Capítulo 8

Selección de variables

8.1. Introducción

En muchos problemas de regresión es posible considerar un número importante de variables predictoras. Un empresario podría estudiar los factores que afectan la calidad de su producto a través de un conjunto de predictores como la experiencia de los empleados, proveedores de materia prima, algunas características del proceso de producción y muchas otras. En un estudio clínico para modelar el tamaño de un tumor se podrían tener muchos predictores potenciales que describen el estatus del paciente y factores ambientales que podrían ser relevantes.

Si se tienen muchos predictores es necesario identificar aquellos considerados importantes o *activos* y los que no son importantes o *inactivos*. Los métodos de selección de variables desarrollados en este capítulo son usados para este propósito. Las estimaciones y predicciones son, por lo general, más precisas cuando el modelo se estima usando los términos relevantes.

La selección de variables para regresión lineal no es la única metodología al problema de modelamiento de una variable respuesta como función de una gran cantidad de términos o predictores. Las áreas de *machine learning* y *data mining* proporcionan algunas alternativas para este problema y en algunas circunstancias los métodos desarrollados en estas áreas podrían dar mejores resultados. Una introducción se pueden consultar en Hastie, Tibshirani y Friedman (2001).

8.2. Los términos activos

Sea Y una variable respuesta y X un conjunto de términos obtenidos a partir de los predictores. El objetivo del proceso de selección de variables es dividir $X = (X_{\mathcal{A}}, X_{\mathcal{I}})$ donde $X_{\mathcal{A}}$ es el conjunto de términos activos y $X_{\mathcal{I}}$ es el conjunto de términos inactivos. Suponga que la función media es:

$$\mathrm{E}(Y|X = \mathbf{x}) = \boldsymbol{\beta}_{\mathcal{A}}^{'}\mathbf{x}_{\mathcal{A}} + \boldsymbol{\beta}_{\mathcal{I}}^{'}\mathbf{x}_{\mathcal{I}} \tag{8.2.1}$$

Para los téminos inactivos se considera que $\boldsymbol{\beta}_{\mathcal{I}} = \mathbf{0}$. Si se tiene un tamaño de muestra lo suficientemente grande es posible identificar fácilmente los términos activos como aquellos cuyos coeficientes estimados son diferentes de cero y los términos inactivos como aquellos cuyos coeficientes son cercanos a cero.

8.2.1. Colinealidad

Dos términos X_1 y X_2 son exactamente *colineales*, o linealmente independientes, si existe una ecuacion lineal tal que:

$$c_1X_1 + c_2X_2 = c_0 \tag{8.2.2}$$

donde c_0, c_1 y c_2 son constantes. La colinealidad entre X_1 y X_2 se mide por el cuadrado de su coeficiente de correlación muestral r_{12}^2. La colinealidad exacta corresponde a $r_{12}^2 = 1$ y la no colinealidad a $r_{12}^2 = 0$.

Un conjunto de términos X_1, X_2, $\cdots$, X_p es exactamente colineal si:

$$c_1X_1 + c_2X_2 + \cdots + c_pX_p = c_0$$

donde c_0, c_1, $\cdots$, c_p son constantes tales que al menos una es diferente de cero.

8.2.2. Colinealidad y variancias

Cuando $p > 2$, la varianza del $j-$ésimo coeficiente es:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2}\frac{1}{S_{X_jX_j}} \tag{8.2.3}$$

donde $S_{X_jX_j} = \sum(x_{ij} - \overline{x}_j)^2$. La cantidad $(1 - R_j^2)^{-1}$ es llamado el $j-$ésimo *factor de inflación de variancia* o $\mathcal{VIF}_j$ (Marquardt, 1970) y representa

el incremento en la variancia debido a la correlación entre predictores. En términos prácticos se considera que si el $\mathcal{VIF}_j > 10$ se tiene evidencia de la presencia de colinealidad.

8.3. Selección de variables

El objetivo del proceso de seleción de variables es dividir X en el conjnto de términos activos $X_{\mathcal{A}}$ y el conjunto de términos inactivos $X_{\mathcal{I}}$. Existen dos problemas a resolver. Primero, dado un candidato particular $X_{\mathcal{C}}$ para los términos activos: ¿qué criterio debería ser usado para comparar $X_{\mathcal{C}}$ con otras posibles elecciones para $X_{\mathcal{A}}$? El segundo problema es computacional: ¿como manejar el número de comparaciones que deben realizarse?

8.3.1. Criterios de información

Suponga que se tiene un subconjunto candidato $X_{\mathcal{C}}$. Si $X_{\mathcal{C}} = X_{\mathcal{A}}$, entonces los valores estimados obtenidos con la función media:

$$\mathrm{E}(Y|X_{\mathcal{C}} = \mathbf{x}_{\mathcal{C}}) = \boldsymbol{\beta}_{\mathcal{C}}^{'}\mathbf{x}_{\mathcal{C}} \tag{8.3.1}$$

debería ser similares a los valores estimados de la función media 8.2.1 y la suma de cuadrados del residual para el modelo anterior debería ser muy cercana a la del modelo 8.2.1.

