# Modelado estadístico de datos: Práctica 1

Sergio Soler Rocha
Universidad Nacional de Educación a Distancia

## 1. ¿Qué es un estudio cruzado ("cross-over study")? Se pide además ilustrar este concepto con algún ejemplo numérico. En el documento que se entregue habrá que incluir el código utilizado

Un estudio cruzado (cross-over study) es un diseño de mediciones repetidas en el cual cada unidad experimental (generalmente un paciente) recibe distintos tratamientos durante diferentes períodos de tiempo. En otras palabras, los pacientes cambian de un tratamiento a otro a lo largo del estudio. Esto contrasta con el diseño paralelo, donde los pacientes son asignados al azar a un tratamiento y permanecen en ese mismo tratamiento durante toda la duración del estudio. De esta manera, cada paciente actúa como su propio control, experimentando ambos tratamientos en una secuencia de periodos establecida de manera aleatoria y separados entre sí por un periodo de lavado. El periodo de lavado es un intervalo de tiempo que se introduce entre la administración de diferentes tratamientos. Su propósito principal es permitir que los efectos del tratamiento anterior se eliminen antes de comenzar con el siguiente tratamiento.

La razón para considerar un estudio cruzado es que si se cumplen las condiciones necesarias, podría permitir una comparación de tratamientos más eficiente que un diseño paralelo. Por lo tanto, podrían necesitarse menos pacientes en el estudio cruzado para lograr el mismo nivel de precisión que en un diseño paralelo. La gran ventaja es que cada paciente actúa como su propio control, recibiendo tanto el tratamiento A como el B. Los estudios cruzados son populares, especialmente en medicina, pero también se utilizan en muchas otras disciplinas. Se compara la respuesta del mismo paciente en A versus B.

Como ventajas, se tiene, como ya hemos mencionado antes, que cada paciente actúa como su propio control, lo que significa que se obtendrá una menor variabilidad en las comparaciones entre los efectos de los tratamientos en distintos pacientes. Esto implica que se puede reducir considerablemente el tamaño muestral. Además, la estimación será menos sesgada, ya que la respuesta en un estudio paralelo será más variable al medirse en pacientes diferentes.

Como desventajas, se observa que este tipo de diseños son menos flexibles que los estudios paralelos. En caso de tratarse de alguna enfermedad, esta debe ser estable y crónica; es decir, no se puede llevar a cabo un estudio cruzado con tratamientos destinados a curar una enfermedad. Por ejemplo, si el tratamiento A cura a un paciente, carecería de sentido que el mismo paciente experimentara con el tratamiento B. Por tanto, los tratamientos deben estar destinados a aliviar la enfermedad. Otra desventaja podría ser la duración de los diferentes periodos que deben seguir los pacientes durante el estudio. Al tener que recibir el tratamiento A, luego pasar por un periodo de lavado y, finalmente, tomar el B, aumenta la probabilidad de que se produzcan bajas en los participantes, algo que no ocurriría en un estudio paralelo. También como desventaja, hay que considerar el efecto residual que persiste después de aplicar un tratamiento. A pesar de que haya un periodo de lavado para eliminar los efectos del primer tratamiento, estos podrían no disminuir completamente.

Existen diferentes diseños de estudios cruzados, siendo el más popular (que se estudiará más adelante) el diseño de dos secuencias, dos períodos y dos tratamientos. Los pacientes se asignan aleatoriamente a ambas secuencias. En la secuencia AB, reciben el tratamiento A en el primer período y el tratamiento B en el segundo período. En la otra secuencia, BA, reciben el tratamiento B en el primer período y, posteriormente, el tratamiento A en el segundo período, Cuadro 1.

Supongamos que los pacientes se asignan aleatoriamente al tratamiento A, seguido del tratamiento B (Secuencia AB), o al tratamiento B seguido del tratamiento A (Secuencia BA). Su-

Cuadro 1: Diseño AB-BA

| | Periodo 1 | Periodo 2 |
|---|---|---|
| Secuencia AB | A | B |
| Secuencia BA | B | A |

pongamos que hay $n_{AB}$ y $n_{BA}$ pacientes en cada secuencia, con un tamaño de muestra total de $N = n_{AB} + n_{BA}$. Aunque se espera que los pacientes estén estables, un modelo estadístico para un diseño cruzado debe incluir parámetros denominados efectos de período, que representan la diferencia entre el período 2 y el período 1, independientemente del orden del tratamiento, ya que la salud de los pacientes puede cambiar con el tiempo. En este contexto, se pueden identificar dos fuentes de variación: la variación entre pacientes y la variación dentro de cada paciente.

Si $Y_{ij}$ es la respuesta del paciente $i^{th}$ durante el período $(j = 1, 2)$, El modelo para un estudio cruzado AB-BA se define de la siguiente manera:

$$\begin{aligned} y_{ij} &= \mu_A + \eta_i + \epsilon_{ij} && \text{Periodo 1 en la Secuencia AB} \\ y_{ij} &= \mu_B + \phi + \eta_i + \epsilon_{ij} && \text{Periodo 2 en la Secuencia AB} \\ y_{ij} &= \mu_B + \eta_i + \epsilon_{ij} && \text{Periodo 1 en la Secuencia BA} \\ y_{ij} &= \mu_A + \phi + \eta_i + \epsilon_{ij} && \text{Periodo 2 en la Secuencia BA} \end{aligned}$$

- $\mu_A$ es la media del tratamiento A
- $\mu_B$ es la media del tratamiento B
- $\phi$ es el efecto periodo
- $\eta_i$ es una variable aleatoria para el paciente i con $E[\eta_i] = 0$
- $\epsilon_{ij}$ es una variable aleatoria para el paciente i en el periodo j con $E[\epsilon_{ij}] = 0$

En ocasiones, los estudios cruzados se analizan mediante el uso de una prueba t de muestra única aplicada a la diferencia entre los dos tratamientos, también llamada prueba t pareada.

