Regresión: Detalles de las cuestiones básicas

Sea n la cantidad de observaciones y m la de parámetros (incluyendo la constante si la hay); los predictores están en la matriz $\mathbf{X}$ de $n \times m$, y las respuestas en el vector $\mathbf{y}$. Suponemos $\mathbf{X}$ fija. "Var" indicará la varianza de una variable aleatoria y "**Var**" la matriz de covarianzas de un vector aleatorio.

Sean $\widehat{\beta}$ el estimador de mínimos cuadrados, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$ el vector de valores ajustados y $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$ los residuos.

Sean $\mathbf{H}$ la "hat matrix"

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

y h_i $(i = 1, .., n)$ sus elementos diagonales: $h_i = H_{ii}$ (el "leverage"). Entonces $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.

Bajo el modelo lineal standard se supone

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{e} \tag{1}$$

donde β_0 es el vector de parámetros desconocido, $\mathrm{E}\mathbf{e} = \mathbf{0}$ y $\mathbf{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$; aquí se cumple

$$\mathrm{Var}(r_i) = \sigma^2(1 - h_i).$$

1 Gráficos de residuos

Sea $\mathbf{r} = \mathbf{y} - \mathbf{X}\widehat{\beta}$ el vector de residuos. El método clásico para detectar outliers es usar $t_i = r_i/\mathrm{dt}(r_i)$ $(i = 1, ..., n)$, donde $\mathrm{dt}(r_i)$ es una estimación de la desviación de r_i. Pero esto no es muy confiable. Es mejor usar los residuos de "cross validation" (o "leave-one-out"). Sea para cada i : $\widehat{\beta}_{-i}$ el estimador de mínimos cuadrados calculado *sin* usar la observación i, y r_{-i} el residuo correspondiente:

$$r_{-i} = y_i - \mathbf{x}_i'\widehat{\beta}_{-i}$$

donde $\mathbf{x}_i'$ es la fila i-ésima de $\mathbf{X}$. Para esto no hace falta recalcular todo cada vez, pues se puede mostrar que

$$r_{-i} = \frac{r_i}{1 - h_i}.$$

Uno de los gráficos útiles es el de r_{-i} contra $\widehat{y}_i$.

El otro es el QQ-plot de residuos. Para ello conviene que estén normalizados. Como r_{-i} es un múltiplo de r_i, al normalizarlo da lo mismo, o sea

$$t_i = \frac{r_i}{s\sqrt{1 - h_i}},$$

donde s es la tradicional estimación de σ. Pero es mejor usar

$$t_{-i} = \frac{r_i}{s_{-i}\sqrt{1 - h_i}},$$

donde s_{-i} es la s calculada sin la observación i, o sea

$$s^2_{-i} = \frac{1}{n-m-1} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i' \widehat{\beta}_{-i} \right)^2 .$$

Se prueba que

$$s^2_{-i} = \frac{1}{n-m-1} \left[(n-m)s^2 - \frac{r_i^2}{1-h_i} \right] .$$

Ver Belsley-Kuh-Welsch, *Regression Diagnostics*, Wiley, pg. 14.

Lo mejor es hacer un gráfico de los $t_{(i)}$ con la distribución normal y vigilar los puntos que se alejan de una recta. Para algo rápido, un $\left| t_{(i)} \right| > 2.5$ se puede considerar sospechoso.

Para investigar la posible heteroscedasticidad, y ver si $\mathrm{Var}\,(e_i)$ depende de $\mathbf{x}_i$, conviene graficar los $|t_i|$ vs. $\widehat{y}_i$ (notar que bajo homosedsticidad los t_i tienen la misma varianza pero los r_i no).

2 Selección de variables para predicción

La idea es que quiero un vector de parámetros que me dé una buena predicción para una y "nueva", es decir, que no pertenece a la muestra. Se usará el criterio de error medio cuadrático (EMC). Para definirlo tenemos dos casos. Supondremos que $(\mathbf{X}, \mathbf{y})$ cumplen (1).

