MODELOS LINEALES

MODELOS LINEALES
Introducción

Douglas Rivas
Universidad de Los Andes

Clases

PART I

INTRODUCCIÓN

Capítulo 1

MODELOS LINEALES

> The sheer volumne of answers can often stifle insight...The purpose of computing is insight, not numbers.
>
> —Hamming [**?**]

1.1 Introducción

Uno de los objetivos de la ciencia es describir y predecir eventos en el mundo en el cual vivimos. Una manera de alcanzarlo es al hallar una formula o ecuación que relaciona cantidad en el mundo real. Podemos estar interesados, por ejemplo, en la relación entre la temperatura y la presión en un proceso químico o en la relación entre el número de manzanas en varios árboles en una hectárea y la cantidad de fertilizante que cada árbol recibe, etc.

Esta en la naturaleza humana tratar de entender los fenómenos físicos y naturales que ocurren a nuestro alrededor. Cuando las observaciones sobre un fenómeno pueden ser cuantificadas, tal intento de entendimiento a menudo implica la construcción de un modelo matemático, incluso si es sólo un intento simplista para capturar los elementos esenciales. Ya sea debido a nuestra ignorancia o con el fin de mantenerlo simple, muchos factores relevantes pueden dejarse fuera. También los modelos necesitan ser validado a través de las mediciones, y tales mediciones a menudo vienen con error. Con el fin de tener en cuenta la medición o los errores observacionales así como los factores que pueden quedarse fuera,

uno necesita un modelo estadístico el cual incorpora alguna cantidad de incertidumbre.

1.2 Modelo Matemático y Estadístico

Para nuestro propósito, describimos un **modelo matemático** como una relación funcional entre variables. En particular, estamos interesados en modelos que relacionan un conjunto de variables de entrada a un conjunto de variables de salida. Pensemos en una respuesta como la salida de un proceso que depende de una o más entradas. Esta idea se muestra en la figura ***. La caja indica un proceso en el cual las tres entradas son transformadas en una sola salida. Aquí consideramos una sola salida pero pueden ocurrir varias salidas. Matemáticamente, describimos la relación como

$$y = g(x_1, x_2, x_3) \tag{1.1}$$

donde y denota la salida, x_1, x_2 y x_3 denotan las entradas y $g(x_1, x_2, x_3)$ denota la relación funcional por la cual las entradas son convertidas en la salida. Nos referimos a esta como la función respuesta. Los conceptos son ilustrados en el siguiente ejemplo

EJEMPLO 1.1

En la fabricación de tableros, pequeñas partículas de madero son mezcladas con un adhesivo, formando laminas de un espesor dado, y cocidas en un horno. La compañía está interesada en la relación de la resistencia de los tableros a la temperatura de cocción. Para examinar esta relación, varios tableros son producidos cada uno a diversas temperaturas y la resistencia, y y la temperatura, t son medidas. Basado en esta información, el analista desea determinar la relación funcional.

La búsqueda de la relación funcional en el ejemplo 1.1 puede ser facilitada por un gráfico de dispersión de la resistencia versus la temperatura. (Aquí se puede hacer porque solo es una variable de entrada). El gráfico podría sugerir que la relación es cercanamente lineal, pero que hay datos alejados de la linealidad. Además, por lo general ocurre que observaciones a la misma temperatura no obtienen la misma resistencia. **¿A qué se deberá ese comportamiento?**. Esta diferencia podría ser causada por otros factores que influyen en la resistencia y no han sido tomados en cuenta en el estudio. En muchos casos, estas diferencias no explicadas pueden solamente atribuirse a la variabilidad natural en el material o al procesos que no tiene una explicación matemática. Para incluir tal variabilidad se introduce el concepto de un **modelo estadístico**. En particular, consideremos la extensión del modelo matemático agregándole una variable aleatoria en la lado donde están las entradas en la ecuación 1.1. Por lo tanto, escribimos el modelo como

