MODELO DE REGRESIÓN MÚLTIPLE

Julián de la Horra
Departamento de Matemáticas U.A.M.

1 Introducción

Abordaremos en este capítulo el modelo de regresión lineal múltiple, una vez que la mayor parte de las técnicas básicas han sido presentadas y desarrolladas en el modelo de regresión lineal simple. Por supuesto, también se pueden considerar extensiones a modelos no lineales, mediante transformaciones de las variables, como se indicó en el capítulo anterior. En general, el objetivo de la regresión múltiple es tratar de expresar una variable respuesta (numérica) en función de varias posibles variables explicativas (todas ellas numéricas).

Ejemplos

Podemos estar interesados en expresar el peso de los ejemplares de cierta especie de ave en función de su longitud y de su envergadura.

Podemos estar interesados en explicar el nivel de cierto contaminante en función de la densidad industrial y de las lluvias medias mensuales.

2 Modelo. Hipótesis del modelo

Disponemos de los siguientes elementos para el estudio estadístico:

Una variable respuesta (o dependiente), Y, que será una variable numérica (o cuantitativa): es la variable que se quiere analizar. Formalmente, será una variable aleatoria de tipo continuo.

Varias posibles variables explicativas (o independientes), $X_1, ..., X_k$, que serán variables numéricas (o cuantitativas). Recuérdese que los factores en el modelo de diseño de experimentos eran siempre variables cualitativas.

Finalmente, necesitamos datos. Supondremos que disponemos de n conjuntos de datos:

$$(y_i, x_{1i}, ..., x_{ki}) \qquad \text{para } i = 1, ..., n$$

Por supuesto, sigue siendo absolutamente necesario que los datos vayan unidos en el sentido de que $(y_i, x_{1i}, ..., x_{ki})$ representan los valores de Y, X_1 ,..., X_k en el i-ésimo individuo o unidad muestral.

El modelo de regresión lineal múltiple es de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_j x_{ji} + ... + \beta_k x_{ki} + u_i \qquad \text{para } i = 1, ..., n$$

Significado de los parámetros:

β_0= Valor medio de la variable respuesta cuando $X_1 = ... = X_k = 0$.

Muy a menudo, el parámetro β_0 no tiene una interpretación intuitiva de interés.

β_j= Mide la variación media que experimenta la variable respuesta cuando X_j aumenta una unidad ($j = 1, ..., k$).

La interpretación intuitiva de β_j ($j = 1, ..., k$) siempre es muy interesante.

u_i= Término de error = Efecto adicional debido a otras variables que no se incluyen en el modelo por no ser consideradas relevantes.

Para poder obtener y utilizar herramientas estadísticas que nos permitan tomar decisiones objetivas y razonadas, necesitamos que el modelo se ajuste a unas determinadas hipótesis. Estas hipótesis iniciales del modelo son las siguientes:

Normalidad: Las observaciones Y_i siguen una distribución Normal,
Linealidad: Los valores medios de la variable respuesta dependen linealmente de los valores de $X_1, ...X_k$: $E[Y_i] = \beta_0 + \beta_1 x_{1i} + ... + \beta_j x_{ji} + ... + \beta_k x_{ki}$,
Homogeneidad o igualdad de varianzas (homocedasticidad): $V(Y_i) = \sigma^2$,
Las observaciones son independientes.

Todas estas hipótesis se pueden expresar abreviadamente de la siguiente forma:

$$Y_i \sim N(\beta_0 + \beta_1 x_{1i} + ... + \beta_j x_{ji} + ... + \beta_k x_{ki}; \sigma^2) \quad \text{independientes.}$$

Ausencia de multicolinealidad: No existen relaciones lineales entre las variables explicativas $X_1, ..., X_k$.

La ausencia de multicolinealidad constituye una hipótesis completamente nueva y su significado es el siguiente:
Por un lado, si alguna de las variables explicativas fuera combinación lineal de las otras, el modelo, obviamente, podría simplificarse. Pero eso no es lo más importante. La importancia práctica de exigir ausencia de multicolinealidad procede del hecho de que, si alguna de las variables explicativas está fuertemente correlacionada con otras, se pueden producer distorsiones en los resultados.

Es importante que estas hipótesis iniciales del modelo se cumplan (aproximadamente) para que las conclusiones que obtengamos no sean una barbaridad.