Sea $c_i = y_{i2} - y_{i1}$ para los $i \in AB$ y $c_i = y_{i1} - y_{i2}$ para los $i \in BA$ podemos definir el estadístico:

$$\hat{\tau} = \bar{c} = \frac{\sum_{i=1}^{n} c_i}{N} \tag{1}$$

con el error estandar:

$$\hat{SE}[\bar{c}] = \sqrt{\frac{s_c^2}{N}} \tag{2}$$

Definiendo el nuevo estadístico como:

$$t_{exp} = \frac{\bar{c}}{\hat{SE}[\bar{c}]} \tag{3}$$

Si $c_i$ tiene una distribución normal, el test estadístico $T_C$ sigue una t-student con $N-1$ grados de libertad. Sin embargo la elección de este estadístico puede llevar a problemas se sesgos, ya que $E[\bar{c}]$ está sesgado.

$$\begin{aligned} E[\bar{c}] &= E\left[\frac{\sum_{i\in AB} y_{i2} - y_{i1}}{N} + \frac{\sum_{i\in BA} y_{i1} - y_{i2}}{N}\right] = E\left[\frac{n_{AB}}{N}(\bar{y}_{i2} - \bar{y}_{i1}) + \frac{n_{BA}}{N}(\bar{y}_{i1} - \bar{y}_{i2})\right] = \\ &= \frac{n_{AB}}{N}(\mu_B - \mu_A + \phi) + \frac{n_{BA}}{N}(\mu_B - \mu_A - \phi) = \mu_B - \mu_A + \frac{n_{AB} - n_{BA}}{N}\phi \end{aligned} \tag{4}$$

Hemos tenido en cuenta que los dos primeros sumatorios pertenecen a la secuencia AB, mientras que los dos últimos pertenecen a la secuencia BA, y que $E[\eta_i] = 0$ y $E[\epsilon_{ij}] = 0$. Observamos que hay un sesgo debido al efecto de período que rara vez se puede descartar, lo que lleva a la necesidad de utilizar los mismos pacientes en la secuencia AB que en la secuencia BA. Sin embargo, como ya hemos explicado, puede ocurrir que algunos pacientes no completen el proceso, lo que podría

causar un desequilibrio. Para corregirlo, sería necesario descartar otras pruebas que podrían ser útiles para el estudio. Por lo tanto, no se recomienda utilizar el estadístico $\hat{\tau}$.

Para encontrar un estimador no sesgado, definimos $d_i = y_{i2} - y_{i1}$, teniendo:

$$\bar{d}_{AB} = \frac{\sum_{i \in AB} d_i}{n_{AB}} \quad \text{y} \quad \bar{d}_{BA} = \frac{\sum_{i \in BA} d_i}{n_{BA}} \tag{5}$$

estudiamos ahora la esperanza de $\bar{d}_{AB} - \bar{d}_{BA}$:

$$E\left[\bar{d}_{AB} - \bar{d}_{BA}\right] = E\left[\bar{d}_{AB}\right] - E\left[\bar{d}_{BA}\right] = \mu_B + \phi - \mu_A + \mu_B - \mu_A - \phi = 2(\mu_B - \mu_A) \tag{6}$$

Entonces,

$$E\left[\frac{\bar{d}_{AB} - \bar{d}_{BA}}{2}\right] = \mu_B - \mu_A \tag{7}$$

Por lo que definimos estadístico no sesgado $\hat{\tau} = \frac{\bar{d}_{AB} - \bar{d}_{BA}}{2}$.

Calculamos ahora las varianzas de $d_i$ en la seqcuencias AB y BA respectivamente,

$$\sigma^2_{AB} = Var[d_i] = Var[y_{i2} - y_{i1}] = Var[\epsilon_{i2} - \epsilon_{i1}] = 2\sigma^2_\epsilon \tag{8}$$

$$\sigma^2_{BA} = Var[d_i] = Var[y_{i2} - y_{i1}] = Var[\epsilon_{i2} - \epsilon_{i1}] = 2\sigma^2_\epsilon \tag{9}$$

Hemos aplicado la propiedad de que la varianza de una constante es igual a cero, asumiendo que la covarianza entre $\epsilon_{i2}$ y $\epsilon_{i1}$ es también cero.