2.1 Definición del EMC

1) Supongamos que $\mathbf{X}$ es fija y que Sea $\mathbf{y}_0 \in R^n$ una nueva observación que cumple (1), independiente de $(\mathbf{X}, \mathbf{y})$; o sea, $\mathbf{y}_0 = \mathbf{X}\beta_0 + \mathbf{e}_0$. Entonces para un estimador $\widehat{\beta} = \widehat{\beta}\,(\mathbf{X}, \mathbf{y})$, el EMC es

$$\mathrm{EMC}\left(\widehat{\beta}\right) = \frac{1}{n} \mathrm{E} \left\| \mathbf{y}_0 - \mathbf{X}\widehat{\beta} \right\|^2 . \tag{2}$$

2) Si $\mathbf{X}$ es aleatoria, se supone que $(\mathbf{x}_i, y_i)$ son independientes igualmente distribuidos. Sea $(\mathbf{x}_0, y_0)$ independiente de $(\mathbf{X}, \mathbf{y})$ y tal que $y_0 = \mathbf{x}_0'\beta_0 + e_0$. Entonces se define

$$\mathrm{EMC}\left(\widehat{\beta}\right) = \mathrm{E}\left(y_0 - \mathbf{x}_0' \widehat{\beta} \right)^2 .$$

Los métodos que siguen se aplican en ambos casos.

Ahora para cada subconjunto de variables $A \subset \{1, ...m\}$ calculo el estimador de mínimos cuadrados, y me quedo con el que da menor EMC. Para eso tengo que estimar el EMC.

2.2 Estimación del EMC

De los métodos posibles, prefiero la "validación cruzada" (cross-validation). Para cada (sub)muestra de estimación ("learning sample") tenemos una muestra de evaluación ("test sample"). El caso más simple es cuando ésta es de tamaño 1 ("leave-one-out"). Aquí tenemos para el estimador de mínimos cuadrados

$$\widehat{\mathrm{EMC}} = \frac{1}{n}\sum_{i=1}^{n} r_{-i}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i^2}{(1-h_i)^2}. \tag{3}$$

Sea $p = \#(A)$; entonces $\sum_{i=1}^{n} h_i = p$, y por lo tanto el promedio de los h_i es p/n. Una simplificación de (3) consiste en reemplazar h_i por p/n, lo que da

$$\frac{n}{(1-p)^2}\sum_{i=1}^{n} r_i^2.$$

El resultado se llama "generalized cross-validation" (GCV). Una simplificación extra se hace teniendo en cuenta que para $p << n$ es $(1-p/n)^2 \approx 1-2p/n$. El criterio que resulta es llamado por algunos autores S_p :

$$S_p = \frac{1}{(n-2p)}\sum_{i=1}^{n} r_i^2.$$

En la mayoría de los casos, lo más probable es que los tres criterios conduzcan a selecciones semejantes, pero yo prefiero (3).

2.3 Selección de un subconjunto

Para seleccionar predictores describo el método "backwards stepwise regression".

Primero tomamos $p = m$.

Dado el subconjunto A, sean $\mathbf{X}_A$ la matriz de $n \times p$ con los predictores seleccionados, $\widehat{\beta}_A = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$ el vector de parámetros estimados por mínimos cuadrados, $r_{A,i}$ $(i = 1, ..., n)$ los residuos. Sea $\widehat{\mathrm{EMC}}_A$ el estimador del EMC correspondiente. Sean

$$s_A^2 = \frac{\sum_{i=1}^{n} r_{A,i}^2}{n-p}, \quad \mathbf{C}_A = s_A^2(\mathbf{X}_A'\mathbf{X}_A)^{-1}.$$

Aquí $\mathbf{C}_A$ es el estimador de la matriz de covarianzas de $\widehat{\beta}_A$. Los "estadísticos t" de los estimadores son

$$T_j = \frac{\widehat{\beta}_{A,j}}{\sqrt{C_{A,jj}}} \ (j = 1, .., p),$$

donde C_{jj} son los elementos diagonales de $\mathbf{C}_A$. Entonces la variable a eliminar es la "menos significativa", es decir, la que tiene $|T_j|$ minimo.

