Tema 5
Técnicas de regresión avanzadas (I)
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

- Podemos generalizar
- Tenemos

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR},$$

donde SST es la suma de cuadrados total, SSE la suma de cuadrados del error y SSR la suma de cuadrados de la regresión.

- SST mide la variabilidad de la muestra Y respecto a la media muestral.
- SSE mide la variabilidad de los datos de Y respecto a los valores ajustados.
- SSR es la parte de la varianza que explican todas las variables predictoras X_j del modelo.

El coeficiente de determinación general

- Podemos definir el coeficiente de determinación general:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

- En este caso general, no resulta útil para poder elegir entre diferentes modelos con un número de variables predictoras distinto.
- A medida que añadimos variables, por definición R^2 aumenta.
- Podemos definir un coeficiente de determinación ajustado:

$$R^2_{\text{adj}} = 1 - \frac{(n-1)\text{SSE}}{(n-m-1)\text{SST}}.$$

- Con este nuevo coeficiente, al aumentar m el valor de R^2_{adj} decrece.
- Este coeficiente penaliza modelos complejos.

Colinealidad

- Una hipótesis extra a comprobar es la colinealidad.
- Colinealidad: relación lineal entre dos o más variables.
- Problemas de identificabilidad: ¿a qué se deben los cambios en la variable respuesta?
- Pequeños cambios en la muestra pueden provocar grandes cambios en la estimación de los parámetros.
- El VIF (factor de inflación de la varianza) detecta esta colinealidad:

$$\text{VIF} = \frac{1}{1 - R^2}.$$

- A mayor VIF, mayor colinealidad. Para VIF> 10 puede haber un problema de colinealidad.
- El análisis puede hacerse antes de la regresión o después con las variables que se hayan usado para ajustar el modelo.

Situaciones:

- Algunas variables predictoras no aportan información.
- Algunsa variables están muy relacionadas entre sí (colinealidad).
- Si hay dos modelos con capacidad predictiva similar, escogemos el más sencillo (parsimonia/Ockham).

Best subset:

- Fijado el número de variables se analizan todos los modelos posibles y se obtiene el mejor según capacidades predictivas (p.ej. R^2, R^2_{adj}...).
- Para m grande tiene un coste de computación elevado (crecimiento exponencial: 2^m modelos).

Stepwise selection:

- Computacionalmente más económica.
- Se busca el mejor modelo de manera secuencial.
- Selección hacia adelante: partiendo del modelo nulo se añade en cada paso la variable que añada más información al modelo. El proceso termina cuando añadir una variable no mejora el modelo.
- Selección hacia atrás: partiendo del modelo con todas las variables, se elimina en cada paso la variable tal que el modelo mejora más. El proceso termina cuando al eliminar variables no mejora el modelo.
- Si $m > n$ solamente se puede usar la selección hacia adelante y terminar en el paso n.

Problemas en la práctica. Datos faltantes

- Datos perdidos, medidas no tomadas o tomadas mal, etc.
- Una primera opción es eliminar las observaciones con algún dato faltante en alguna de las variables.
- La usaremos con muestras grandes y pocos datos faltantes, o si estimarlos es difícil.
- Perdemos información que teníamos.
- Como alternativa, podemos usar una regresión para poder imputar el dato faltante o imputarlo por la mediana o la moda.

Pasos:

- Centrar las variables (restar la media).
- Estandarizar (restar la media y dividir por la desviación estándar).
- Escalar la variable a $[0, 1]$.