Para la evaluación de hipótesis, tememos $H_0$: $\mu_B - \mu_A = 0$ versus $H_1$: $\mu_B - \mu_A \neq 0$. Con el estadístico:

$$T = \frac{\frac{\bar{D}_{AB} - \bar{D}_{BA}}{2} - (\mu_B - \mu_A)}{EE\left[\frac{\bar{D}_{AB} - \bar{D}_{BA}}{2}\right]} \tag{10}$$

$$t_{exp} = \frac{\frac{\bar{d}_{AB} - \bar{d}_{BA}}{2}}{\hat{EE}\left[\frac{\bar{d}_{AB} - \bar{d}_{BA}}{2}\right]} \tag{11}$$

Donde,

$$\hat{EE}\left[\frac{\bar{d}_{AB} - \bar{d}_{BA}}{2}\right] = \frac{1}{2}\hat{EE}\left[\bar{d}_{AB} - \bar{d}_{BA}\right] = \frac{1}{2}\sqrt{\hat{s}^2_c\left(\frac{1}{n_{AB}} + \frac{1}{n_{BA}}\right)} \tag{12}$$

Donde aplicamos el caso de homocedasticidad,

$$\hat{s}^2_c = \frac{(n_{AB} - 1)\hat{s}^2_{AB} + (n_{BA} - 1)\hat{s}^2_{BA}}{(n_{AB} - 1) + (n_{BA} - 1)} \tag{13}$$

Siendo $\hat{s}^2_{AB}$ y $\hat{s}^2_{BA}$ las cuasivarianzas.

Para el Intervalo de Confianza tenemos:

$$IC_{1-\alpha}(\mu_B - \mu_A) = \frac{1}{2}(\bar{d}_{AB} - \bar{d}_{BA}) \pm \frac{1}{2}t_{1-\alpha/2}\hat{EE}[\mu_B - \mu_A] \tag{14}$$

## 1.1. Ejemplo numérico

Para ilustrar el concepto de estudio cruzado, utilizaremos los datos del archivo 'senn_32.txt' como un ejemplo práctico de este tipo de estudio. En este caso, se trata de un estudio cruzado que involucra broncodilatadores. Los datos de la tabla se originan en un ensayo cruzado aleatorio AB-BA de dos períodos, donde el número '1' en la primera columna indica pertenencia a la secuencia AB, mientras que '2' indica la secuencia BA.

En este estudio, los pacientes consumen dos tipos de broncodilatadores, salbutamol (S) y formoterol (F), como se muestra en la columna 'tratamiento'. El resultado analizado es el flujo espiratorio máximo, que se registra en la columna 'pef'. Los participantes son asignados aleatoriamente para

recibir primero formoterol y luego salbutamol, o viceversa. En la columna 'periodo', el número '1' indica el primer periodo de tratamiento, mientras que '2' indica el segundo periodo.

La columna 'sujeto' se utiliza para identificar a cada paciente individualmente, asignándoles un número de identificación único.

Una vez importados los datos (véase ejercicio1.R), para cada sujeto restamos la cantidad de flujo espiratorio máximo (pef) del periodo 2 a la cantidad del periodo 1, luego dividimos el resultado entre dos y los separamos en dos vectores según el grupo al que pertenezcan. En resumen, se generan dos vectores de diferencias (divididas entre dos), donde las diferencias del primer vector corresponden a la secuencia AB y las del segundo vector corresponden a la secuencia BA. Posteriormente, aplicamos el test de Shapiro-Wilk a ambos vectores para evaluar su normalidad. En ambos casos, no se encuentran evidencias suficientes para rechazar la hipótesis de normalidad. Seguidamente, empleamos el test de Levene, el cual tampoco arroja evidencia de diferencias significativas en las varianzas entre ambos grupos.

Ahora, aplicamos la prueba test-student, de la cual obtenemos el intervalo de confianza:

$$IC_{95\,\%}(\mu_B - \mu_A) = (-70,32619, -22{,}88810)$$

Al realizar el contraste de hipótesis, obtenemos un valor de $p_{valor} = 0{,}001205 < 0{,}05$ lo que nos lleva a aceptar la hipótesis alternativa de que las diferencias entre las medias no son iguales a cero, por lo tanto hay evidencias para afirmar que el tratamiento A causa mayor efecto que el tratamiento B.

Por último, es importante señalar la existencia de diversas variantes del modelo que hemos estudiado, tales como ABB-BBA, AAB-ABA-BAA, entre otras. Además, se han llevado apueden llevar a cabo estudios cruzados con más de dos tratamientos, en muchos casos utilizando el test ANOVA.

## 2. Sea $\mathbf{z} \in \mathfrak{M}_{nx1}$ vector aleatorio, entonces se verifica que $E\left[\mathbf{z}\mathbf{z}^T\right] = E\left[\mathbf{z}\right] E\left[\mathbf{z}^T\right]$

Sea,

$$\mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \quad \text{y} \quad \mathbf{z}^T = \begin{pmatrix} z_1 & \cdots & z_n \end{pmatrix}$$

Entonces,

$$\mathbf{z}\mathbf{z}^T = \begin{pmatrix} z_1^2 & z_1 z_2 & \cdots & z_1 z_n \\ z_2 z_1 & z_2^2 & \cdots & z_2 z_n \\ \vdots & \vdots & \ddots & \vdots \\ z_n z_1 & z_n z_2 & \cdots & z_n^2 \end{pmatrix}$$

Por lo tanto,

$$E\left[\mathbf{z}\mathbf{z}^T\right] = \begin{pmatrix} E\left[z_1^2\right] & E\left[z_1 z_2\right] & \cdots & E\left[z_1 z_n\right] \\ E\left[z_2 z_1\right] & E\left[z_2^2\right] & \cdots & E\left[z_2 z_n\right] \\ \vdots & \vdots & \ddots & \vdots \\ E\left[z_n z_1\right] & E\left[z_n z_2\right] & \cdots & E\left[z_n^2\right] \end{pmatrix}$$

