

PLENARIA

MODELAMIENTO ESTADÍSTICO DE DATOS EXPERIMENTALES TEORÍA BÁSICA CON APLICACIONES

Eduardo Dávila

M.Sc. en estadística, candidato a Ph.D.

jedavilas@unal.edu.co

Camilo Niño

Ingeniero Agrónomo

Camilo.nino@talex.com.co

Resumen

En esta ponencia se presenta una sinopsis de la metodología relacionada con el análisis de datos provenientes de experimentos. Después de definir el concepto básico de modelamiento, los autores presentan los pasos esenciales que todo investigador debe tener en cuenta para un correcto análisis de datos, cuando se hace uso de métodos paramétricos. Mediante un ejemplo, se presentan los errores que pueden cometerse, cuando no se validan los supuestos asociados a cada modelo estadístico. Finalmente, el lector contará con las herramientas mínimas para que el modelo que busque ajustar tenga validez en la comunidad científica internacional.

Palabras clave: Modelos Estadísticos, Validación de Supuestos, Selección del Modelo e Inferencia.

Abstract

This paper presents an overview of the methodology related to data analysis from designed experiment. After defining basic concept of modeling, the authors present the essential steps that all researchers must take into account for proper analysis, when dealing with the use of parametric methods. Through one example, it is presented the typical errors committed once the assumptions are not validated in relation to the selected statistical model. Finally, the reader will have the minimal tools for modeling with acceptance in the scientific community.

Key Words: Statistical Models, Verify Assumptions, Model selection and Inference.

INTRODUCCIÓN

Existe en la literatura muchos textos sobre modelos estadísticos, que ofrecen alternativas para el análisis de diferentes tipos de datos, obtenidos mediante encuestas o por medio de experimentos diseñados para tal fin; en el primero o segundo caso, la mayoría de autores dedican sólo unos cuantos renglones al proceso inicial que busca responder cómo escoger el correcto método estadístico de análisis. Según Crawley (2007), la parte más difícil del proceso es iniciar, lo cual incluye la dura tarea de atreverse a escoger algunos tipos de modelos. En términos generales, se puede organizar el proceso de modelamiento estadístico en las siguientes partes consecutivas:

- a) seleccionar el tipo correcto de método de análisis estadístico,
- b) de cada método seleccionado, determinar los valores más apropiados de los parámetros que lleven a proponer el “mejor” modelo que se ajuste a los datos,
- c) verificar el modelo con base en los supuestos asumidos,
- d) inferir.

La primera parte depende del tipo de datos y del contexto científico; la segunda está en el esquema de la selección del modelo, asociado al término muy usado de “bondad de ajuste”; el tercer aparte va adherido a la definición de modelo estadístico; finalmente, la inferencia es el fin mismo del modelamiento que lleva a la verificación de hipótesis científicas, que son la base sobre las que se construye la teoría científica y, por tanto, el conocimiento formal.

Con lo anterior, el objetivo de esta ponencia es revisar cada uno de las partes que llevan a un correcto

modelamiento estadístico. El artículo se presenta en tres secciones, estando la Sección 1 dedicada a la revisión formal del modelamiento, partiendo de la definición del modelo estadístico; la Sección 2 muestra una aplicación, paso a paso, haciendo uso del programa R de libre distribución; en la tercera Sección se hace una breve reflexión sobre la evolución del modelamiento y se concluye.

CONCEPTOS BÁSICOS DEL MODELAMIENTO ESTADÍSTICO

Después de presentar una definición de modelo estadístico, se revisarán los pasos formales que llevan a un apropiado análisis paramétrico de datos.

Definición de Modelo Estadístico

De forma general un modelo estadístico es una familia de distribuciones sobre un espacio muestral; formalmente, un modelo estadístico es una tripleta:

$$M = (\Omega, F, P_{\theta \in \mathcal{T}}),$$

Donde Ω es un espacio muestral, F es una sigma-algebra y $P_{\theta \in \mathcal{T}}$ es una familia de distribuciones, dentro de la cual se espera esté la densidad que genera los datos y que se asocia a la variable aleatoria Y (Liese and Miescke, 2008). Una definición bien elaborada de un modelo particular se presenta a continuación.

