Modelo lineal general II

Alvaro José Flórez

[1]Escuela de Estadística
Facultad de Ingeniería

Marzo - Julio 2023

1. Recordemos...
2. Ejercicio
3. Lo que viene...

Pasos en la modelación

1. Especificación del modelo (relación entre la variable respuesta y las covariables).
2. Estimación de los parámetros del modelo.
3. Evaluación de la adecuación del modelo (cumplimiento de los supuestos) e identificación de atípicos y/o puntos influyentes.
4. Inferencia de acuerdo con los objetivos.

En caso de tener problemas en el paso (3), se puede proponer otro modelo (por ejemplo, usando transformaciones).

Modelo de regresión múltiple

Modelo:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{p-1,i}\beta_{p-1} + \varepsilon_i$$

Donde:

- $E(\varepsilon_i) = 0$.
- $V(\varepsilon_i) = \sigma^2$, para $i = 1, \ldots, n$.
- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, para $i \neq j$.
- $\varepsilon_i \sim$ Normal.

Matricialmente:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ donde } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Modelo de regresión múltiple

La estimación de $\boldsymbol{\beta}$ se hacer por medio del estimador por MCO:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Por lo cuál, se requiere que las columnas de $\boldsymbol{X}$ no sean linealmente dependientes.

Si los supuestos del modelo se cumplen, se tiene que:

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Ademas, el estimador por MCO es el mejor estimador lineal insesgado (teorema de Gauss-Markov).

Supuestos del modelo linea múltiple

El incumplimiento de los supuestos incide en las propiedades de los estimadores por MCO.

Si no se cumple el supuestos (1) se obtienen estimaciones sesgadas.

Si no se cumplen (2) y (3), los estimadores MCO pierden la condición de optimalidad.

Si no se cumple (4) se pierde eficiencia y puede imposibilitar la aplicación de inferencias basadas en normalidad.

Supuestos del modelo

Los supuestos del modelo se pueden verificar de forma gráfica (residuos estudentizados, residuos parciales, ...) o por medio de pruebas formales (White, Durbin-Watson, Shapiro-Wilks, ...).

En caso de incumplirse algún supuesto, se pueden hacer transformaciones sobre las variables (método Box-Cox) o o considerar otro tipo de estimador (por ejemplo, MCP).

Puntos atípicos e influyentes

Puede que algunas pocas observaciones (atípicas y/ influyentes) afecten dramáticamente el modelo (estimaciones, pruebas de hipótesis, supuestos). Por lo cuál es importante su identificación.

Estos puntos pueden ser detectados de forma gráfica y/o usando algunos indicadores (Distancia de Cook, DFBETAS, DFFITS, COVRATIO).

Ejemplo 1

Objetivo: ajustar un modelo para predecir del % de grasa corporal (proceso complejo).

$n = 252$ hombres.

Variables observadas:

variable	Descripción	variable	descripción
`age`	edad (años)	`weight`	peso (libras)
`height`	altura (pulgadas)	`neck`	circunferencia del cuello (cm)
`chest`	circunferencia del pecho (cm)	`abdom`	circunferencia del abdomen (cm)
`hip`	circunferencia de la cadera (cm)	`thigh`	circunferencia del muslo (cm)
`knee`	circunferencia de la rodilla (cm)	`ankle`	circunferencia del tobillo (cm)
`biceps`	circunferencia del bíceps ext. (cm)	`forearm`	circunferencia del antebrazo (cm)
`wirst`	circunferencia de la muñeca (cm)		

`Brozek`**(Respuesta):** % de grasa corporal medido con precisión mediante una técnica de pesaje bajo el agua.

Datos: `data(fat)` de la librería `faraway`

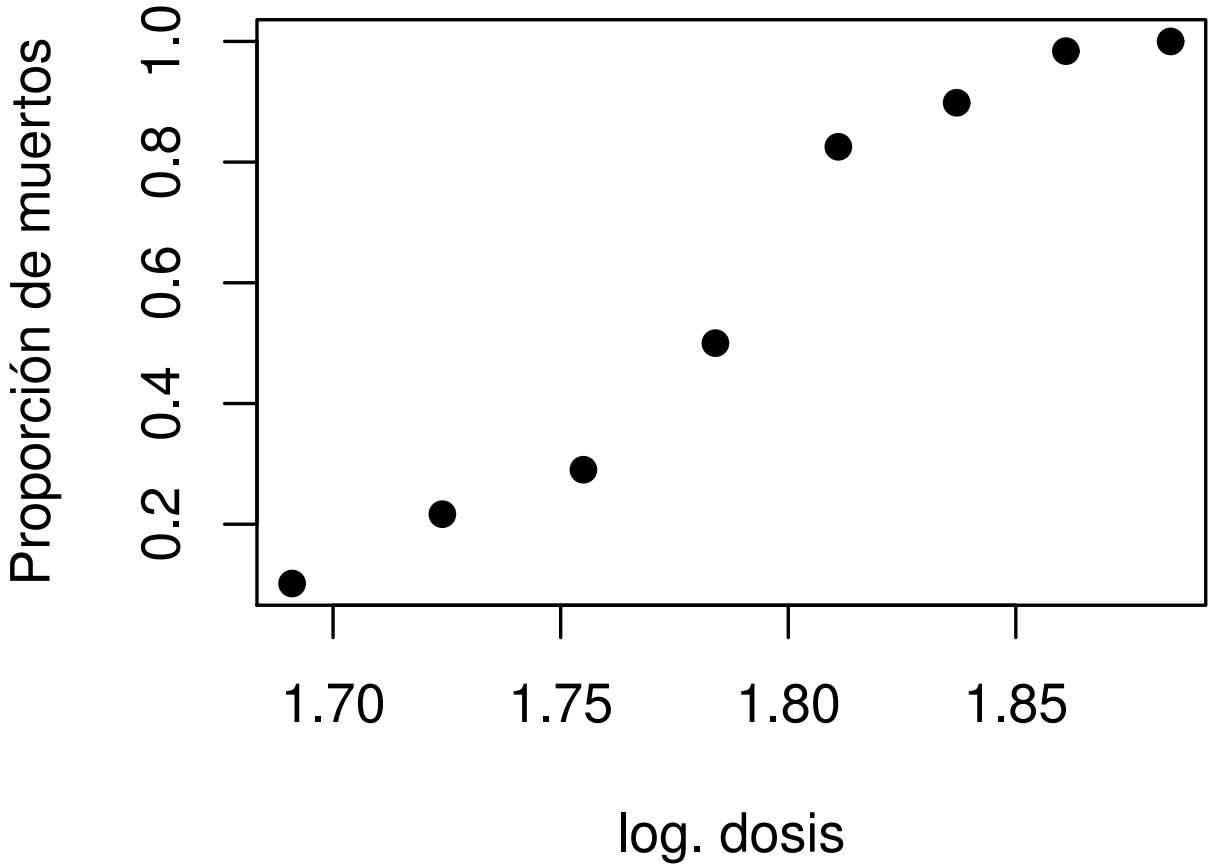
Ejemplo 2

Número de escarabajos muertos después de cinco horas de exposición a disulfuro de carbono gaseoso ($CS_2 mgl^{-1}$) en diversas concentraciones

log. dosis	Número de escarabajos	
	total	muertos
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

¿Hay una relación entre la dosis y la mortalidad de escarabajos?

Mortalidad de escarabajos



Ejemplo 3

Datos: `Puromycin`.

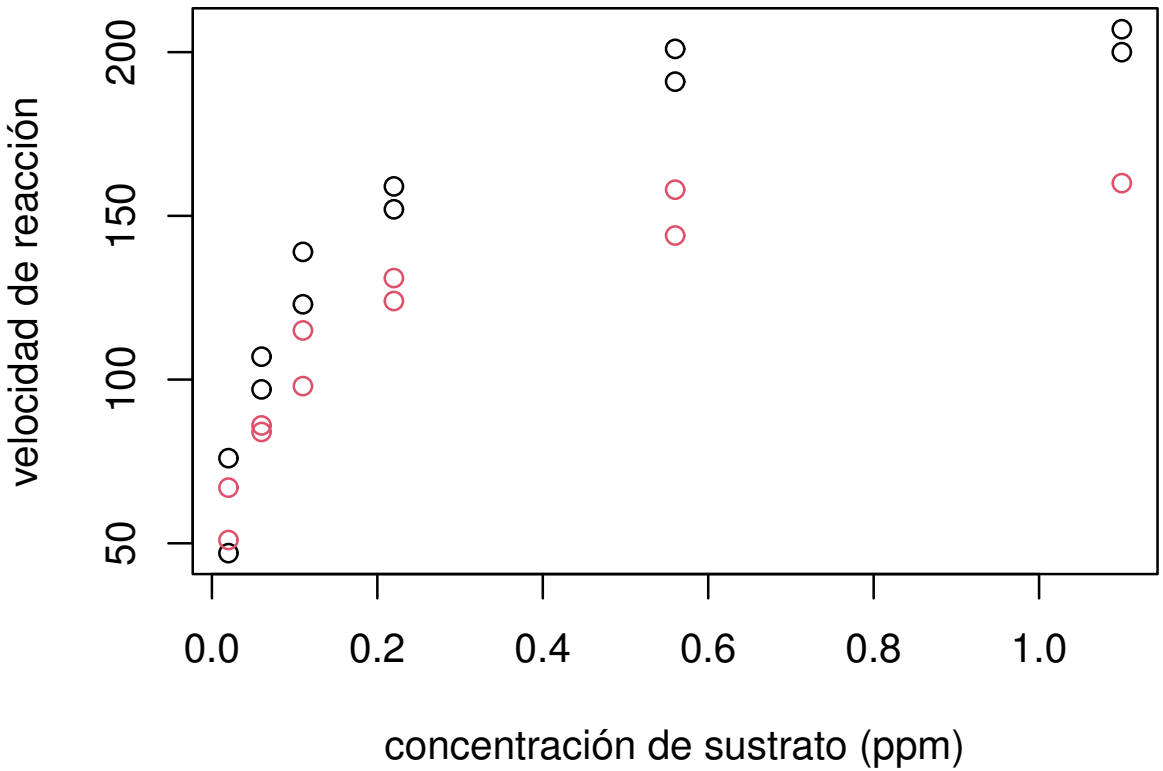
Objetivo: evaluar la velocidad de una reacción enzimática de células tratadas con Puromicina.

Se midió la reacción enzimática (qué tan rápido ésta cataliza la reacción que convierte un sustrato en producto) de 23 encimas (12 tratadas con Puromicina).

Las variables son:

- `conc`: concentración de sustrato (ppm).
- `rate`: velocidad de reacción instantáneas (conteo/min^2).
- `state`: tratado y no tratado.

Puromicina



Encimas tratadas (puntos negros) - Encimas no tratadas (puntos rojos)

Temas del curso

Algunas preguntas que nos podemos planear son:

~~¿Cómo ingreso covariables categóricas dentro del modelo de regresión?~~

Si hay problemas de multicolinealidad, ¿cómo se puede corregir?

Si tengo muchas covariables, ¿cómo puedo hacer la "mejor selección" de ellas para ajustar el modelo?

Si la relación entre la variable respuesta y las covariables se puede explicar por medio de una función no lineal, ¿cómo estimo los parámetros?

Si mi variable respuesta es binaria o de conteo, ¿cómo puedo proponer un modelo basado en una distribución diferente a la normal?

Temas del curso

Temas principales:

- Modelos polinomiales.
- ~~Variables indicadoras~~.
- Multicolinealidad (identificación y corrección).
- Métodos de selección de variables.

Temas adicionales (se toman de forma introductoria y si el tiempo lo permite):

- Modelos no lineales.
- Modelo lineal generalizado (principalmente modelo logístico y Poisson).
- modelos con efectos aleatorios (modelo lineal mixto).

Evaluación

Evaluación del curso:

- Tareas (60 %)
- Trabajo final con sustentación (40 %)

Referencias

Texto Guía

- *Introduction to Linear Regression Analysis*, Fifth Ed., 2012, by Montgomery, D. C., Peck, E. A., and Vining, G. G.

Textos de Referencia

- *Applied Regression Analysis*, Third Ed., 1998, Draper, N. R. and Smith, H.
- *Applied Linear Regression*, Fourth Ed., 2013, Weisberg, S.
- *Linear Regression Analysis*, Second Ed., 2003, by Seber, G. A. F. and Lee, A. J.
- *Applied Linear Statistical Models*, Fifth Ed., 2005, by Kutner, M., Nachtsheim, C. J., Neter, J., and Li, W.