Modelos ANOVA

Los modelos ANOVA, llamados así por analysis-of-varaince (análisis de varianza), fueron creados para el análisis de experimentos biológicos. En este contexto, supongamos que un investigador aplica varios tratamientos distintos a unidades experimentales elegidas aleatoriamente, y posteriormente desea comparar las medias de cierta respuesta asociadas a los distintos tratamientos, más adelante veremos varios ejemplos de esta situación. En el modelo ANOVA hacemos uso de modelos lineales para realizar la comparación de estas medias. Este modelo frecuentemente es expresado con más parámetros de los que pueden ser estimados, lo cual resulta en una matriz $\boldsymbol{X'X}$ que no puede invertirse, por lo que alternativas a los métodos de estimación y pruebas basadas en mínimos cuadrados deben ser tratadas. Nos enfocaremos en modelos balanceados, donde se tiene el mismo número de observaciones para cada tratamiento.

ANOVA de una vía

Supongamos que un investigador a desarrollado dos químicos aditivos para incrementar el millaje de la gasolina. Para formular el modelo podemos plantear que que sin aditivos un galón de gasolina brinda un millaje promedio de μ kilómetros. Sii el aditivo 1 es añadido, entonces el millaje promedio incrementa τ_1 kilómetros por galón; y si el aditivo 2 es añadido, el millaje promedio aumenta τ_2 kilómetros por galón. Si y_1 denota el kilometraje por galón asociado a un tanque de gasolina con el aditivo químico 1, y similarmente y_2 para el aditivo 2; entonces el modelo puede ser epresado como sigue:

$$y_1 = \mu + \tau_1 + \epsilon_1, \quad y_2 = \mu + \tau_2 + \epsilon_2$$

Con ϵ_1 y ϵ_2 errores gaussianos. En este conteto es natural querer realizar estimaciones de los parámetros μ, τ_1 y τ_2, y pruebas de hipótesis como $H_0 \ : \ \tau_1 = \tau_2$.

Supongamos que el investigador diseña un experimento en el cual se observa el kilometraje por galón para 3 tanques con el aditivo 1 y 3 tanques con el aditivo 2. Esto se puede escribir como

$$y_{1,1} = \mu + \tau_1 + \epsilon_{1,1}, \qquad y_{1,2} = \mu + \tau_1 + \epsilon_{1,2}, \ \ y_{1,3} = \mu + \tau_1 + \epsilon_{1,3},$$
$$y_{2,1} = \mu + \tau_2 + \epsilon_{2,1}, \qquad y_{2,2} = \mu + \tau_2 + \epsilon_{2,2}, \ \ y_{2,3} = \mu + \tau_2 + \epsilon_{2,3},$$

o alternativamente escrito como

$$y_{i,j} = \mu + \tau_i + \epsilon_{i,j}, \quad i = 1, 2; \ j = 1, 2, 3$$

En forma matricial escribimos esto como

$$\begin{pmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{2,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{1,3} \\ \epsilon_{2,1} \\ \epsilon_{2,2} \\ \epsilon_{2,3} \end{pmatrix}$$

o alternativamente escrito como

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

Observamos que $\boldsymbol{X'X}$ no es invertible por lo que el estimador de mínimos cuadrados no existe para este modelo.

In [1]:
```
X = zeros(6,3) # Matriz de ceros para construir X
for i in 1:3
    X[i,:] = [ 1, 1, 0 ] # renglones asociados al aditivo 1
end
for i in 4:6
    X[i,:] = [ 1, 0, 1 ] # renglones asociados al aditivo 2
end
X'*X
```

Out[1]:
```
3×3 Array{Float64,2}:
 6.0  3.0  3.0
 3.0  3.0  0.0
 3.0  0.0  3.0
```

La primer columna de $\boldsymbol{X}'\boldsymbol{X}$ menos la segunda resulta en la tercera columna por lo que la matrix no es invertible.