$$y = g(x_1, x_2, x_3) + \varepsilon \tag{1.2}$$

donde ε denota la variable aleatoria agregada, a menudo llamado **termino de error**. Las propiedades de esta variable aleatoria dependerá de la situación, pero a menudo se asume que sigue una distribución normal univariante con media cero y varianza constante (σ^2). Cómo consecuencia de este supuesto, y, puede ser vista como una variable aleatoria con media, $g(x_1, x_2, x_3)$, y varianza (σ^2). Así podemos escribir la parte deterministica del modelo, usando $E(Y)$ para denotar el valor esperado, como

$$E(y) = g(x_1, x_2, x_3) \tag{1.3}$$

En muchas situaciones, la forma funcional del modelo matemático puede ser conocida, es decir se puede saber que la relación es lineal, pero los parámetros pueden ser desconocidos. Si denotamos como β al vector de parámetros y a x el vector de entradas, escribimos la parte deterministica como

$$E(y) = g(x, \beta) \tag{1.4}$$

Si agregamos el supuesto de independencia a las variables aleatorias asociadas con la respuesta individual, los datos para el ejemplo 1.1 pueden ser visto como una muestra aleatoria de una población con media dada por 1.4 y varianza σ^2. En este caso, la población solamente esta definida conceptualmente como la colección de posibles tableros que podrían ser producidos. En otros casos, puede ser posible enumerar la población, tal como la población de estudiantes en la facultad de ingeniería de cierta universidad. En ese caso podríamos muestrear la población en vez de observar todos los estudiantes o, alternativamente, podríamos ver este grupo de estudiantes como una muestra de la colección de todos los posibles estudiantes de ingeniería en esta universidad. Los datos podrían haber sido recogidos para desarrollar un modelo para la relación entre la nota de un examen como una función de la nota promedio hasta el momento.

El concepto de una variable de entrada es muy general. Por ejemplo, cuando modelamos la cantidad diaria de agua usada en una refineria de aceite, las entradas pueden inluir el tamaño de la refinería, la cantidad de aceite crudo procesado, el número de torres de calentamiento y los tipos de productos. En general, las entradas pueden ser **cuantitativas**, esto es, medidas tales como temperatura o cantidad, o ellas pueden ser **cualitativas**, indicando la presencia o ausencia de un factor o el tipo de producto.

Las observaciones pueden surgir como resultado de un **experimento diseñado** cuidadosamente. Por ejemplo, si deseamos evaluar el efecto de la temperatura sobre la resistencia de un producto, podemos conducir un experimento controlado en el cual el proceso de producción es corrido a diferentes temperaturas mientras mantienen los otros factores en valores fijos. De manera alternativa, los datos pueden venir den un **estudio observacional**. En tales situaciones, las mediciones son tomadas sobre una muestra aleatoria de individuos, o unidades experimentales, desde una población dada. Por ejemplo, podemos seleccionar una muestra aleatoria de estudiantes mujeres en un grupo de edad especifica y medir el porcentaje de gordura y ciertas características físicas. En este caso el objetivo es desarrollar un modelo para predecir la gordura desde características de más fácil medición. Trataremos cada una de estas situaciones de la misma manera, reconociendo que en el experimento diseñado, las salidas desde el modelo asumido pueden solamente ser causados por la variabilidad natural del material, mientras que en el estudio observacional, puede incluir otros factores que no hemos incluido como entradas.

Un análisis de modelos estadísticos incluye hacer inferencias acerca de los parámetros desconocidos y usar el modelo para predecir futuras observaciones.

1.3 El Modelo Lineal

Una pregunta importante que a menudo uno intenta responder a través de modelos estadísticos es la siguiente: ¿Cómo puede una cantidad observada y ser explicada por otras cantidades $x_1, x_2, ...x_p$?. Quizás el modelo más simple que es usado para responder esta

pregunta es el *modelo lineal*:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{1.5}$$

donde $\beta_0, \beta_1, ..., \beta_p$ son constantes y ϵ es un término de error para explicar la incertidumbre. Nos referimos a y como la variable respuesta. Esta es también conocida como la *variable dependiente*, *variable endógena* o *variable criterio*. Nos referimos a $x_1, x_2, ...x_p$ como *variables explicatorias*. Estas son también llamadas *variables independientes* o *variables exógenas*. En el contexto de algunos casos especiales estas son llamadas *regresores*, *predictores* o *factores*. Los coeficientes $\beta_0, \beta_1, ..., \beta_p$ son lo parámetros del modelo. Note que el lado derecho de 1.5 el cual es una función lineal de las variables explicatorias, también puede ser vista como una función lineal de los parámetros.