Por otro lado tenemos,

$$E\left[\mathbf{z}\right] = \begin{pmatrix} E\left[z_1\right] \\ \vdots \\ E\left[z_n\right] \end{pmatrix} \quad \text{y} \quad E\left[\mathbf{z}^T\right] = \begin{pmatrix} E\left[z_1\right] & \cdots & E\left[z_n\right] \end{pmatrix}$$

Por lo que,

$$E[\mathbf{z}]\,E\left[\mathbf{z}^T\right] = \begin{pmatrix} E^2[z_1] & E[z_1]E[z_2] & \cdots & E[z_1]E[z_n] \\ E[z_2]E[z_1] & E^2[z_2] & \cdots & E[z_2]E[z_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[z_n]E[z_1] & E[z_n]E[z_2] & \cdots & E^2[z_n] \end{pmatrix}$$

Si suponemos cierta la igualdad del enunciado, tenemos,

$$E\left[\mathbf{z}\mathbf{z}^T\right] - E[\mathbf{z}]\,E\left[\mathbf{z}^T\right] = \mathbf{0},$$

$$\begin{pmatrix} E\left[z_1^2\right] & E[z_1z_2] & \cdots & E[z_1z_n] \\ E[z_2z_1] & E\left[z_2^2\right] & \cdots & E[z_2z_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[z_nz_1] & E[z_nz_2] & \cdots & E\left[z_n^2\right] \end{pmatrix} - \begin{pmatrix} E^2[z_1] & E[z_1]E[z_2] & \cdots & E[z_1]E[z_n] \\ E[z_2]E[z_1] & E^2[z_2] & \cdots & E[z_2]E[z_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[z_n]E[z_1] & E[z_n]E[z_2] & \cdots & E^2[z_n] \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$\begin{pmatrix} E\left[z_1^2\right] - E^2[z_1] & E[z_1z_2] - E[z_1]E[z_2] & \cdots & E[z_1z_n] - E[z_1]E[z_n] \\ E[z_2z_1] - E[z_2]E[z_1] & E\left[z_2^2\right] - E^2[z_2] & \cdots & E[z_2z_n] - E[z_2]E[z_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[z_nz_1] - E[z_n]E[z_1] & E[z_nz_2] - E[z_n]E[z_2] & \cdots & E\left[z_n^2\right] - E^2[z_n] \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$\begin{pmatrix} Var(z_1) & Cov(z_1,z_2) & \cdots & Cov(z_1,z_n) \\ Cov(z_1,z_2) & Var(z_2) & \cdots & Cov(z_2,z_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(z_1,z_n) & Cov(z_2,z_n) & \cdots & Var(z_n) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Es decir, para que sea verdadero la matriz covarianza tiene que ser $\mathbf{0}$, que solo ocurre si el vector aleatorio es un vector constante. Por tanto concluimos que la igualdad del enunciado es **FALSA**.

## 3. Sea $\mathbf{z} \in \mathfrak{M}_{nx1}$ vector aleatorio y $\mathbf{A} \in \mathfrak{M}_{nxn}$ una matriz no aleatoria, entonces se verifica que $E\left[\mathbf{z}^T\mathbf{A}\mathbf{z}\right] = E\left[\mathbf{z}^T\right]\mathbf{A}E\left[\mathbf{z}^T\right]$

$$\mathbf{z}^T\mathbf{A}\mathbf{z} = \begin{pmatrix} z_1 & \cdots & z_n \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} =$$

$$= \begin{pmatrix} z_1 & \cdots & z_n \end{pmatrix} \begin{pmatrix} z_1a_{11} & + & \cdots & + & z_na_{1n} \\ & & \vdots & & \\ z_1a_{n1} & + & \cdots & + & z_na_{nn} \end{pmatrix} =$$

$$= a_{11}z_1^2 + \cdots + a_{1n}z_1z_n + \cdots + a_{n1}z_nz_1 + \cdots + a_{nn}z_n^2 =$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}z_iz_j$$

Aplicando la esperanza al término obtenido, tenemos,

$$E\left[\mathbf{z}^T\mathbf{A}\mathbf{z}\right] = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}E[z_iz_j] \tag{15}$$

Ahora desarrollamos análogamente el segundo miembro de la igualdad,

$$E\left[\mathbf{z}^T\right]\mathbf{A}E\left[\mathbf{z}^T\right] = \begin{pmatrix} E[z_1] & \cdots & E[z_n] \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} E[z_1] \\ \vdots \\ E[z_n] \end{pmatrix}$$

$$E\left[\mathbf{z}^T\right]\mathbf{A}E\left[\mathbf{z}^T\right] = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}E\left[z_i\right]E\left[z_j\right] \tag{16}$$

si suponemos cierta la igualdad entre tenemos,

$$\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}E\left[z_i z_j\right] - \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}E\left[z_i\right]E\left[z_j\right] = 0$$

$$\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}\left(E\left[z_i z_j\right] - E\left[z_i\right]E\left[z_j\right]\right) = 0$$

$$\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}Cov(z_i, z_j) = 0 \tag{17}$$

La igualdad sería cierta bajo ciertas condiciones, por ejemplo, si $\mathbf{z}$ fuera un vector constante, o si $\mathbf{A}$ fuera una matriz antisimétrica en la que todos los términos de su diagonal fueran siempre cero, entre otros casos. Sin embargo, en general, la igualdad no es verdadera para cualquier vector aleatorio $\mathbf{z}$ y cualquier matriz $\mathbf{A}$, por tanto, la igualdad es **FALSA**. Si tomamos como ejemplo $a_{11} = 1$ y el resto $a_{ij} = 0$, y un vector aleatorio tal que $z_1$ siga una distribución normal $N(\mu, \sigma)$, y el resto $z_i = 0$, en (17) obtenemos el valor de $\sigma^2 \neq 0$.