Llegados a este punto, se puede abordar la cuestión de si tenemos suficientes datos (suficiente información muestral) para abordar el análisis estadístico de este modelo. La regla básica para responder a esto es muy fácil de recordar (y de entender): en general, necesitaremos al menos tantos datos como parámetros queremos estimar en el modelo. En este modelo, tenemos:

Número de datos= n
Número de parámetros= k+2

Por lo tanto, necesitamos, al menos, $n = k + 2$ conjuntos de datos.

3 Metodología

La metodología o plan de trabajo que seguiremos en el análisis estadístico de un modelo de regresión múltiple es el siguiente:

(1) Diagnosis de las hipótesis iniciales del modelo.

Al final del capítulo, se indicarán las herramientas estadísticas que se pueden utilizar para llevar a cabo la diagnosis de las hipótesis previas del modelo.

(2) Estimación puntual de los parámetros del modelo.

(3) Intervalos de confianza para estimar los parámetros del modelo.

(4) Contrastes de hipótesis.

(5) Análisis de la varianza.

(6) Evaluación del ajuste proporcionado por el modelo de regresión ajustado.

4 Estimación puntual de los parámetros

La metodología estadística para obtener estimadores puntuales de los parámetros es la siguiente:

Se aplica el método de máxima verosimilitud, y el estimador obtenido se corrige (en caso necesario) para que sea insesgado.

Estimación de $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$:

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

donde X recibe el nombre de matriz de diseño. En realidad, no usaremos el cálculo matricial para obtener estas estimaciones, sino que serán obtenidas por paquetes estadísticos (como el SPSS).

El modelo de regresión ajustado o estimado sería:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_j x_j + ... + \hat{\beta}_k x_k$$

Estimación de σ^2:

$$\begin{aligned} \hat{\sigma}^2 &= S_R^2 = \frac{1}{n-k-1}\sum_i (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-k-1}\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_j x_{ji} - ... - \hat{\beta}_k x_{ki})^2 \end{aligned}$$

El estimador de σ^2, S_R^2, recibe habitualmente el nombre de *varianza residual* y merece algún comentario adicional. El nombre de varianza residual obedece a que es una varianza que calculamos a partir de los residuos de cada dato. El *residuo* de cada dato depende del modelo estadístico que estemos utilizando, pero responde siempre a la misma filosofía:

"Residuo" = "Valor observado" - "Estimación del valor esperado"
$= y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_j x_{ji} - ... - \hat{\beta}_k x_{ki}$

5 Intervalos de confianza

Los estimadores puntuales son interesantes, pero son demasiado rígidos. Cuando decimos que estimamos que el parámetro β_1 vale, por ejemplo, 1,15, lo que estamos diciendo en realidad es que pensamos que vale, aproximadamente, 1,15. La forma en que los métodos estadísticos cuantifican este "aproximadamente" de forma automática y objetiva es a través de los intervalos de confianza.

Aplicando el método de la cantidad pivotal, se obtienen los siguientes intervalos de confianza para estimar β_0, β_1,...,β_k:

$$IC_{1-\alpha}(\beta_j) = \left(\hat{\beta}_j \pm t_{n-k-1;\alpha/2}(\text{error típico de } \hat{\beta}_j)\right)$$

6 Constrastes de hipótesis

En esta sección, vamos a considerar los contrastes de hipótesis necesarios para estudiar la influencia individual de cada una de las presuntas variables explicativas. El tipo de pregunta que nos planteamos es de la siguiente forma:

¿Disponemos de suficiente evidencia muestral para afirmar que X_j tiene una influencia significativa sobre Y? Dado que la posible influencia de X_j desaparecería si su coeficiente β_j se anulase, esto nos lleva a elegir entre las posibilidades $\beta_j = 0$ y $\beta_j \neq 0$ y, por tanto, a un contraste de hipótesis donde:

$H_0 : \beta_j = 0$ (X_j no influye)
$H_1 : \beta_j \neq 0$ (X_j sí influye)

Elegiremos un nivel de significación α para tomar una decisión al final del estudio. Esta decisión la podemos tomar utilizando el intervalo de confianza $IC_{1-\alpha}(\beta_j)$:

Si el valor cero está contenido en $IC_{1-\alpha}(\beta_j)$, aceptamos H_0, y la conclusión es que no hay evidencia estadística para afirmar que X_j tiene una influencia significativa sobre Y.