Luego de eliminar el predictor j, se pasa de p a $p-1$ y se repite, hasta $p=1$ (la constante puede forzarse a permanecer si se desea).

Así se obtiene una sucesión de los $\widehat{\mathrm{EMC}}_A$, y se elige el A que da el mínimo.

Dado que $\widehat{\mathrm{EMC}}_A$ tiene error aleatorio, esto no hay que tomárselo estrictamente: se puede tomar el menor p que da un valor dentro de –digamos– el 5% del S_p mínimo. Una manera de formalizar esto (optativa) es la "one SD rule" que se describe como sigue.

Cualquiera de las versiones de $\widehat{\mathrm{EMC}}_A$ es un promedio:

$$\widehat{\mathrm{EMC}}_A = \frac{1}{n}\sum_{i=1}^{n} u_i = \overline{u}$$

donde en particular para (3) es $u_i = \left(r_i/\left(1-h_i\right)\right)^2$. Sea s_u el desvío estimado de los u :

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(u_i-\overline{u}\right)^2.$$

Entonces el desvío estimado de $\widehat{\mathrm{EMC}}_A = \overline{u}$ es $d_A = s_u/\sqrt{n}$.

Sea $A_{\min}$ el subconjunto que minimiza $\widehat{\mathrm{EMC}}_A$, sean $\widehat{\mathrm{EMC}}_{\min}$ el correspondiente EMC, y $d_{\min} = d_{A_{\min}}$. Entonces la regla es tomar el subconjunto A más chico de entre los que cumplen $\widehat{\mathrm{EMC}}_A \leq \widehat{\mathrm{EMC}}_{\min} + d_{\min}$.

Ver Hastie-Tishirani-Friedman *The elements of Statistical Learning*, Springer.

3 Otras cuestiones

1) Consideremos una situación heteroscedástica:

$$y_i = \mathbf{x}_i'\beta + e_i, \quad \text{con } \mathrm{Var}\left(e_i\right) \quad = \sigma_i^2.$$

Las σ_i las conozco o estimo. Para estimar uso mínimos cuadrados pesados. Una manera de hacerlo es "normalizar" los datos, dividiendo cada uno por el correspondiente desvío. O sea, calculo $\widetilde{y}_i = y_i/\sigma_i$ y $\widetilde{\mathbf{x}}_i = \mathbf{x}_i/\sigma_i$, y luego aplico cuadrados mínimos ordinarios a $\left(\widetilde{\mathbf{x}}_i, \widetilde{y}_i\right)$.

Cuando el modelo tiene intercept, el primer elemento del vector $\mathbf{x}_i$ es 1. ¡¡Ese "1" también hay que dividirlo por σ_i!!. Olvidar esto suele conducir a resultados disparatados.

2) Si no tengo razones para pensar que la intercept es nula, debo ponerla en el modelo. El costo es casi nulo, y aparte de darme un modelo más flexible, tiene una ventaja extra. Si los errores no tienen media nula (o sea $\mathrm{E}e_i = \mu \neq 0$) el estimador tiene sesgo. Pero si hay intercept, todo el sesgo se lo lleva la intercept, y las pendientes (que son lo que suele importar más) quedan libres. La demostración es fácil.

3) Salvo que se diga específicamente, en los datasets el orden cronológico de las observaciones es desconocido. Por lo tanto aplicar tests para autocorrelación, como Durbin-Watson, puede dar cualquier cosa.

4) Recordar que en el modelo lineal con x aleatoria no se hace ninguna suposición sobre la distribución de las x ni de las y. En todo caso, se supone normalidad de los e_i para poder deducir tests e intervalos de confianza. Pero eso no implica que las y sean normales.

4 ¿Qué buscar: modelo o predictor?

En regresión, los objetivos de predecir y de "elegir las variables relevantes" parecen semejantes, pero no son para nada equivalentes, por lo que conviene tener clara la diferencia. Vamos a ver una situación donde llevan a conclusiones distintas.