In [2]:
```
(X'*X)^(-1)
```

```
LinearAlgebra.SingularException(3)

Stacktrace:
 [1] checknonsingular at /buildworker/worker/package_linux64/build/usr/share/julia/stdlib/v1.2/LinearAlgebra/src/factorization.jl:12 [inlined]
 [2] #lu!#118(::Bool, ::typeof(LinearAlgebra.lu!), ::Array{Float64,2}, ::Val{true}) at /buildworker/worker/package_linux64/build/usr/share/julia/stdlib/v1.2/LinearAlgebra/src/lu.jl:41
 [3] #lu! at ./none:0 [inlined]
 [4] #lu#122 at /buildworker/worker/package_linux64/build/usr/share/julia/stdlib/v1.2/LinearAlgebra/src/lu.jl:142 [inlined]
 [5] lu at /buildworker/worker/package_linux64/build/usr/share/julia/stdlib/v1.2/LinearAlgebra/src/lu.jl:142 [inlined] (repeats 2 times)
 [6] inv(::Array{Float64,2}) at /buildworker/worker/package_linux64/build/usr/share/julia/stdlib/v1.2/LinearAlgebra/src/dense.jl:730
 [7] literal_pow(::typeof(^), ::Array{Float64,2}, ::Val{-1}) at /buildworker/worker/package_linux64/build/usr/share/julia/stdlib/v1.2/LinearAlgebra/src/generic.jl:925
 [8] top-level scope at In[2]:1
```

Observamos que el modelo tiene 3 parámetros a estimar pero $\text{rango}(\boldsymbol{X}) = 2$ por lo que decimos que el modelo está sobreparametrizado. Esto se puede ilustrar con el siguiente ejemplo. Si $\mu = 15$, $\tau_1 = 1$ y $\tau_2 = 3$ obtenemos

$$y_{1,j} = 15 + 1 + \epsilon_{1,j} = 16 + \epsilon_{1,j}, \quad j = 1, 2, 3$$
$$y_{2,j} = 15 + 3 + \epsilon_{2,j} = 18 + \epsilon_{2,j}, \quad j = 1, 2, 3$$

es equivalente a si $\mu = 10$, $\tau_1 = 6$ y $\tau_2 = 8$ con lo que

$$y_{1,j} = 10 + 6 + \epsilon_{1,j} = 16 + \epsilon_{1,j}, \quad j = 1, 2, 3$$
$$y_{2,j} = 10 + 8 + \epsilon_{2,j} = 18 + \epsilon_{2,j}, \quad j = 1, 2, 3$$

y claramente hay una infinidad de valores para μ, τ_1 y τ_2 que deterrminan el mismo modelo.

Veremos dos alternativas para tratar el modelo ANOVA, pero primero discutiremos el modelo ANOVA de dos vías.

ANOVA de dos vías

Supongamos que un investigador quiere investigar el efecto de dos vitáminas y de dos mecanismo de administración con respecto a la ganancia en peso para pollos. Sean α_1, α_2 los efectos en la ganacia de peso para pollos asociados a las vitaminas y β_1, β_2 los efectos relacionados a los mecanismo de administracion. Entonces se puede proponer el siguiente modelo donde los efectos son aditivos

$$y_{1,1} = \mu + \alpha_1 + \beta_1 + \epsilon_{1,1}, \quad y_{1,2} = \mu + \alpha_1 + \beta_2 + \epsilon_{1,2}$$
$$y_{2,1} = \mu + \alpha_2 + \beta_1 + \epsilon_{2,1}, \quad y_{2,2} = \mu + \alpha_2 + \beta_2 + \epsilon_{2,2}$$

Para simplificar la exposición, suponemos que el investigador observa una observación de cada combinación vitamina - administración. En forma matricial esto se expresa como

$$\begin{pmatrix} y_{1,1} \\ y_{1,2} \\ y_{2,1} \\ y_{2,2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,1} \\ \epsilon_{2,2} \end{pmatrix}$$

o alternativamente

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

Observamos que $\boldsymbol{X}'\boldsymbol{X}$ no es ivertible

In [6]:
```
X = [ [1,1,1,1] [1,1,0,0] [0,0,1,1] [1,0,1,0] [0,1,0,1] ]
```

Out[6]:
```
4×5 Array{Int64,2}:
 1  1  0  1  0
 1  1  0  0  1
 1  0  1  1  0
 1  0  1  0  1
```

In [7]:
```
X' * X
```

Out[7]:
```
5×5 Array{Int64,2}:
 4  2  2  2  2
 2  2  0  1  1
 2  0  2  1  1
 2  1  1  2  0
 2  1  1  0  2
```

La primera columna de $\boldsymbol{X}'\boldsymbol{X}$ menos la segunda resulta en la tercera columna por lo que la matriz no es invertible.