Por el contrario, si el valor cero no está contenido en $IC_{1-\alpha}(\beta_j)$, rechazamos H_0, y la conclusión en este caso es que disponemos de suficiente evidencia estadística para afirmar que X_j tiene una influencia significativa sobre Y.

7 Análisis de la varianza

En esta sección, vamos a considerar el contraste de hipótesis necesario para estudiar la validez global del modelo. La pregunta que nos planteamos ahora es la siguiente:

¿Disponemos de suficiente evidencia muestral para afirmar que el modelo, globalmente considerado, es válido? Dicho de otra manera, ¿podemos afirmar que el modelo es globalmente (o conjuntamente) explicativo? Ésto nos lleva al siguiente contraste de hipótesis:

$H_0 : \beta_1 = ... = \beta_k = 0$ (el modelo no es conjuntamente válido)
$H_1 :$ Algún $\beta_j \neq 0$ (el modelo sí es conjuntamente válido)

Este contraste de hipótesis, que se conoce también con el nombre de *contraste de la regresión*, se va a abordar mediante la técnica estadística del análisis de la varianza (ANOVA).

La descomposición de la variabilidad o análisis de la varianza en el caso del modelo de regresión lineal es siempre en dos partes (tanto en la regresión lineal simple como en la múltiple):

"Variabilidad total de los datos"= SCT= $\sum_i (y_i - \bar{y})^2$
$= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$
= SCE + SCR

SCE = $\sum_i (\hat{y}_i - \bar{y})^2$ que es la variabilidad asociada al modelo (o explicada por el modelo).

SCR = $\sum_i (y_i - \hat{y}_i)^2$ que es la variabilidad residual (o no explicada por el modelo).

La decisión de aceptar o rechazar H_0 se va a tomar en base al estadístico que se obtiene a partir de este análisis de la varianza:

$$F = \frac{SCE/k}{SCR/(n-k-1)}$$

Este estadístico tiene una distribucion $F_{k;n-k-1}$ (bajo H_0) y, por tanto, la regla de decisión es de la siguiente forma:
Rechazaremos H_0, al nivel de significación α, cuando

$$F = \frac{SCE/k}{SCR/(n-k-1)} > F_{k;n-k-1;\alpha}$$

También podemos alcanzar una decisión razonando con el p-valor o significación de los datos. La manera más sencilla de "interpretar" y utilizar el p-valor es entendiendo el p-valor como el "apoyo que los datos dan a H_0". De este modo:

Si el p-valor$< \alpha$, el apoyo a H_0 es insuficiente, y rechazaremos H_0 (al nivel de significación α).
Si el p-valor$> \alpha$, el apoyo a H_0 es suficiente, y aceptaremos H_0 (al nivel de significación α).

Por supuesto, obtendremos la misma decisión, tanto si trabajamos con el estadístico F como si trabajamos con el p-valor.

Es tradicional, y así lo podemos ver en libros y salidas de ordenador, organizar los cálculos correspondientes a un análisis de la varianza en una tabla: la tabla ANOVA, que suele ser del siguiente tipo:

Suma de cuadrados	G.l.	Varianza	Estadístico
$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	k	$\frac{SCE}{k}$	$F = \frac{SCE/k}{SCR/(n-k-1)}$
$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-k-1$	$\frac{SCR}{n-k-1}$	
$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$		

Finalizamos esta sección con algunos comentarios sobre los diferentes contrastes de hipótesis que abordamos en los modelos de regresión:

(a) En la regresión lineal múltiple, el intervalo de confianza $IC_{1-\alpha}(\beta_j)$ sirve para estudiar la influencia individual de la variable X_j, mientras que el análisis de la varianza (ANOVA) sirve para estudiar la validez global del modelo. Recordemos que, en la regresión lineal simple, las dos técnicas constituían técnicas alternativas (pero equivalentes) para estudiar el mismo contraste: el contraste de la regresión.

(b) En la regresión lineal múltiple, se comprende mejor la importancia de determinar si una variable explicativa tiene una influencia significativa o no sobre la variable respuesta. El motivo es sencillo: se introducirán en el modelo varias presuntas variables explicativas y, posteriormente, los datos se encargarán de decirnos cuáles son realmente relevantes.

