TECNICAS DE INVESTIGACION DE MODELOS ESTADISTICOS

Por Foster Cady

En cada experimento, el investigador se enfrenta con el problema de descubrir sus resultados en términos cuantitativos. El presente trabajo describe los diferentes orígenes y las etapas básicas en la formulación del modelo estadístico; con un modelo estadístico, cada observación realizada en un experimento puede darse a entender y el error al azar puede señalarse. También, en el trabajo se exponen algunas técnicas para manejar diferentes tipos de modelos y se discuten ciertas complicaciones en esta operación.

¿Qué es un modelo?

Un modelo matemático es una presentación cuantitativa de un fenómeno de la naturaleza. Por ejemplo, se ha encontrado que la altura de los niños es una función lineal de la altura de los padres, o sea: $\mu = \alpha + \beta x$. En economía, una relación que se usa comúnmente es el Cobb-Douglas, $\mu = \alpha X^{\beta}$, donde μ es el valor verdadero de la salida, X es el valor de la entrada y α y β son coeficientes. En agronomía, una relación entre el rendimiento y el fertilizante es la ecuación famosa de Mitscherlich, $\log \alpha - \log (\alpha - \mu) = \beta X$, donde α es el mayor rendimiento posible, μ es el rendimiento verdadero que se obtiene de un nivel dado de X y β es un coeficiente. Nótese que éstos son modelos matemáticos. En cada modelo, la "variable dependiente" es μ, el valor verdadero.

Si se trabaja todo el tiempo con los valores verdaderos, entonces no se tendrían problemas; se medirían varias $\mu's$ para los valores dados de las $X's$ y entonces se resolverían para los coeficientes. Ahora bien, en las ciencias biológicas y en las ciencias sociales, no se puede medir μ, pero se puede medir Y, el valor observado. La relación entre μ y Y es:

$$Y = \mu + \epsilon$$

donde ϵ es un componente al azar.

Así, la Rama de Estadística nació con ϵ. Podemos hablar acerca de un modelo estadístico, que es una descripción cuantitativa de un valor observado. Comprende todas las componentes que son partes del valor observado, incluyendo la ϵ.

Por ejemplo, aquí están seis lápices nuevos. Queremos hacer una descripción de cada valor medido del largo de cada lápiz. Cada valor medido sería el mismo, excepto para el error al azar. El error al azar incluye cosas tales como la falla de la persona al leer los calibres correctamente, la falla de la fábrica al producir lápices exactamente del mismo largo, etc. Entonces Y_i es la suma de dos cantidades:

$$Y_i = \mu + \epsilon_i .$$

Este es un modelo estadístico. Los dos componentes son los únicos que se necesitan para la descripción de un valor observado. Sin embargo, supongamos que se ponen en agua tres lápices durante dos días y tres lápices, no. Después de los

dos días medimos los largos. ¿Cuál es el modelo, o cuál es la descripción para cualquiera de las observaciones?

Podríamos considerar que hay dos poblaciones: una población de los lápices que han estado en agua, y otra, de los que no han estado.

Por consiguiente:

$$Y_i = \mu_i$$

Sabemos que el error al azar está presente; esto es, los tres valores de los lápices en el agua serán diferentes y los tres valores para los otros lápices, también serán diferentes.

Entonces:

$$Y_{ij} = \mu_i + \epsilon_{ij}; \quad i = 1, 2; \quad j = 1, 2, 3,$$

usualmente no nos gusta el modelo en esta forma y escribimos:

$$Y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

o $$Y_{ij} = \mu + \lambda_i + \epsilon_{ij}$$

Usando este modelo, queremos calcular los estimadores y el análisis de la varianza y hacer pruebas de hipótesis.

Obsérvese el orden de los pasos. El experimento → el modelo → la estimación de los parámetros → el análisis de la varianza → las pruebas de la hipótesis. El orden es muy importante. El modelo depende del experimento. El análisis depende del modelo. Hay una correspondencia de uno a otro entre los pasos. Por ejemplo, el análisis tiene que estar de acuerdo con el modelo: ellos son la misma cosa.

Podemos escribir el modelo en la forma de regresión múltiple:

$$Y_i = \mu X_{oi} + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \epsilon_i$$

o $$= \beta_o X_{oi} + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Es conveniente usar modelos que puedan expresarse en la forma del modelo de regresión múltiple, porque sabemos cómo manejar este tipo de modelo. Usamos el método de los mínimos cuadrados para encontrar los estimadores. Nótese que en este modelo, los parámetros son lineales. En la Teoría Estadística, sabemos mucho acerca de las propiedades de los estimadores y cómo probar las funciones lineales de los mismos.

Hay otra razón por la cual nos gusta este tipo de modelo: es muy flexible. Por ejemplo, es posible tener el siguiente modelo:

$$Y_i = \beta_o X_{oi} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$$

donde

$$X_{3i} = X_1^2, \; X_{4i} = X_{2i}^2 \qquad \text{y} \qquad X_{5i} = X_1 X_2$$

De hecho, cualquier cosa puede usarse como X y el modelo de regresión dará la relación lineal entre Y y X ajustada para los otros factores en el modelo.

¿Como se escribe el modelo? De maneras diferentes

1.—*De un experimento diseñado.* Por ejemplo, en un diseño bloques al azar, sabemos que los factores importantes son los tratamientos y las estratificaciones del material experimental.

2.—*De la teoría en algunos de los campos.* Por ejemplo, Mitscherlich, basado en observaciones en un invernadero, presentó la teoría de que la respuesta a un aumento dado de fertilizante era proporcional a la disminución del rendimiento máximo. Mitscherlich formuló esta ecuación diferencial:

$$\frac{\alpha\,\mu}{\alpha\,X} = \beta\,(\alpha - \mu)$$

De esta ecuación, el modelo dado anteriormente se formula por integración.

3.—*De los datos.* Unas veces la teoría no existe y otras el investigador es un poco perezoso. Por consiguiente, se coleccionan muchos datos antes de fijarse en la pregunta: ¿Cuál es el modelo? La decisión se toma poniendo todas las variables en un programa de regresión múltiple y se tiene el modelo.

Los problemas en la construccion y uso del modelo

1.—*Modelos no-lineales.* Unas veces, el modelo que se formula de la teoría en un campo resulta un modelo con un parámetro que no es lineal. Por ejemplo, en la función Cobb-Douglas, $Y = \alpha\,X^{\beta}$, nótese que el parámetro es un exponente. Por consiguiente, no es posible estimar los parámetros en la manera que él acostumbra, o sea por el método de los mínimos cuadrados. Hay en la literatura otros caminos para la estimación, pero hay problemas en estimar la varianza del parámetro no-lineal.

Usualmente, se toma el log de ambos miembros:

$$\log Y = \log \alpha + \beta \log X$$

ahora tenemos una ecuación lineal donde el $\log \alpha$ es la interacción y β es la pendiente de la relación lineal entre $\log Y$ y $\log X$.

Sin embargo, para probar una hipótesis acerca de los estimadores, necesitamos suponer que el error es aditivo en el modelo logarítmico, lo cual quiere decir que el error es multiplicativo en el primer modelo; esto es, $Y = \alpha\,X^{\beta}\,E$. Un "error multiplicativo" quiere decir que el error es proporcional a Y. Esta es una suposición que puede ser correcta en algunos casos, pero no aplicable a todos.

2.—*El error para probar la significación de los estimadores.*

2a.—*Ejemplo.* Un experimento en que se usa un diseño bloques al azar para varias localidades. El análisis para cada localidad es:

Bloques	$(b-1)$
Tratamientos.	$(t-1)$
Error	$(b-1)\,(t-1)$

Si las localidades se combinan:

	g.l.	*E.C.M.*
Localidades (L).	$(1-1)$	
Bloques $/L$	$1(b-1)$	
Tratamientos	$(t-1)$	$\sigma^2 + b\,\sigma^2_{\lambda L} + bl\,\sigma^2\,\lambda$
$T \times L$	$(1-1)\,(t-1)$	$\sigma^2 + b\,\sigma^2_{\lambda L}$
Error	$1(b-1)\,(t-1)$	σ^2

Cuando los datos se manejan con el análisis de la varianza, no hay problema en saber cuál es el uso correcto para probar. Los tratamientos se prueban contra $T \times L$ y la interacción se prueba contra el error. Sin embargo, cuando el análisis se hace empleando un programa de regresión múltiple, el programa normalmente usará las desviaciones de regresión como error y probará todos los estimadores con el mismo error.