EJEMPLO 1.2

La cuenta del hospital de un paciente es probablemente a ser más grande si el paciente tiene que estar más dias en el hospital. La cuenta también depende de otros factores incluyendo la naturaleza del tratamiento, si cuidado intensivo es necesario y así sucesivamente. Algunos factores pueden ser desconocidos (como la codicia del hospital). Un modelo simple que puede ser usado aquí es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

donde y es la cantidad de la cuenta del hospital, x_1 es la duración de la estadía en el hospital (excluyendo la estadía en la unidad de cuidados intensivos) y x_2 es la duración de la estadía en la unidad de cuidados intensivos. El término de error ϵ representan todos los factores que no están especificamente incluidos, tales como la naturaleza de los tratamientos, pruebas y la variación de un hospital a otro. El modelo anterior es un caso especial de 1.5.

EJEMPLO 1.3

La estatura de una persona adulta varía de un grupo étnico homogéneo a otro. Este también depende del género de la persona. Una comparación de dos grupos en términos de la estatura puede hacerse sobre la base del siguiente modelo, el cual nuevamente es un caso especial de (1.5):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

donde y es la estatura medida de un adulto, x_1 es una variable binaria que representa el grupo étnico y x_2 es otra variable binaria que representa el género. El término de error ϵ representan una combinación de errores de medida y la variación en la estatura que existe entre los adultos de un género particular en un grupo étnico dado.

EJEMPLO 1.4

El rendimiento de té en un acre de una plantación de té depende de varios tipos de prácticas de agricultura (tratamientos). Un experimento puede ser planificado donde varias parcelas están sujetas a solo uno de dos posibles tratamientos sobre un período

de tiempo. El rendimiento de té luego de aplicar los tratamientos es registrado. Un modelo para el rendimiento port-tratamiento (y) es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

donde la variable binaria x_1 representa el tipo de tratamiento y la variable real x_2 es el rendimiento pre-tratamiento. El término de error ϵ principalmente consiste de factores no tomados en cuenta. La inclusión de x_2 es para reducir el efecto de los factores no tomados en cuenta tales como el tipo de aceite o las diferencias inherentes en los arbustos de té.

Es importante tener claro el término **lineal** en la definición del modelo. Cuando decimos que el modelo es lineal, nos referimos a que es lineal en los parámetros y no en las variables. Por ejemplo el siguiente modelo es un modelo lineal como el que requerimos $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, pues es lineal en los parámetros. Por el contrario el modelo $y = \beta_0 + \beta_1^2 x_1 + \beta_2 x_2 + \varepsilon$ no cumple con nuestra definición de lineal porque no es lineal en los parámetros aunque si lo sea en las variables. Ahora bien, para ver si quedo claro la linealidad de un modelo, ¿será este un modelo lineal $y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2 + \varepsilon$?, ¿ y este $y = \beta_0 + \beta_1 ln(x_1) + e^{\beta_2} x_2 + \varepsilon$?.

1.4 ¿Por qué un modelo lineal?

El modelo (1.5) es justo uno de los posibles modelos que pueden ser usados para explicar la respuesta en términos de las variables explicatorias. Algunas de las razones por las que llevamos a cabo un estudio detallado del modelo lineal son las siguientes.

1. Debido a su simplicidad, el modelo lineal es mas entendible y más fácil de interpretar que los otros modelos competidores, además los métodos de análisis y las inferencias son mejor desarrolladas. Por lo tanto, si no hay una razón particular para presuponer otro modelo, el modelo lineal puede ser usado al menos como un primer paso.