# 4. Se pide comentar qué hace el siguiente código en R e interpretar el resultado que proporciona

```
rm(list=ls())

library(daewr)
library(car)
datos = read.table('senn_32.txt', header =T)
attach(datos)

sapply(datos , class)

grupof <-as.factor(grupo)
sujetof <-as.factor(sujeto)
periodof <-as.factor(periodo)
tratf <-as.factor(trat)

options(contrasts=c(factor ="contr.treatment", ordered ="contr.poly"))

summary(lm(pef ~ sujetof + tratf ))
detach(datos)
```

`rm(list=ls())` borra todos los objetos guardados en la memoria actual del entorno de trabajo, útil al comenzar un nuevo script para así evitar confusiones.

`library(daewr)` proporciona funciones y marcos de datos utilizados en el libro *Design and Analysis of Experiments with R*.

`library(car)` proporciona una variedad de herramientas para el análisis de regresión, incluyendo gráficos de diagnóstico, pruebas de hipótesis y análisis de varianza.

`datos = read.table('senn_32.txt', header =T)` lee el archivo 'senn_32.txt' y guarda su contenido en el objeto llamado `datos`, `header = T` indica que en la primera fila se encuentran el nombre de las columnas; `attach(datos)` hace que se puedan acceder a alguna columna de `datos` sin la necesidad de escribir `datos$columna` cada vez que se quiera acceder a la misma.

`sapply(datos , class)` devuelve una lista de las clases de cada columna del dataframe `datos`, es decir, del tipo de datos de cada una de las columnas, en este caso concreto nos dice que como

'integer' tenemos a las columnas grupo, sujeto, periodo y pef, mientras que como 'character' tenemos a la columna trat.

Con grupof <-as.factor(grupo), sujetof <-as.factor(sujeto), periodof <-as.factor(periodo), tratf <-as.factor(trat), se crean nuevos vectores transformando las columnas grupo, sujeto, periodo y trat a 'factor', con el objetivo de trabajar con ellas como variables categóricas.

El código options(contrasts=c(factor ='contr.treatment', ordered ='contr.poly')) establece las opciones de contraste cuando trabajamos con variables categóricas (factores) o variables ordenadas, contr.treatment se elije una variable como nivel de referencia y se crean variables dummy que representan la presencia o ausencia de cada nivel, contr.treatment se usa para variables ordenadas, utilizando contrastes ortonormales.

lm(pef  sujetof + tratf ) hace un ajuste lineal donde se evalúa la variable pef en función de las variables predictoras categóricas sujetof y trat, la función summary() nos hace un resumen que comentaremos en breve sobre el ajuste lineal realizado.

Finalmente detach(datos) se utiliza para eliminar el dataframe datos adjuntado anteriormente.

Lo obtenido en summary(lm(pef  sujetof + tratf )) se puede ver en la Figura 1

```
Residuals:
   Min     1Q Median     3Q    Max
-42.31 -11.15   0.00  11.15  42.31

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       312.69      21.06  14.846 4.37e-09 ***
sujetof2           87.50      28.70   3.048 0.010114 *
sujetof3           65.00      28.70   2.265 0.042858 *
sujetof4           -5.00      28.70  -0.174 0.864616
sujetof5          105.00      28.70   3.658 0.003277 **
sujetof6           45.00      28.70   1.568 0.142915
sujetof7          110.00      28.70   3.832 0.002386 **
sujetof9           15.00      28.70   0.523 0.610773
sujetof10         -60.00      28.70  -2.090 0.058539 .
sujetof11          75.00      28.70   2.613 0.022678 *
sujetof12          10.00      28.70   0.348 0.733581
sujetof13        -135.00      28.70  -4.703 0.000511 ***
sujetof14          57.50      28.70   2.003 0.068264 .
tratfsalbutamol   -45.38      11.26  -4.031 0.001666 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.7 on 12 degrees of freedom
Multiple R-squared:  0.9286,    Adjusted R-squared:  0.8513
F-statistic: 12.01 on 13 and 12 DF,  p-value: 6.182e-05
```

Figura 1: Resultado del ajuste lineal del código del ejercicio

La primera sección (Residuals) muestra las discrepancias entre los valores observados y aquellos predichos por el modelo. El hecho de que la mediana sea cero y exista simetría entre los cuartiles, así como entre los residuos mínimos y máximos, sugiere que hay un equilibrio en la distribución de los residuos.