2b.—Muchas veces al investigador se le olvida que las desviaciones de la regresión, usando el programa de regresión múltiple, son una mezcla de tipos diferentes del error, es decir, error experimental y error de muestreo. En un análisis de la varianza, el investigador generalmente separará estos dos tipos de error y usará el error experimental para probar. Pero en muchos casos, usando el programa de regresión múltiple con un modelo, usará las desviaciones de la regresión, que constituyen principalmente error de muestreo, y no pensará que usando este error resultarán más estimadores significativos de los que hay realmente.

3.—*Usando los datos para encontrar el modelo. Una situación.* Un investigador tiene unos datos y su escuela tiene una computadora. Sus datos incluyen una Y y $20X's$. Entonces, sin pensar, se usa este modelo:

$$Y_i = \beta_o\,X_o + \beta_1\,X_1 + \beta_2\,X_2 + \ldots + \beta_{2o}\,X_{2o} + E_i$$

Este es el primer problema. Este modelo dice que la relación entre Y y cada X es lineal. Quizás X_{10} es una variable importante pero la relación no es lineal.

Por consiguiente, el primer paso es hacer muchos diagramas para que las relaciones no lineales y las interacciones entre las variables puedan encontrarse. El estadístico puede ayudar, pero solamente el investigador puede describir el modelo.

Después de este paso, todavía hay muchos problemas. Usualmente las variables X no son independientes en el sentido estadístico. Entonces, los estimadores de los parámetros en el modelo a veces tienen valores que no son racionales. Por ejemplo, el investigador piensa que la relación entre Y y X_8 sería positiva, pero el estimador b_8 es negativo debido a las interrelaciones entre las variables $X's$. El investigador tiene que recordar que las veinte variables deben considerarse en conjunto y es muy difícil interpretar una variable aisladamente. La interpretación correcta de una b es la relación entre Y y X ajustada por las otras variables en el modelo. Nada más.

Supongamos que el investigador quiere empezar con las 20 variables, pero eliminando las que no son importantes. Entonces, hace muchos modelos diferentes.

Hay varios métodos para hacer esta reducción en el número de las variables. Un artículo reciente de Biometrics (junio de 1966, pág. 268), compara cuatro métodos. Todos éstos tienen el problema de que, después de tratar muchos modelos, los niveles de significación para probar los estimadores en cada modelo no son los niveles escogidos de 5 o 1%, porque el mismo conjunto de los datos se usa para estimar los parámetros en cada modelo.

Una solución es usar la mitad de los datos para encontrar el modelo reducido; entonces, la otra mitad se usa para probar la significación estadística de los estimadores en el modelo final.

4.—*Al probar que un cierto modelo es suficiente. Una situación.* Un investigador quiere probar que un modelo cuadrático es suficiente para describir la relación entre Y y X_1 y X_2, por ejemplo, nitrógeno y fósforo. El modelo es:

$$Y_i = \beta_o X_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon_i$$

El número mínimo de las combinaciones de tratamientos es seis. Sin embargo, con sólo seis observaciones o con repetición de las mismas, el modelo ajusta los datos exactamente.

El investigador debe tener más combinaciones de tratamientos para que haya oportunidad de que el modelo no ajuste los datos exactamente.

Las repeticiones son necesarias para una estimación del error experimental. Supongamos 15 combinaciones de tratamientos usados con dos repeticiones. El análisis de la varianza es:

Modelo (no incluyendo la media)	5
Falta de ajuste	9
Error	15

Si la relación F de falta de ajuste contra el error no es significativa, entonces el modelo cuadrático es suficiente.

Referencias citadas

ANÓNIMO. 1961. *Status and methods of research in economic and agronomic aspects of fertilizer response and use.* Publication 918. National Academy of Sciences. National Research Council Washington, D. C.

GRAYBILL, FRANKLIN. 1961. *An introduction to linear statistical models.* McGraw-Hill Co.

HEADY, EARL O. & JOHN L. DILLON. 1961. *Agricultural production functions.* Iowa State University Press, Ames, Iowa.

MASON, DAVID D. 1956. *Functional models and experimental designs for characterizing response curves and surfaces.* Chapter 5 in "Methodological procedures in the economic analysis of fertilizer use data." Iowa State University Press, Ames, Iowa.

WEINER, JOHN M. & OLIVE JEAN DUNN, 1966. *Elimination of variates in linear discrimination problems.* Biometrics, Vol. 22: 268-275.