Reparametrización

Veremos que si tranformamos el modelo podemos llegar a uno donde el estimador usual de mínimos cuadrados para el modelo de regresión lineal múltiple puede ser calculado. Transformamos el modelo $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ donde $\boldsymbol{X}$ es $n \times p$ dimensional con rango $k < p \leq n$ en un modelo $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ con $\boldsymbol{Z}$ $n \times k$ dimensional con rango k y $\boldsymbol{\gamma} = \boldsymbol{U}\boldsymbol{\beta}$ con los renglones de $\boldsymbol{U}$ linealmente independientes. Entonces

$$\boldsymbol{Z}\boldsymbol{\gamma} = \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}$$

por lo que $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{U}$. Dado que $\boldsymbol{U}$ es $k \times p$ dimensional con rango $k < p$, la matriz $\boldsymbol{U}'\boldsymbol{U}$, análogamente a cuando discutíamos el modelo de regresión lineal múltiple, es no singular por lo que

$$\boldsymbol{Z}\boldsymbol{U} = \boldsymbol{X} \implies \boldsymbol{Z}\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{X}\boldsymbol{U}' \implies \boldsymbol{Z} = \boldsymbol{X}\boldsymbol{U}'(\boldsymbol{U}\boldsymbol{U}')^{-1}$$

Observamos que $\boldsymbol{Z}$ es de rango completo ya que $\text{rango}(\boldsymbol{Z}) \geq \text{rango}(\boldsymbol{ZU}) = \text{rango}(\boldsymbol{X}) = k$ y $\text{rango}(Z) \leq \min\{n, k\} = k$ por lo que \pmb{Z} $= k$ y podemos usar el estimador de mínimos cuadrados $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$.

Condiciones adicionales

Consideramos el modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ donde $\boldsymbol{X}$ es $n \times p$ dimensional con rango $k < p \leq n$. La deficiencia en el rango de $\boldsymbol{X}$ es de $p - k$ por lo que consideramos condiciones adicionales dadas por $\boldsymbol{T\beta} = \boldsymbol{0}$ con $\boldsymbol{T}$ $(p - k) \times p$ dimensional tal que los correspondientes renglones son linealmente independientes entre si y entre los renglones de $\boldsymbol{X}$. Entonces podemos considerar las ecuaciones

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon 0} \qquad = \boldsymbol{T\beta} + \boldsymbol{0}$$

o alternativamente

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{0} \end{pmatrix}$$

La matriz $\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}$ es entonces $(n + p - k) \times p$ dimensional de rango p por lo que $\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}' \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}$ es $p \times p$ dimensional de rnago p y las ecuaiones normales

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}' \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix} \hat{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}' \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}$$

tienen la solucion única

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}' \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix} \right)^{-1} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix}' \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \\ &= \left((\boldsymbol{X}', \boldsymbol{T}') \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{T} \end{pmatrix} \right)^{-1} (\boldsymbol{X}', \boldsymbol{T}') \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \\ &= \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{T}'\boldsymbol{T} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y}, \boldsymbol{T}'\boldsymbol{0} \right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{T}'\boldsymbol{T} \right)^{-1} \boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

Ejercicio de tarea

Consideramos el modelo ANOVA

$$y_{i,j} = \mu + \tau_i + \epsilon_{i,j}, \quad i = 1, 2; \; j = 1, 2;$$

escrito en forma matricial como

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,1} \\ \epsilon_{2,2} \end{pmatrix}$$

De los estimadores de mínimos cuadrados haciendo uso tanto del método de reparametrización tanto como el de condiciones adicionales, encontrando las matrices $\boldsymbol{U}$ y $\boldsymbol{T}$. Intente encontrar reparametrizaciones y condiciones que tengan una interpretación respecto a los parámetros del modelo.

In [ ]: