



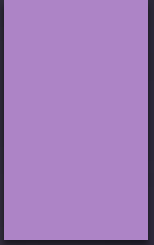
Selección de variables y formación del modelo

ANÁLISIS DE REGRESIÓN

Construcción del modelo

En la mayoría de los problemas prácticos el analista tiene un grupo de regresores candidatos, que deberían incluir a todos los factores influyentes, y debe determinar el subconjunto real de regresores que debe usarse en el modelo. La definición de un subconjunto adecuado de regresores para el modelo es lo que se llama problema de selección de variables.

El proceso de encontrar un modelo que sea un término medio entre: 1) incluir tantos regresores como sea posible (para no perder información) y 2) que el modelo incluya los menos regresores posibles (modelos más sencillos con menor costo de recolección) se llama selección de la mejor ecuación de regresión.



Criterios para evaluar modelos de regresión

SELECCIÓN DE VARIABLES

Coeficiente de determinación múltiple

Una medida de la adecuación del modelo es el coeficiente de determinación múltiple para un modelo de regresión con subconjunto de p términos, esto es, $p-1$ regresores, definido como:

$$R_p^2 = \frac{SSR_p}{SST}$$

donde SSR_p representa la suma de cuadrados de la regresión para un modelo de subconjunto de p términos y SST la suma de cuadrados total para el modelo completo.

Nótese que hay $\binom{k}{p-1}$ valores de R_p^2 , para cada valor de p , uno para cada posible modelo de subconjunto de tamaño p . En general, no es directo el uso de R^2 como criterio para escoger la cantidad de regresores que se incluirán en el modelo, sin embargo, para una cantidad fija p de variables, se puede usar R_p^2 para comparar los $\binom{k}{p-1}$ modelos generados, prefiriendo modelos con valores grandes de R_p^2 .

R^2 ajustada

Para evitar las dificultades del uso de R^2 , algunos analistas prefieren el uso de R^2_{adj} , definida para una ecuación de p términos como sigue:

$$R^2_{adj,p} = 1 - \left(\frac{n-1}{n-p} \right) (1 - R^2_p)$$

Un criterio para seleccionar un modelo con subconjunto óptimo es elegir el que tenga $R^2_{adj,p}$ máxima.

Cuadrado medio de residuales, MSE

El cuadrado medio de residuales para un modelo de regresión de subconjunto, definido como $MSE_p = \frac{SSE_p}{n-p}$, se puede usar como criterio para evaluar un modelo.

La selección se basa en MSE_p mínimo o el valor de p tal que MSE_p sea aproximadamente igual a MSE para el modelo completo. El modelo de regresión para subconjunto, que minimiza MSE_p también maximizará $R^2_{adj,p}$.

Estadístico C_p de Mallows

Mallows ha propuesto un criterio que se relaciona con el error cuadrático medio de un valor ajustado. Denotado por C_p y definido como:

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p$$

Para calcular C_p se necesita un estimador insesgado de σ^2 , con frecuencia se usa el cuadrado medio de residuales para la ecuación completa, así se supone que el modelo completo tiene sesgo despreciable. En general, se prefieren valores pequeños de C_p , en modelos sin sesgo $C_p = p$, entonces para k regresores un valor $C_p \cong p = k + 1$ son “buenos”.



Selección de Variables

TODAS LAS REGRESIONES POSIBLES

Todas las regresiones posibles

Para determinar el conjunto de variables que se van a usar en la ecuación del modelo final, es natural la consideración de ajuste de modelos con varias combinaciones de los regresores candidatos. Este procedimiento requiere que el analista ajuste todas las ecuaciones de regresión, que tengan un regresor candidato, dos regresores candidatos, etc. Esas ecuaciones se evalúan de acuerdo con algún criterio adecuado y se selecciona el “mejor” modelo de regresión.

Si se supone que el término de ordenada al origen β_0 se incluye en todas las ecuaciones, entonces si hay k regresores candidatos, hay 2^k ecuaciones en total por estimar y examinar.