4.1 Un poco de teoría

Considero el modelo lineal con predictores fijos:

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{e} \tag{4}$$

donde $\mathbf{X} = [\mathbf{x}_{ij}]$ es de $n \times p$ y

$$\mathrm{E}\mathbf{e} = \mathbf{0}, \quad \mathrm{Var}\left(\mathbf{e}\right) = \sigma^2\mathbf{I} \tag{5}$$

donde "Var" es la matriz de covarianzas.

Tengo un estimador $\widehat{\beta} = \widehat{\beta}\left(\mathbf{X}, \mathbf{y}\right)$ que produce un predictor $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$. Para evaluarlo uso el error medio cuadrático (EMC). Sea

$$\mathbf{y}_* = \mathbf{X}\beta_0 + \mathbf{e}_*$$

donde $\mathbf{e}_*$ tiene la misma distrubución que $\mathbf{e}$ y es independiente de $\mathbf{e}$. Entonces se define

$$\mathrm{EMC} = \frac{1}{n}\mathrm{E}\left\|\mathbf{y}_* - \widehat{\mathbf{y}}\right\|^2.$$

Sean $\mathbf{b} = \mathrm{E}\widehat{\beta} - \beta_0$ el sesgo de $\widehat{\beta}$, y $\mathbf{V} = \mathrm{Var}\left(\widehat{\beta}\right)$. Entonces se deduce que

$$\mathrm{EMC} = \sigma^2 + \frac{1}{n}\left(\left\|\mathbf{X}\mathbf{b}\right\|^2 + \mathrm{tr}\left(\mathbf{V}\mathbf{X}'\mathbf{X}\right)\right) \tag{6}$$

donde "tr" es la traza.

Si $\widehat{\beta}$ es el estimador de mínimos cuadrados con todas las variables, es $\mathbf{b} = \mathbf{0}$ y $\mathbf{V} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, por lo que en este caso

$$\mathrm{EMC} = \sigma^2\left(1 + \frac{p}{n}\right). \tag{7}$$

4.2 Un ejemplo

Consideramos ahora (4) con $p = 2$, tal que

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \quad (j = 1, 2)$$

y

$$\frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2} = \rho.$$

Supongo $\beta_0 = (\beta_0, \beta_0)'$. Aquí los dos coeficientes de β_0 son *iguales*, de modo que si buscamos las "variables relevantes" (o "el modelo correcto") debemos quedarnos con las dos. En tal caso, el EMC es (7) con $p = 2$:

$$\text{EMC} = \sigma^2 \left(1 + \frac{2}{n}\right). \tag{8}$$

Ahora considero usar sólo la primera variable; o sea, la regresión (por el origen) de $\mathbf{y}$ sobre las x_{i1}. Esto da

$$\widehat{\beta} = \left(\widehat{\beta}_0, 0\right)' \quad \text{con} \quad \widehat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i x_{i1}}{\sum_{i=1}^{n} x_{i1}^2} = \frac{1}{n}\sum_{i=1}^{n} y_i x_{i1}. \tag{9}$$

Entonces resulta (¿se animan a hacer la cuenta?)

$$\mathbf{b} = \beta_0 \begin{pmatrix} \rho \\ -1 \end{pmatrix}, \quad \mathbf{V} = \frac{\sigma^2}{n} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

por lo que finalmente

$$\text{EMC} = \sigma^2 \left(1 + \frac{1}{n}\right) + \beta_0^2 \left(1 - \rho^2\right). \tag{10}$$

De aquí resulta que (10) es menor que (7) cuando

$$n\frac{\beta_0^2}{\sigma^2}\left(1 - \rho^2\right) < 1. \tag{11}$$

El estimador univariado tiene más sesgo pero menos varianza que el completo. Cuando n es grande, la varianza es chica y por lo tanto el sesgo domina. Cuando la "relación señal-ruido" $|\beta_0|/\sigma$, también el sesgo domina. Cuando $|\rho|$ está cerca de 1 ambas variables tienen aproximadamente la misma información.

Naturalmente, en un caso real uno no sabe si (11) se cumple, pero un buen estimador del EMC puede llevar a la elección adecuada.

Aquí la diferencia entre (8) y (10) es chica; pero puede ser importante cuando hay muchas variables.