La segunda sección (Coefficients) muestra, en la primera columna (Estimate), los coeficientes de cada variable en el modelo. La segunda columna (Std. Error) presenta los errores asociados a cada uno de estos coeficientes. En la tercera columna (t value) se visualizan los valores t, obtenidos al dividir el coeficiente estimado entre su error estándar. Por último, la columna $Pr(>|t|)$ representa la probabilidad asociada al valor t; un valor p bajo (generalmente $< 0.05$) podría sugerir que el coeficiente es significativamente diferente de cero, lo que implicaría que la variable es importante para predecir la variable de respuesta.

En la tercera sección, el Error estándar residual (Residual standard error) se presenta como 28.7 unidades, indicando que, en promedio, los valores reales pueden desviarse aproximadamente por esa cantidad con respecto a los valores predichos por el modelo. El R-cuadrado Múltiple (Multiple R-squared) es una medida de la capacidad del modelo de regresión lineal para ajustarse a los datos.

Se calcula como la suma de los R-cuadrados de cada variable independiente en el modelo. Por otro lado, el R-cuadrado ajustado, similar al R-cuadrado múltiple, considera el número de variables independientes presentes. Esta medida es una versión ajustada del R-cuadrado. Un valor de R-cuadrado cercano a 1 indica que el modelo explica una gran parte de la variabilidad presente en la variable de respuesta. El Estadístico F (F-statistic) es una evaluación de la significancia global del modelo. Este estadístico analiza si al menos una de las variables independientes tiene un efecto significativo en la variable dependiente. Un valor cercano a 1 podría sugerir que ninguna de las variables independientes influye en la variable dependiente. El p-value, la probabilidad asociada al Estadístico F, cuando es bajo indica que el modelo es significativo y al menos una de las variables independientes tiene un efecto significativo en la variable dependiente.

El primer problema que identificamos es la alta cantidad de parámetros (14) en relación con la cantidad de datos disponibles (26). Esta disparidad puede conducir a un fenómeno conocido como 'sobreajuste' (overfitting), donde el modelo se ajusta excesivamente a los datos de entrenamiento y puede perder capacidad para generalizar a nuevos datos. En este contexto, los valores de p, que determinan la significancia de las variables, pueden carecer de fiabilidad y sentido. Los modelos sobreajustados pueden resultar en coeficientes que varían significativamente si los datos de entrenamiento se modifican, lo que afecta la robustez de los resultados del modelo. Por lo tanto, es prudente no depender únicamente de los asteriscos (indicadores de significancia) para evaluar la importancia de los parámetros en el modelo. Esto también influye en la interpretación del R cuadrado, dado que, aunque su valor se acerque a 1, es importante considerar que a medida que aumenta el número de parámetros en el modelo, el R-cuadrado puede incrementar incluso cuando las variables asociadas a esos parámetros no aportan información relevante. El enfoque de utilizar el estadístico F para probar la asociación entre las variables independientes y dependiente es válido cuando el valor de p (número de parametros) es relativamente pequeño en comparación con el tamaño de la muestra (n). Sin embargo, a pesar de obtener un valor de $F = 12{,}01$, aunque alejado de 1, tampoco podemos confiar plenamente en esta medida debido a lo mencionado anteriormente.

El segundo problema evidente es la falta de sentido en la creación de variables dummy en la categoría de 'sujeto'. Uno de los objetivos clave al realizar un ajuste lineal es lograr generalizar el modelo, es decir, hacer predicciones precisas con nuevos datos de manera que se adapten bien a la realidad. Sin embargo, si intentáramos predecir el flujo espiratorio máximo (pef) de un nuevo paciente a partir del modelo creado, se volvería imposible realizar dicha predicción. Esto se debe a que este nuevo paciente tendría un número diferente asignado en la categoría de 'sujeto'. Por lo tanto, para hacer una predicción, tendríamos que asignar un valor de 0 a cada una de las variables dummy 'sujetof', pero esta situación ocurre únicamente para el sujeto número 1. En resumen, estaríamos evaluando a todos los pacientes nuevos como si fueran el paciente número 1, lo que va en contra del propósito de generalizar el modelo para realizar predicciones precisas con una variedad de casos. Si eliminamos todas las variables dummy 'sujetof', nos quedaríamos únicamente con la variable dummy 'tratfsalbutamol'. En este caso, esta variable siempre proporcionaría solo dos valores distintos: uno para cuando un paciente tome formoterol y otro para cuando un paciente tome salbutamol, entonces pasaríamos directamente a un modelo con subajuste (underfitting), es decir, un modelo que es incapaz de capturar patrones importantes en los datos y no se ajusta adecuadamente ni siquiera a los datos de entrenamiento.

En resumen, estamos frente a un modelo que aparentemente muestra buenos resultados en su evaluación. Sin embargo, esta aparente eficacia se debe al uso de un número excesivo de parámetros en comparación con la cantidad de datos disponibles, lo cual conduce a un sobreajuste (overfitting). Aunque podríamos considerar la aplicación de técnicas para reducir la cantidad de parámetros y evitar el sobreajuste, esta acción nos llevaría a un subajuste debido a la falta de sentido en las variables 'sujetof'. En conclusión, este problema no parece ser adecuado para un modelo de regresión lineal (o de regresión en general). No obstante, los datos podrían ser útiles para un enfoque de clasificación, es decir, para predecir si un paciente toma salbutamol o formoterol.

# 5. Se ha realizado el siguiente diálogo con ChatGPT. Se pide dar una nota numérica de 0 a 10 justificando dicha puntuación a través de una rúbrica.