(c) Los resultados sobre la influencia individual de cada X_j y sobre la validez conjunta del modelo los utilizaremos también en la diagnosis de las hipótesis iniciales del modelo.

8 Evaluación del ajuste

A partir de los datos, podemos obtener siempre el modelo de regresión ajustado:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_j x_j + ... + \hat{\beta}_k x_k$$

Este modelo, en algunos casos se ajustará muy bien a los datos que tenemos, y en otros casos se ajustará peor. Cuando el ajuste sea bueno, tendremos una cierta seguridad de que ese modelo representa razonablemente bien la relación entre Y y las variables explicativas $X_1, ..., X_k$.

Es útil disponer de alguna medida numérica que nos evalúe, de manera sencilla, si el ajuste es bueno o no. Para hacer ésto, en la regresión lineal múltiple tenemos una herramienta que ya fue introducida en la regresión lineal simple:

Coeficiente de determinación.- Este coeficiente procede del Analisis de la Varianza y tiene una definición y una interpretación muy sencillas:

$$\text{“Coeficiente de determinación”} = R^2 = \frac{SCE}{SCT} \in [0, 1]$$

El significado es obvio: R^2 mide la proporción de variabilidad explicada por el modelo.

Su interpretación también es obvia:

Cuando R^2 toma un valor próximo a cero, la proporción de variabilidad explicada por el modelo es pequeña, y el ajuste es malo.

Cuando R^2 toma un valor próximo a uno, la proporción de variabilidad explicada por el modelo es grande, y el ajuste es bueno.

Finalmente, veamos la estrecha relación que hay entre el coeficiente de determinación R^2 y el valor del estadístico F del análisis de la varianza:

$$F = \frac{SCE/k}{SCR/(n-k-1)} = \frac{SCE}{SCT}\,\frac{SCT}{SCR}\,\frac{n-k-1}{k}$$

$$= \quad R^2 \; \frac{1}{(SCT - SCE)/SCT} \; \frac{n-k-1}{k} = \frac{R^2}{1-R^2} \; \frac{n-k-1}{k}$$

Por supuesto, para $k = 1$ (regresión lineal simple), tenemos:

$$F = \frac{R^2}{1-R^2} \; \frac{n-2}{1} = (n-2) \; \frac{R^2}{1-R^2}$$

9 Diagnosis de las hipótesis del modelo

Como se indicó en la Sección 3 (Metodología), es conveniente hacer una diagnosis de las hipótesis iniciales del modelo: Normalidad, Linealidad, Homogeneidad de Varianzas y Ausencia de Multicolinealidad. Algunos de estos diagnósticos se pueden llevar a cabo nuevamente utilizando unos análisis gráficos sencillos de los residuos. Para llevar a cabo estos análisis gráficos, necesitamos:

¶Los residuos de cada dato, que en este modelo son de la forma:
"Residuo" $= y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_j x_{ji} - ... - \hat{\beta}_k x_{ki}$
¶Los valores pronosticados o estimados para cada dato, que en este modelo son de la forma:
"Valor pronosticado" $= \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_j x_{ji} + ... + \hat{\beta}_k x_{ki}$

Con los residuos y los valores pronosticados podemos hacer un análisis visual de los siguientes gráficos:

(a) Histograma de los residuos.
La hipótesis de Normalidad de los datos será aceptable cuando este histograma muestre un razonable parecido con la curva Normal.

(b) Gráfico de probabilidades normales de los residuos (probability plot).
La hipótesis de Normalidad de los datos será aceptable cuando los puntos del gráfico estén razonablemente cerca de la diagonal del cuadrado.

(c) Gráfico de dispersión de los residuos sobre los valores pronosticados.
La hipótesis de Homocedasticidad (o igualdad de varianzas) de los datos será aceptable cuando la anchura vertical del gráfico de dispersión se mantenga razonablemente constante.
La hipótesis de Linealidad de los datos será aceptable cuando la línea central del gráfico de dispersión sea razonablemente recta.

La gran ventaja de estos análisis gráficos es su sencillez. Sus grandes inconvenientes son que con pocos datos (como suele ser frecuente) no nos dicen prácticamente nada, y aunque dispongamos de muchos datos, las conclusiones son inevitablemente subjetivas (salvo situaciones muy claras que no son demasiado frecuentes con los datos reales).