2. La formulación de un modelo lineal es usado aún para ciertos modelos no lineales los cuales pueden ser reducidos a la forma (1.5) a través de una transformación.

3. Los resultados obtenidos por el modelo lineal sirven como un trampolín para el análisis de una clase amplia de modelos relacionados tales como modelos de efectos mixtos, espacio de estados y otros modelos de series de tiempo.

4. Suponga que la respuesta es modelada como una función no lineal de las variables explicatorias más el error. En muchas situaciones prácticas solamente una parte del dominio de esta función es de interés. Por ejemplo, en un proceso de manufactura, uno esta interesado, en una pequeña región centrada alrededor del punto de operación. Si la función anterior es razonablemente suave en esta región, un modelo lineal sirve como una buena primera aproximación para lo que es goblamente el modelo no lineal.

1.5 Descripción del Modelo lineal y notaciones

Si uno usa (1.5) para un conjunto de n observaciones de la respuesta y variables explicatorias, la forma explicita de las ecuaciones debe ser

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i, \qquad i = 1, 2, ..., n \tag{1.6}$$

donde para cada i, y_i es la i-ésima observación de la respuesta, x_{ij} es la i-'esima observación de la j-ésima variable explicatoria $(j = 1, 2, ..., p)$, y ϵ_i es el error no observable conrrespondiente a esta observación. Este conjunto de n ecuaciones puede ser escrito en la siguiente forma compacta al usar matrices y vectores

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{1.7}$$

En este modelo,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Para completar la descripción del modelo, algunos supuestos sobre la naturaleza de los errores son necesarios. Se asume que los errores tienen media cero y sus varianzas y covarianzas son conocidas como un factor de escala. estos supuestos se resumen en la forma matriz-vector como

$$E(\epsilon) = 0, \qquad D(\epsilon) = \sigma^2 V \tag{1.8}$$

donde la notación E es para valor esperado y D representa matriz de dispersión (o matriz de varianza-covarianza). El vector 0 denota un vector con elementos cero (en este caso n elementos) y V es una matriz conocida de orden $n \times n$. El parámetro σ^2 no esta especificado, al igual que el parámetro β. Los elementos de β son reales, mientras que σ^2 es no negativa.

Usaremos la tripleta $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 V)$ como una manera corta para el modelo lineal (1.6)-(1.7). Cuando los errores $\epsilon_1, \epsilon_2, \epsilon_n$ están descorrelacionados y cada uno tiene varianza σ^2, tenemos el caso especial $V = I$, la matriz identidad $n \times n$. En este caso especial, llamamos al modelo $(y, X\beta, \sigma^2 I)$ el modelo lineal *homocedastico*. Cuando es necesario diferenciar el modelo $(y, X\beta, \sigma^2 V)$ del modelo homocedastico, nos referimos al primero como el modelo lineal general.

Por lo tanto definimos un modelo lineal general de la siguiente manera

Definición 1.1 *Un modelo lineal general se define como*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

donde $\boldsymbol{y}$ es un vector $n \times 1$ de observaciones aleatorias, $\boldsymbol{X}$ es una matriz $n \times p$ de constantes conocidas llamada matriz del modelo o matriz de diseño, $\boldsymbol{\beta}$ es un vector $p \times 1$ de parámetros fijos no observables, y ev es un vector $n \times 1$ de errores aleatorios no observables. Ambos $\boldsymbol{y}$ y $\boldsymbol{\varepsilon}$ son vectores aleatoris. Asumimos que $E(\boldsymbol{\varepsilon} = 0)$ y $Cov(\boldsymbol{\varepsilon}) = \sigma^2 V$ donde V es una matriz definida no negativa.