Para hacer la evaluación de las respuestas, vamos a puntuar cada pregunta del 0 al 10. Cada respuesta tiene em mismo peso, por lo que la puntuación final será la suma de las puntuaciones

individuales de cada respuesta dividido entre el número total de preguntas. Para tener una guía más o menos coherente a la hora de evaluar las respuestas vamos a seguir la siguiente rúbrica (Cuadro 2 ).

| **Puntuación** | **Criterios de Evaluación** |
|---|---|
| [0, 2.5) | Contenido insatisfactorio: La respuesta carece completamente de información relevante. La información carece de sentido. Contiene información incorrecta o no relacionada al tema. |
| [2.5, 5) | Contenido incompleto: poco desarrollado, respuestas muy generales, esquivando ser específico |
| [5, 7.5) | Contenido adecuado: proporciona información adecuada y relevante. Presenta argumentos razonables o detalles que respaldan el tema. Puede carecer de profundidad o claridad en ciertos aspectos. |
| [7.5, 10] | Contenido bueno: información completa y bien desarrollada. Presenta argumentos claros y detallados, respaldados por evidencia o ejemplos. Proporciona una comprensión sólida del tema sin imprecisiones. |

Cuadro 2: Rúbrica para Evaluar las respuestas

1. **¿Qué se entiende por multicolinealidad en un modelo de regresión lineal simple?**

   Menciona la presencia de una alta correlación entre dos o más variables predictoras. No obstante, esta afirmación carece de fundamentos, dado que en un modelo de regresión lineal simple solo interviene una variable predictora, por lo que la multicolinealidad no es un fenómeno presente en este tipo de modelo. Además, se comete un error al sugerir que este fenómeno podría afectar la precisión de las predicciones, ya que, en realidad, la multicolinealidad no incide directamente en la precisión predictiva del modelo. Incluso en presencia de multicolinealidad, el modelo debería mantener un nivel aceptable de precisión al predecir la variable objetivo.

   Siguiendo los criterios establecidos en la tabla de evaluación, esta respuesta se sitúa claramente en el nivel más bajo, como se indica en el Cuadro 2, debido a su falta de coherencia. Se otorga una puntuación de 1, ya que presenta una contradicción evidente en su argumentación.

2. **¿Qué se entiende por multicolinealidad en un modelo de regresión lineal mútiple?**

   En la primera parte de la respuesta se repite el error de la declaración anterior: *'La multicolinealidad en un modelo de regresión lineal múltiple es similar a lo que discutimos antes, pero se refiere específicamente a la correlación alta entre dos o más variables predictoras en un modelo con más de una variable independiente'*. Al afirmar esto, parece limitar un concepto general a un caso particular (aplicado a la regresión lineal múltiple). No es necesario mencionar *'específicamente'* dos o más variables predictoras, ya que esto es válido de manera general en el contexto de la multicolinealidad. En el segundo párrafo, todo lo mencionado es correcto. La multicolinealidad impacta negativamente en la precisión de los coeficientes de regresión, aumentando los errores estándar asociados a estos. Asimismo, conlleva problemas de interpretación, ya que no se puede discernir de manera precisa el efecto individual de cada variable independiente sobre la variable dependiente. En el tercer parrafo lo que se dice también es correcto, además da un simple ejemplo para que se entienda de manera intuitiva.

   La respuesta se encuentra en el cuarto nivel de la Rúbrica. Inicialmente contiene una imprecisión, pero logra explicar con solidez la pregunta planteada, por lo que se le dará una puntuación baja en ese nivel . Se le asigna una calificación de 8.

3. **Me imagino que una causa de multicolinealidad en un modelo de regresión lineal múltiple será la presencia de datos atípicos. ¿Es eso cierto?**

   En la respuesta, se separa adecuadamente el problema de los datos atípicos y la multicolinealidad, reconociendo que son fenómenos que requieren abordajes independientes. Es importante destacar que podemos encontrarnos con modelos exentos de multicolinealidad pero con datos atípicos presentes (como en un modelo de regresión lineal simple) o viceversa. No está comprobado que la presencia de datos atípicos cause multicolinealidad. Sin embargo, la definición

proporcionada sobre los datos atípicos, mencionándolos como valores extremos, podría ser más precisa si se explica que son aquellos puntos cuyos valores reales difieren significativamente de los valores predichos por el modelo. Además, se menciona que estos datos atípicos pueden afectar las estimaciones de los coeficientes de regresión; sin embargo, es común que un valor atípico tenga un impacto limitado en el ajuste de mínimos cuadrados. Es importante destacar que el término 'extremo' y su posible impacto en las estimaciones de los coeficientes podrían confundirse con un dato influyente (High Leverage Point). Es correcto afirmar que la presencia de datos atípicos puede impactar en la bondad de ajuste del modelo.

La respuesta se encuentra en el tercer nivel de la Rúbrica, hace bien en tratar ambos fenómenos de manera independiente, sin embargo es algo confusa la definición de valor atípico. Por esto obtiene una puntuación de 7.5

4. **Me imagino que una causa de multicolinealidad en un modelo de regresión lineal múltiple será la presencia de datos influyentes. ¿Es eso cierto?**

   En el primer párrafo, es acertado diferenciar ambos conceptos, aunque podría eliminarse el término 'ligeramente' para una mayor precisión. En el segundo párrafo, para una definición más completa de un dato influyente, podría añadirse que este tipo de dato tiene un valor predictivo que se aleja del rango normal de las observaciones. En el tercer párrafo, se menciona acertadamente que los datos influyentes pueden contribuir a la multicolinealidad, lo cual es respaldado por evidencia: *'From the examples, figures, and simulation results we observe that even a single high leverage point can generate huge multicollinearity'*[5]. Sin embargo, se genera cierta confusión al mencionar que los datos influyentes deben ser extremos (posiblemente refiriéndose a valores atípicos) y que podrían afectar la estimación de la relación entre las variables, lo que potencialmente contribuiría a problemas de multicolinealidad. Esto parece contradecir la afirmación anterior que negaba la relación directa entre valores atípicos y multicolinealidad. Además, no queda claro cómo la estimación de la relación entre una variable dependiente e independiente podría conllevar a problemas de multicolinealidad en la segunda parte del párrafo.

   La respuesta se encuentra en el tercer nivel de la Rúbrica, ya que acierta en la respuesta y separa los fenomenos, pero la definición de dato influyente podriía ser más precisa y llega a la contradicción de que tiene que ser un dato atípico para provocar multicolinealidad. Obtiene una puntuación de un 6.

   Como promedio final obtiene una calificación de **5.625**.

# 6. Se pide diseñar un estudio a nivel nacional para saber si a los españoles les gusta más la tortilla con cebolla o sin cebolla.

Para la realización del ejercicio, presentamos un ejemplo ficticio basado en un estudio realizado por el CIS (Centro de Investigaciones Sociológicas) [6], donde se planteaba la siguiente pregunta: ¿Cree Ud. que la tortilla española debe hacerse con cebolla o sin cebolla? Aunque la pregunta no coincide exactamente con la del ejercicio, su similitud nos permite tomar resultados ligeramente modificados para analizar los resultados en nuestro estudio ficticio.

## 6.1. Población

La **población** elegida será el conjunto de españoles residentes en España mayores de 18 años.

## 6.2. Muestreo

El tamaño de la muestra será de 4000 personas y se empleará un método de muestreo por afijación proporcional, utilizando cuotas de sexo y edad. Los estratos corresponden a las comunidades autónomas. Se aplicará un nivel de confianza del 95

## 6.3. Diseño del cuestionario

El cuestionario constará de la pregunta ¿Cómo prefiere le gusta más la tortilla con cebolla o sin cebolla? Teniendo las siguientes respuestas (solo se puede elegir una):

1. Me es indiferente
2. Me gusta más la tortilla con cebolla
3. Me gusta más la tortilla sin cebolla

## 6.4. Realización de las encuestas

Se seleccionan de manera aleatoria números de teléfonos fijos (20 %) y números de teléfonos móviles (80 %), aplicamos las cuotas de sexo y edad. Los estratos se forman a partir de las 17 comunidades autónomas y las dos ciudades autónomas.

## 6.5. Análisis de datos

Los resultados de la encuesta son los siguientes:

| Respuesta | Proporción |
|---|---|
| Me es indiferente. | 0.08 |
| Me gusta más la tortilla con cebolla | 0.70 |
| Me gusta más la tortilla sin cebolla | 0.22 |

Cuadro 3: Resultados de la encuesta

Ahora realizamos una prueba de hipótesis para determinar si la opción 'Me gusta más la tortilla con cebolla' es mayoritaria, es decir, representa más del 50 % (superando en preferencia a las otras dos opciones juntas). Si esto se confirma, se podría concluir que esta opción es más favorable que 'Me gusta más sin cebolla'.

$$H_0 : p = 0{,}5$$
$$H_1 : p > 0{,}5$$

Aplicamos el test clásico de proporción:

$$z = \frac{0{,}7 - 0{,}5}{\sqrt{\frac{0{,}7 * 0{,}3}{4000}}} = 27{,}60$$

Obteniendo un $p$-valor $< 2{,}2 \cdot 10^{-16}$, el cual es mucho menor que $\alpha = 0{,}05$, y un intervalo de confianza $(0{,}6854858, 0{,}7141271)$. Por tanto, existen evidencias suficientes para rechazar la hipótesis nula, es decir, que la proporción de personas a las que les gusta más la tortilla con cebolla es más del 50 %, indicando que hay más personas españolas residentes en España que prefieren la tortilla con cebolla.

# Referencias


[1] Gareth James, Daniela Witten, Trevor Hastie y Robert Tibshirani «An Introduction to Statistical Learning, Second Edition» , 2023.

[2] Chris Robert´s University of Manchester: Crossover. Recuperado de https://personalpages.manchester.ac.uk/staff/chris.roberts/MATH38071/Crossover.pdf

[3] PennState Eberly College of Science: Lesson 15: Crossover Designs. Recuperado de https://online.stat.psu.edu/stat509/book/export/html/749

[4] Ignacio Cascos Fernández: Vectores aleatorios. Recuperado de https://halweb.uc3m.es/esp/personal/personas/icascos/esp/resumen_vectores.pdf

[5] Md. Kamruzzaman y A. H. M. Rahmatullah Imon «High leverage point: another source of multicollinearity», Journal of Statistics , 2002.

[6] Centro de Investigaciones Sociológicas «Turismo y Gastronomía», pag 8 , 2023.