También es importante poder hacer un diagnóstico previo de la nueva hipótesis de ausencia de multicolinealidad (ausencia de relaciones lineales entre las variables explicativas $X_1, ..., X_k$). Podemos hacer dos tipos de diagnósticos:

(a) Podemos echar un vistazo a los diagramas de dispersión de X_i sobre X_j (para cada par de variables explicativas), con la finalidad de detectar correlaciones fuertes entre ellas. Esto resulta suficiente cuando sólo tratamos con dos variables explicativas, pero no es así cuando tratamos con tres o más variables explicativas.

(b) También podemos comparar los resultados sobre la influencia individual de cada X_j y sobre la validez conjunta del modelo:

¶ Si el modelo resulta globalmente válido o explicativo, y algunas o todas las X_j resultan individualmente explicativas, los resultados son coherentes, y lo único que tendremos que hacer es simplificar el modelo (si es necesario).

¶ Si el modelo no resulta globalmente válido o explicativo, pero algunas o todas las X_j sí resultan individualmente explicativas, los resultados son incoherentes. Esto suele ocurrir como consecuencia de problemas de multicolinealidad, es decir, de la existencia de fuertes relaciones lineales entre algunas de las variables explicativas. En este caso, se hace necesario revisar el modelo con vistas a eliminar alguna de las presuntas variables explicativas.

¶ Lo mismo ocurre si el modelo es globalmente válido, pero ninguna X_j es individualmente explicativa.

10 Extensión del modelo a datos procedentes de dos poblaciones

Seguimos considerando el problema de explicar una variable respuesta Y en función de k variables explicativas, pero supongamos ahora que disponemos de datos procedentes de dos poblaciones A y B, y no estamos seguros de si los datos son homogéneos o no. Las dos poblaciones pueden ser dos ciudades o dos países, pueden ser niños y adultos, pueden ser hombres y mujeres, pueden ser dos subespecies de plantas o animales,...

Para tratar esta situación, tenemos varias posibilidades:

(a) Trabajar con todos los datos como si fueran homogéneos.

Ventaja: Utilizamos todos los datos.

Inconveniente: Si los datos no son homogéneos, el modelo de regresión ajustado puede no servir ni para unos ni para otros.

(b) Trabajar con los datos de A por un lado, y con los datos de B por otro.

Ventaja: Hallaremos un modelo de regresión útil para A y otro para B.

Inconveniente: Estamos trabajando con menos datos para cada población, y con menos datos los resultados son menos fiables.

(c) Trabajar con todos los datos, pero añadiendo una nueva variable (ficticia o *dummy*) que incorpore al modelo la información de que los datos proceden de dos poblaciones no necesariamente homogéneas.

Se recomienda la utilización de esta vía, ya que reúne las ventajas de las dos vías anteriormente descritas. En resumen, haríamos lo siguiente:

Disponemos de n conjuntos de datos sobre Y, X_1,...,X_k, procedentes de dos poblaciones A y B. Definimos una nueva variable (ficticia o *dummy*), X_{k+1}, del siguiente modo:

$X_{k+1} = 1$, si el dato procede de la población A,
$X_{k+1} = 0$, si el dato procede de la población B.

Por lo tanto, nuestro modelo de regresión lineal múltiple sería ahora de la forma:

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \beta_{k+1} x_{k+1,i} + u_i \qquad \text{para } i = 1, ..., n$$

A partir de aquí, se procede como se ha explicado durante todo el capítulo, y una vez efectuado el estudio, se puede tomar una decisión sobre si interesa mantener la variable ficticia. Para esto, consideramos el contraste:

$H_0 : \beta_{k+1} = 0$ (la variable ficticia no es relevante)
$H_1 : \beta_{k+1} \neq 0$ (la variable ficticia sí es relevante)

Si aceptamos H_0, la variable X_{k+1} no tiene una influencia significativa, y eliminaríamos X_{k+1} del modelo, considerando que los datos de A y de B son razonablemente homogéneos.

Si, por el contrario, rechazamos H_0, la variable X_{k+1} tiene una influencia significativa, y no podríamos eliminar X_{k+1} del modelo, ya que los datos de A y de B no son homogéneos.