Los criterios para comparar varios subconjuntos candidatos estan basados en la falta de ajuste de un modelo y su *complejidad*. La falta de ajuste para un subconjunto candidato $X_{\mathcal{C}}$ se mide usando la suma de cuadrados del residual $SCR_{\mathcal{C}}$. La complejidad de un modelo de regresión lineal mútiple se mide por el número de términos $p_{\mathcal{C}}$ en $X_{\mathcal{C}}$ incluyendo el intercepto. El *criterio de información de Akaike* es:

$$AIC = n\log(SCR_{\mathcal{C}}/n) + p_{\mathcal{C}} \tag{8.3.2}$$

Los mejores subconjuntos de candidatos tienen menores valores para este indicador. Otra alternativa es el *criterio de información Bayesiano*:

$$BIC = n\log(SCR_{\mathcal{C}}/n) + p_{\mathcal{C}}\log(n) \tag{8.3.3}$$

que proporciona un balance diferente entre la falta de ajuste y la complejidad. Nuevamente se prefieren valores pequeños para este indicador.

Un tercer criterio es llamado C_p *de Mallows*:

$$C_{p_{\mathcal{C}}} = \frac{SCR_{\mathcal{C}}}{\hat{\sigma}^2} + 2p_{\mathcal{C}} - n \tag{8.3.4}$$

donde $p_{\mathcal{C}}$ es el número de términos en $X_{\mathcal{C}}$ y $\hat{\sigma}^2$ se obtiene usando el modelo 8.2.1.

El *coeficiente de determinación ajustado* se define por:

$$R^2_{aj} = 1 - \frac{SCR/(n - p_{\mathcal{C}} - 1)}{SC_{YY}/(n-1)} \tag{8.3.5}$$

y a diferencia de los criterios anteriores se prefieren valores altos para este indicador.

8.3.2. Validación cruzada

Es posible usar el proceso de *validación cruzada* para comparar un subconjunto candidato de funciones media. La data se divide aleatoriamente en dos partes: un conjunto de *entrenamiento* y un conjunto de *prueba*. El conjunto de entrenamiento es usado para estimar los parámetros en la función media que luego son usados para estimar los valores en el conjunto de prueba. La media de las diferencias entre la variable respuesta y los valores estimados al cuadrado para el conjunto de prueba son usados como un indicador para el subconjunto candidato. Los buenos candidatos para $X_{\mathcal{A}}$ tienen los menores errores por validación cruzada.

Otra opción es considerar los *residuales de predicción* para $X_{\mathcal{C}}$, obtenidos como la diferencia entre el caso i y su estimación usando la ecuación de regresión que no lo incluye. La suma de cuadrados de estos valores es llamada *suma de cuadrados de los residuales de predicción*, o *PRESS*:

$$PRESS = \sum_{i=1}^{n} (y_i - \mathbf{x}'_{\mathcal{C}_i} \hat{\boldsymbol{\beta}}_{\mathcal{C}_{(i)}})^2 = \sum_{i=1}^{n} \left(\frac{\hat{e}_{\mathcal{C}_i}}{1 - h_{\mathcal{C}_{ii}}} \right)^2 \tag{8.3.6}$$

donde $\hat{e}_{\mathcal{C}_i}$ y $h_{\mathcal{C}_{ii}}$ son, respectivamente, el residual y el leverage para el i-ésimo caso.

8.3.3. Métodos computacionales

El algoritmo para el método *best subset selection* es:

1. Sea $\mathcal{M}_0$ que denota el *modelo nulo*, es decir aquel que no tiene predictores.

2. Para $k = 1, 2, \cdots, p$:

   a) Estimar los $\binom{p}{k}$ modelos que contienen exactamente k predictores.

   b) Tomar el mejor de los modelos anteriores, llamado $\mathcal{M}_k$. El mejor modelo es aquel que tiene la menor *SCR* o equivalentemente el mayor R^2.

3. Seleccionar el mejor modelo a partir de $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$ usando *AIC*, *BIC*, C_p, R^2 ajustado o PRESS.

Por razones computacionales *best subset selection* no puede usarse cuando p es grande. Los métodos *stepwise selection* requieren examinar solo un subconjunto de modelos de un tamaño específico. Los métodos stepwise tienen dos variaciones básicas y consideran que el intercepto es el único término presente en todos estos modelos. El algoritmo del método *forward selection* es:

1. Sea $\mathcal{M}_0$ que denota el *modelo nulo*, es decir aquel que no tiene predictores.

2. Para $k = 0, 1, \cdots, p-1$:

   a) Considere todos los $p - k$ modelos que aumentan en uno los predictores en $\mathcal{M}_k$.

   b) Tomar el mejor entre los modelos anteriores y llamarlo $\mathcal{M}_{k+1}$. El mejor modelo es aquel que tiene la menor *SCR* o equivalentemente el mayor R^2.

3. Seleccionar el mejor modelo a partir de $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$ usando *AIC*, *BIC*, C_p, R^2 ajustado o PRESS.

El método *forward selection* puede aplicarse aún cuando $n < p$. El algoritmo del método *backward selection* es:

1. Sea $\mathcal{M}_p$ que denota el *modelo completo*, es decir aquel que contiene a los p predictores.

2. Para $k = p, p-1, \cdots, 1$:

   a) Considere todos los k modelos que contienen todos los predictores en $\mathcal{M}_k$ menos uno .

   b) Tomar el mejor entre los modelos anteriores y llamarlo $\mathcal{M}_{k-1}$. El mejor modelo es aquel que tiene la menor SCR o equivalentemente el mayor R^2.

3. Seleccionar el mejor modelo a partir de $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$ usando AIC, BIC, C_p, R^2 ajustado o PRESS.