Definición 1. Modelos propios de dispersión $PD(\lambda, \mu)$: es una familia de distribuciones cuya densidad, con respecto a la medida de Lebesgue, está dada por

$$p(y; \mu, \lambda) = c(\lambda) V^{-\frac{1}{2}}(y) \exp\left\{-\frac{\lambda}{2} d(y, \mu)\right\}, y, \mu \in \Theta \subseteq \mathbb{R}$$

con $c(\cdot) \geq 0$, función conocida y $V(\cdot)$ la función de varianza.

Dentro de los supuestos incluye una variable aleatoria continua y una función de varianza con ciertas propiedades que caracteriza la familia; para detalles véase (Jørgensen, 1997).

El modelamiento estadístico es un proceso algorítmico que tiene como objetivo primordial seleccionar el modelo más apropiado, que en términos prácticos será aquel que genere la menor cantidad de variabilidad no explicada, sujeto a la restricción que todos sus parámetros sean estadísticamente significativos; además, el modelo buscado deberá cumplir el *principio de parsimonia*, que en términos prácticos significa, según Crawley (2007), que:

- deberá tener la menor cantidad de parámetros posibles,
- de preferencia estará en la clase de los modelos lineales y no en los no lineales,
- estará basado en pocos supuestos,
- iniciará en un grupo de modelos complejos para irse simplificando hasta un nivel adecuado,
- se asociará a explicaciones simples en menor grado que a complejas.

Entendidos el objetivo y los principios del modelamiento, a continuación se estudiará cada parte de dicho proceso.

Etapas del Modelamiento Estadístico

Antes de iniciar con el proceso de modelamiento propiamente dicho, es importante seleccionar el método estadístico apropiado y luego, dentro del método, buscar los modelos mejor indicados. En cuanto al método a usar, debe haber claridad sobre qué tipo de variable respuesta se va a analizar (continua, conteo, proporción, binomial, categórica, tiempo de sobrevida, etc.) y cuál es la naturaleza de las variables explicativas (continuas, categóricas o mezcla de ambas); Dobson (2002) ofrece una buena referencia de métodos de análisis con base en los tipos de variables respuesta y categóricas. Después de un correcto análisis exploratorio de los datos, las etapas del modelamiento deben darse en el orden que se presentan a continuación.

(i) *Formulación del modelo.*

En esta primera etapa se debe relacionar una función de la variable respuesta Y con funciones de las variables explicativas (X_1, \dots, X_p) , por medio de una ecuación, más una distribución de probabilidad para la variable respuesta, en conjunto con los supuestos necesarios para poder considerar que los datos han sido generados por este mecanismo probabilístico.

(ii) *Selección del modelo.*

En esta fase es importante recordar (Crawley, 2007) que el modelo estadísticos se puede organizar dentro de cinco tipos.

- El modelo nulo.* Tiene un solo parámetro, la media general μ , tiene el máximo número de grados de libertad $(n-1)$, no tiene ajuste ni poder explicativo.
- El modelo minimal adecuado.* Es un modelo simplificado cuyo número de parámetros (p^*) está delimitado por $0 \leq p^* \leq p$; su ajuste es menor, aunque significativamente igual con respecto al modelo maximal; tiene $n - p^* - 1$ grados de libertad y su poder explicativo está determinado por la relación entre la variabilidad explicada y la variabilidad total (R^2).
- El modelo maximal.* Contiene todos los p factores, que incluyen interacciones y covariables de interés; sus grados de libertad son $n - p - 1$; su poder explicativo depende de cada caso.
- El modelo saturado.* Contiene un parámetro para cada dato, su ajuste es perfecto, no tiene grados de libertad y no tiene poder explicativo.

Claramente, el objetivo de la selección del modelo es encontrar *el modelo minimal adecuado* a cada situación. El orden lógico será ajustar el *modelo maximal* e iniciar con la eliminación de los términos menos significativos; el proceso se detiene al lograr el modelo con la mayor variabilidad explicada y con el menor número de parámetros. En el ajuste del *modelo maximal* se deberán chequear los supuestos hechos en la etapa de formulación; usualmente en este punto es donde los usuarios de los métodos estadísticos cometen más fallas.

Para la selección del *modelo minimal*, el uso de los *Criterios de Información* ha sido bastante generalizado (Claeskens and Hjort, 2008); de éstos, el Criterio de Información de Akaike (**AIC**), es el más empleado, cuya expresión tiene la forma

$$AIC(M) = 2l_{max}(M) - 2 \dim(M), \quad (1)$$

para cada modelo candidato (M). En (1) $l_{max}(M)$ es el logaritmo de la verosimilitud de cada modelo y $\dim(M)$ es la longitud del vector de parámetros. Claramente el **AIC** es una verosimilitud (logaritmo de) penalizada por el número de parámetros que se incluyen en M . Al igual que en modelamiento, el uso del **AIC** está limitado a modelos tradicionales que incluyen el supuesto de variables independientes e idénticamente distribuidas.

Existen razones matemáticas muy precisas, en cuanto la relación que hay entre el **AIC** y la distancia de Kullback-Leibler, que hacen del uso de este criterio un tanto riesgoso cuando no hay independencia entre observaciones o cuando el modelo que se asume para los datos $f(y; \theta_0)$ está lejos (con respecto a la distancia de Kullback-Leibler) de la “verdadera” densidad que genera los datos $g(y)$; casos típicos se tienen en los datos correlacionados, la sobredispersión en modelos asociados a la familia Poisson y binomial o cuando no se cumplen los supuestos de normalidad en errores, entre otros (Dávila and López, 2010). Cuando no se tienen los requisitos necesarios para aplicar el **AIC**, una buena referencia que incluye extensiones de este criterio es Claeskens and Hjort (2008). Entre las alternativas más empleadas está el Criterio de Información de Takeuchi (**TIC**), cuya penalización es la traza del producto de la matriz de la varianza de los Scores y la inversa de la matriz de la Información de Fisher esperada. En el caso muy usual de sobredispersión, en modelos lineales generalizados (familias Poisson y binomial), se tiene el

$$AICc = 2l_{max}(M) - 2 \dim(M) (1 + d),$$

donde d es el parámetro de sobredispersión. Finalmente, el Criterio de Información de Schwarz o Bayesiano (**BIC**) ofrece una penalización más severa a la complejidad del modelo, éste toma la forma

$$BICc = 2l_{max}(M) - \log(n) \times \dim(M).$$

(iii) Revisión del modelo.

Es importante recalcar que muchos profesionales fallan al atarse a una clase de modelos, luego es importante recordar lo citado en Crawley (2007), esto es:

- Todos los modelos son malos,
- Algunos son mejores que otros,
- El modelo correcto nunca se podrá reconocer con certidumbre total,
- Lo simple que sea el modelo lo hace preferible.

Además, la revisión del modelo seleccionado deberá ser estricta en cuanto a:

- Detectar predicciones muy pobres,
- Diagnosticar heterocedasticidad,
- Mostrar falta de normalidad en los errores,
- Ser fuertemente influenciado por escaso número de datos,
- Mostrar algún patrón sistemáticos en los errores,
- Exhibir sobredispersión.

Existen varios métodos, de los cuales los gráficos tienen muy buena aplicabilidad y son de fácil entendimiento.

Entre los principales gráficos se pueden citar los siguientes:

- Residuales contra valores ajustados=> en busca de heterocedasticidad,
- Residuales contra covariables=> en busca de falta de linealidad,
- Residuales contra el orden de recolección=> en busca de correlación temporal,
- Residuales contra desvíos normales estandarizados=> buscando falta de normalidad.

En la Tabla 1 se presentan algunos de los supuestos que se deben verificar en modelos conocidos. Se recuerda que la prueba de Bartlett es usada para diagnosticar heterocedasticidad y la de Shapiro-Wilks para normalidad, entre otras.

Tabla 1. Algunos supuestos que deben validarse para modelos conocidos.

Clase de Modelos	Supuestos a verificar
Análisis de Varianza Clásico	Normalidad, homocedasticidad, linealidad
Regresión Lineal	Varianza constante y normalidad de errores
Lineal Generalizado	Dispersión nominal, linealidad en la función de enlace.
Copulas	Continuidad en las marginales.