Cuando en la definición anterior $Cov(\boldsymbol{\varepsilon}) = \sigma^2 I$, es decir los errores tienen varianza común y están descorrelacionodos, comúnmente al modelo se le conoce como el **Modelo Lineal Homocedastico**

Las aplicaciones de los modelos lineales caen en dos casos especiales: Análisis de Regresión y Análisis de Varianza. El Análisis de Regresión se refiere a modelos en los cuales la matriz $\boldsymbol{X}'\boldsymbol{X}$ es no singular. Los modelos de Análisis de Varianza son modelos en los cuales la matriz de diseño consiste completamente de cero y unos.

Aunque los valores de las variables explicatorias, x_{ij}, $j = 1, 2, ..., p$, $i = 1, 2, ..., n$ son por lo general controladas en el tiempo de conducción de un experimento, como uno podría esperar en el ejemplo 1.3. Estas pueden también ser cantidades observadas más allá del control del observador, como en el caso de x_1 y x_2 del ejemplo 1.2 y x_2 del ejemplo 1.4. En este caso, los x_{ij} se pueden asumir aleatorios, y el modelo (1.6)-(1.7)se puede interpretar como el modelo condicional de y dado X. Una representación explicita del modelo condicional está dado por

$$E(y|X) = X\beta, \qquad D(y|X) = \sigma^2 V \tag{1.9}$$

Así, el término error en (1.6) es la diferencia $y - E(y|X)$. La media y dispersión del error dado por (1.7) se interpreta como condicionada sobre X. La representación (1.9) es llamado el *modelo de regresión lineal*. En ese contexto, β es llamado el vector de los parámetros de regresión o coeficientes de regresión. Un aspecto importante del modelo de regresión lineal es que el error $y - E(y|X)$ debe estar descorrelacionado de X. (ver ejercicio 1.6)

Suponga que las observaciones $(y_i, x_{i1}, x_{i2}, ..., x_{ip})$ para $i = 1, 2, ..., n$ son estadísticamente independientes (en cuyo caso $V = I$). Entonces el modelo condiccional (1.9) puede escribirse en la forma mas simple

$$\begin{aligned} E(y|x_1, x_2, ..., x_p) &= \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p, \\ var(y|x_1, x_2, ..., x_p) &= \sigma^2 \end{aligned} \tag{1.10}$$

donde $Var(.)$ indica varianza, y establecida para cualquier respuesta observada y $x_1, x_2, ..., x_p$ son las correspondientes variables explicativas.

1.6 Usos de los modelos lineales

Una importante aplicación de los modelos lineales es en análisis de regresión donde la respuesta promedio es explicada a través de otras variables observadas (llamadas regresores en ese contexto). La construcción de un modelo para describir la relación entre las variables es a veces un fin en sí mismo. Por otro lado, el modelo puede ser usado como un vehículo para diversos tipos de análisis, como se describe más abajo.

El objetivo del análisis puede ser sobre un regresor en particular. Por ejemplo, podemos desear hallar específicamente como la longevidad humana esta asociada al nivel de colesterol en la sangre. Uno podría construir un modelo lineal para longevidad de manera que otras variables de influencia potencial tales como genero, factores de desarrollo, estado civil y factores de salud aparte del nivel de colesterol en la sangre son también incluidos. El modelo debería tener la forma (1.5) donde y es la longevidad, x_1 es el nivel de colesterol en la sangre y $x_2, ..., x_p$ representan los otros regresores. De acuerdo a este modelo, el parámetro β es la tasa de cambio de la longevidad promedio con el nivel de colesterol en la sangre, con los otros factores manteniéndose constantes. El análisis de datos obre la

base de modelos lineales deben producir una estimación de β_1, junto a una estimación del error de estimación asociado. Los métodos para obtener estas estimaciones son discutidas en el capítulo 4 y 7.

El análisis de regresión sobre la base de un modelo lineal es también llevado a cabo para examinar estadísticamente ciertas creencias empíricas en relación con el modelo. Por ejemplo, en el contexto del ejemplo 1.2, se puede desear probar la declaración de que la duración de mantenerse en la unidad de cuidados intensivos afecta la cuenta del hospital al menos tres veces mas que la duración fuera de la unidad de cuidados intensivos. La forma cuantitativa de esta declaración es la hipótesis $\beta_2 \geq 3\beta_1$, la cual se puede probar sobre la base de los datos disponibles. La prueba de hipótesis estadística de este tipo son discutidas en los capítulos 5 y 7.

