir el que tenga una $R^2_{aj,p}$ máxima.

CUADRADO MEDIO DE RESIDUALES

El cuadrado medio de residuales se define:

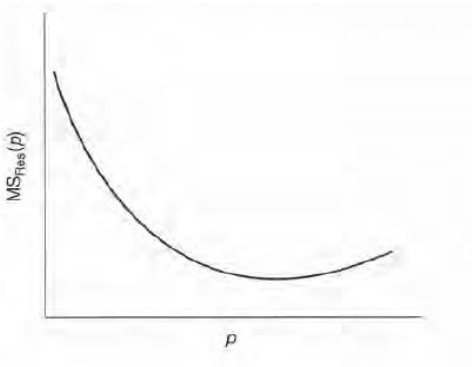
CMRes(p) = SCRes(p)/(n-p)

Podemos utilizar el cuadrado medio de residuales como un criterio de evaluación de un modelo..
Algunos investigadores basa su elección según el comportamiento.

1. El CMRes(p) mínimo.

2. El valor de p tal que CMRes(p) sea aproximadamente igual a CMRes para el modelo completo, o bien

3. Un valor de p cercano al punto en donde crece el CMRes(p) mínimo.



ESTADISTICA Cp de MALLOWS

Los criterios anteriores se basan en el $CMerror$, pero también es interesante tener en cuenta el sesgo en la selección del modelo ya que si se omite una variable regresora importante los estimadores de los coeficientes de regresión son sesgados y los criterios anteriores pueden elegir un modelo que tenga sesgo grande aunque su $CMerror$ sea pequeño.

El estadístico de Mallows se define como:

$$C_p = p + (n-p)\frac{s_p^2}{s^2} - (n-p) = \frac{SSE_p}{s^2} - (n-2p)$$

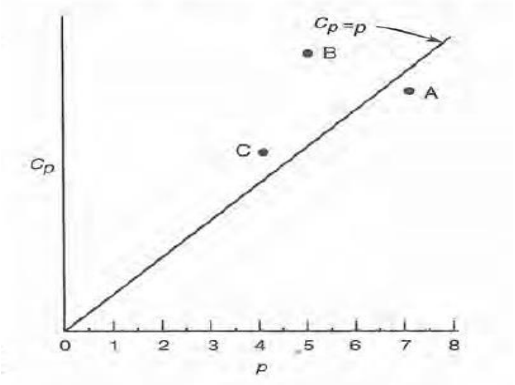
- Ideal $Cp = p$
- Sobreexplicado $Cp > p$
- Infraexplicados $Cp < p$

Mallows propuso un criterio relacionado al error cuadrático medio, con dicho criterio se desea encontrar el modelo que minimice el ECM de predicción para los puntos observados.

Utilice:

- El Cp de Mallows como ayuda para elegir entre múltiple modelos de regresión.
- Porque ayuda a alcanzar un equilibrio importante con el número de predictores en el modelo.
- El Cp de Mallows porque compara la precisión y el sesgo del modelo completo con modelos que incluyen un subconjunto de los predictores.

Se trata de encontrar un modelo donde el sesgo y la varianza sean moderados.



Tarea: Demostrar el desarrollo de la estadística Cp de Mallows