(iv) Inferencia.

Una vez revisado el modelo, el investigador procederá a resolver los problemas de decisión estadísticas que motivaron el uso del modelo, ya sea prueba de hipótesis, estimación puntual, estimación por intervalo, etc.

APLICACIÓN

El caso a analizar corresponde a datos no publicados, propiedad de la empresa **TALEX SAS**, recolectados en un ensayo en el cultivo de rosas, en la sabana de Bogotá. El objetivo del trabajo fue evaluar el efecto de un producto nutricional y profiláctico (**KlingQuel® Raíces**) en la población de ácaros (**Tetranychus**). Para esto, se montó un diseño en bloques completos generalizados (cuatro bloques), con dos Productos (**KlingQuel® Raíces** y Testigo); cada tratamiento o Producto tuvo 16 subrepeticiones, el total de observaciones fueron 128.

Para efecto didáctico, se seguirá paso a paso la metodología descrita, haciendo uso de dos clases de modelos: análisis de varianza clásico (ANOVA) y modelo lineal generalizado (GLM) y sus extensiones.

Formulación de los Modelos

Modelo ANOVA. Considérese el arreglo:

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{(ij)} + e_{ijk}, \quad i = 1,2, \quad j = 1,2,3,4, \quad k = 1,2, \dots, 16,$$

donde μ es la media global, τ_i es el efecto del i -ésimo tratamiento, β_j es el efecto del j -ésimo bloque, $\tau\beta_{(ij)}$ es el efecto del submuestreo y e_{ijk} es el error que se distribuye normal e independiente, esto es $e_{ijk} \sim \text{NID}(0, \sigma^2)$.

Modelo GLM. Considérese que la variable aleatoria sigue una distribución de Poisson, con parámetro $\lambda \in (0, \infty)$, además se tiene el arreglo:

$$\log E(Y_{ij}) = \mu + \tau_i + \beta_j + \tau\beta_{(ij)}, \quad i = 1, 2, \quad j = 1, 2, 3, 4;$$

siendo $E(Y_{ij}) = \text{Var}(Y_{ij})$.

Selección de los Modelos

Al construir los modelos maximales, en cada caso se obtuvo lo siguiente.

Modelo ANOVA: no se cumple con el supuesto de normalidad, luego se hace transformación de Box & Cox, la potencia seleccionada es cero, entonces se hace uso de logaritmo natural. Calculando el **AIC**, su máximo valor (313,67) sugiere el siguiente modelo que se presenta con su tabla de ANOVA.

Tabla 2. Análisis de varianza, variable respuesta ácaros por tallo.

Fuente	GL	Sum. Cuad.	Cuad. Medio	F Calc.	p-valor
Bloques	3	5.185	1.7284	2.7472	0.04593 *
Productos	1	18.129	18.1291	28.8142	3.948e-07 ***
Bloque:Productos	3	14.925	4.9749	7.9071	7.408e-05 ***
Error	120	75.501	0.6292		

El *modelo minimal* coincide con el *modelo maximal*.

Modelo GLM: el supuesto de dispersión nominal no se cumple al ajustar el modelo maximal, ya que hay un desvío residual de 598,8 sobre 120 grados de libertad. De esto se derivan varias alternativas, una son los modelos de Cuasi-verosimilitud, la otra usar una función de varianza tipo binomial negativa. En el primer caso se tiene el siguiente *modelo minimal*, como se registra en la Tabla 3.

Tabla 3. Análisis de quasi-verosimilitud Poisson, variable respuesta ácaros por tallo.

Modelo	GL	Desvío	GL resid.	Desvío Resid.	P(> Chi)
Nulo	127	791,97			
Producto	1	71.03	126	720.96	0.0004902***
Producto:Bloque	6	122.153	120	598.80	0.0019084 **

Es claro que bajo Cuasi-verosimilitud Poisson, el efecto simple de los Bloques no es significativo, luego el *modelo minimal* es un caso especial de *modelo maximal*.