Otro importante uso de los modelos de regresión lineal es en el área de la predicción. Los valores de la variable de principal interés (la variable respuesta) puede ser imposible obtener en el momento del análisis o puede envolver costosas mediciones. Algunas otras variables pueden ser identificadas como variables explicatorias o predictoras. Datos sobre todas las variables son recolectadas para ajustar un modelo de regresión lineal. Valores no observables de la variable respuesta se pueden predecir sobre la base del modelo ajustado y los valores de las variables explicatorias correspondientes a la respuesta no observada. Ver capítulos 5 y 7 para los métodos de predicción en el modelo lineal. Algunas veces valores no observados de las variables explicatorias son predichas sobre la base de la respuesta correspondiente y un modelo ajustado. Este problema de predicción reversa es llamada *calibración* (ver Brown, 1993).

A parte de la situación donde el observador no tiene el control sobre las variables explicatorias, *experimentos diseñados* pueden ser usados también para medir los efectos de ciertas variables explicatorias o para probar estadísticamente creencias empíricas. El modelo lineal puede ser usado por el experimentador como una base para seleccionar los valores de las variables controlables de manera que la respuesta a la pregunta crucial sea la mejor respuesta. El capítulo 6 brevemente cubre lo básico de diseño de experimentos.

Si una de las variables explicatorias es controlable, entonces uno se podría preguntar: ¿cuáles valores de esta variable producirán un nivel deseado de repuesta (dentro de cierto margen de error)?. Esta pregunta, la cual esta relacionada con calibración, es llamado el problema de control. El modelo lineal provee un marco de referencia para resolver este problema (ver Press, 1971,Capítulo 14).

El modelo lineal es también usado como una base para la imputación de datos perdidos. La idea es llenar el vacío utilizando la información de las variables relacionadas, como en el caso de la predicción (ver Titterington y Sedransk, 1987, para detalles). Las herramientas de diagnóstico desarrollados en el contexto del modelo lineal se utilizan a veces para detectar otros defectos en los datos, tales como datos malos o incorrectos (ver Besley et. al, 1980).

1.7 Modelos lineales aplicados

Entre los modelos lineales de mayor aplicación se encuentran los modelos de regresión lineal y los modelos de diseño conocidos también como modelos de análisis de varianza. El objetivo de este curso es el desarrollo de ambos modelos. Pero por el momento veamos algunos ejemplos clásicos de ambos:

EJEMPLO 1.5 Modelo de regresión lineal simple

Es el modelo de regresión más sencillo en el cual la variable respuesta depende de una sola variable explicativa. Por lo tanto el modelo es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, ..., n$$

donde los ϵ_i son variables aleatorias independientes $N(\mu, \sigma^2)$. Supongamos el caso particular en el cual se tienen solo 6 observaciones $(n = 6)$ y que los valores de la variable explicatoria son $(x_1, x_2, x_3, x_4, x_5, x_6) = (1, 2, 3, 4, 5, 6)$ entonces el modelo en notación matricial es

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$

□

EJEMPLO 1.6 Modelo de análisis de varianza de una vía

Es el modelo de diseño más sencillo en el cual la variable respuesta depende de un solo factor. Por lo tanto el modelo es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, ..., n$$

donde los ϵ_i son variables aleatorias independientes $N(\mu, \sigma^2)$. Observemos que el modelo es básicamente el mismo, pero como la matriz de diseño esta compuesta de ceros y unos es común representar este modelo de la siguiente manera

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, ..., n \qquad j = 1, ..., J$$

donde μ es el promedio general, α_j es el efecto del $j - esimo$ tratamiento. Supongamos el caso en el que $i = 1, 2, 3$, $j = 1, ..., N_i$, $(N_1, N_2, N_3) = (3, 1, 2)$, entonces la representación matricial es

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$