Por la otra ruta, al hacer uso de la familia binomial negativa, se obtiene un resultado similar al del ANOVA, pues el *modelo minimal* sugerido por el **AIC** (748,3) es igual al *modelo maximal* (Tabla 4).

Tabla 4. Análisis de desvíos, bajo familia binomial negativa, variable respuesta ácaros por tallo.

Modelo	GL	Desvío	GL resid.	Desvío Resid.	P(> Chi)
Nulo	127	192.07			
Producto	1	13.639	126	178.43	0.0002216 ***
Bloque	3	10.124	123	168.31	0.0175370 *
Producto:Bloque	3	20.670	120	147.64	0.0001233 ***

Revisión de los Modelos e Inferencia

En el caso del modelo ANOVA, las pruebas de homogeneidad de varianzas (Bartlett, p -valor= 0,62), normalidad de errores (Shapiro, p -valor=0,28) y aditividad (gráfico de residuales contra covariables) indican que los supuestos del ANOVA se cumplen, luego la inferencia lleva a afirmar que hay efectos de los tratamientos, lo que se confirma con la prueba de Tukey que da una diferencia altamente significativa (p -valor= $4e-07$) entre **KlingQuel® Raíces** (4,97 ácaros/tallo) y el Testigo (8,85 ácaros/tallo); el resultado propone recomendar el producto en la disminución de la población de ácaros.

Sobre los modelos de Cuasi-verosimilitud Poisson y binomial negativo, ambos presentan buenos diagnósticos en cuanto al análisis gráfico; el binomial negativo compromete ligeramente la linealidad en la transformación, pero sin ser aberrante; ambos se consideran apropiados sobre la base de sus supuestos; además los dos modelos reportan efecto altamente significativo de los Productos.

DISCUSIÓN Y CONCLUSIÓN

Se usaron tres tipos de modelos para analizar un mismo conjunto de datos, los cuales llevaron a universos inferenciales ligeramente diferentes. Es importante recalcar que la variable respuesta analizada, población de ácaros, tiende a estar asociada a conteos de contagio, en los cuales la varianza es muy superior a la que el modelo Poisson sugiere; esto lleva al fenómeno conocido como sobredispersión, que fue diagnosticado en el modelamiento. Al hacer el ANOVA sin transformación de Box & Cox, violando los supuestos, se llega a la misma inferencia del modelo de Cuasi-verosimilitud, rechazando el efecto significativo de los Bloques sobre la variable respuesta (p -valor= 0,222), lo que permite observar, para el caso analizado, que el uso del modelo en el que se debilita el supuesto de la distribución, está llevando a una inferencia incorrecta.

La ausencia de independencia univariada y, en la mayoría de casos, multivariada, está forzando el desarrollo de nuevos modelos estadísticos, para el análisis de variables no normales y no independientes, cuya transformación en busca de normalidad no resuelve satisfactoriamente el problema del modelamiento de la varianza de los estimadores (Dávila and López, 2010).

Uniendo la revisión teórica con los resultados de la aplicación, se puede concluir que el seguimiento al proceso formal de modelamiento, lleva a inferencia más acorde al universo de aplicación, aun haciendo uso de clases diferentes de modelos como fueron el ANOVA clásico y la familia binomial negativa. La validación de los supuestos y la revisión general del modelo asumido, aseguran la generación de conocimiento científico veraz y en los profesionales de la estadística está la responsabilidad para que esto se cumpla.

BIBLIOGRAFÍA

- Claeskens, G. and Hjort, N. (2008). *Model Selection and model Averaging*. Cambridge: Cambridge University Press.
- Crawley, M. (2007). *The R Book*. New York: Wiley.
- Dávila, E. and López, L. (2010). Modeling Multivariate Overdispersed Binomial Data. En: *International Biometrics Conference*. XXV International Biometric Conference. Florianópolis, Brazil 5-10 Dec 2010. The Brazilian Region (RBras) and the Argentinean Region (RArg).
- Dobson, A. (2002). *An Introduction to Generalized Linear Models*. 2 ed. London: Chapman and Hall.
- Jørgensen, B. (1997). *Dispersion Models*. London: Chapman and Hall.