Regresión y ANOVA: Analgésicos Infantiles *

García Prado, Sergio
sergio@garciparedes.me

15 de noviembre de 2017

1. Contexto y Conjunto de Datos

En este trabajo se va a realizar un estudio acerca de la diferencia de medias sobre el conjunto de datos `painkillers`, el cual se refiere a un experimento sobre *Analgésicos Infantiles*. Para ello, se utilizará la técnica de *Análisis de la Varianza (ANOVA)*. Una contextualización más detallada del experimento se describe a partir del siguiente enunciado:

"El departamento de pediatría de un hospital desea analizar la eficacia de cuatro analgésicos infantiles ante las cefaleas. Para ello, realiza un experimento en el que se seleccionan aleatoriamente cinco grupos de cuatro pacientes, de manera que en cada grupo se da un cefalea distinto. A continuación se suministra, también de forma aleatoria, cada analgésico a uno de los pacientes de cada grupo, y se observa el tiempo de remisión de la cefalea, en minutos. Se registran los datos siguientes, en cada uno de los cinco grupos (`tiempo de remisión`, `analgésico` y `cefalea`)."

2. Cuestiones

En esta sección se incluyen una serie de cuestiones que serán resueltas mediante el estudio del conjunto de datos a partir de la técnica *ANOVA*.

2.1. Estudia el tipo de diseño adecuado para esta situación, e identifica las variables, factores y parámetros

Tras analizar el contexto del experimento, se sabe que la variable respuesta Y que se utilizará para el *análisis de la varianza* es `tiempo de remisión`, dado que es la que se utiliza para cuantificar la calidad del tratamiento. En cuanto a las variables a partir de las cuales se pretende explicar el `tiempo de remisión`, estas son el `analgésico` y el `cefalea`.

Sin embargo, estas no han sido seleccionadas de la misma manera, dado que tal y como se indica en el enunciado, se han fijado 4 muestras de *a priori* 5 `tipos de cefalea`, sobre los cuales aplicar los 4 tipos de `analgésico`. Por tanto, podemos interpretar dicha situación diciendo que el factor `tipos de cefalea` representa un **bloque** (que denotaremos por B_β siendo $\beta \in \{1, 2, 3, 4, 5\}$ el identificador de cada uno de los niveles del bloque), y asumiento que el factor `analgésico` representa un **tratamiento** (que denotaremos por T_α siendo $\alpha \in \{A, B, C, D\}$ el identificador de cada uno de los niveles del tratamiento).

*URL: `https://github.com/garciparedes/anova-painkillers`

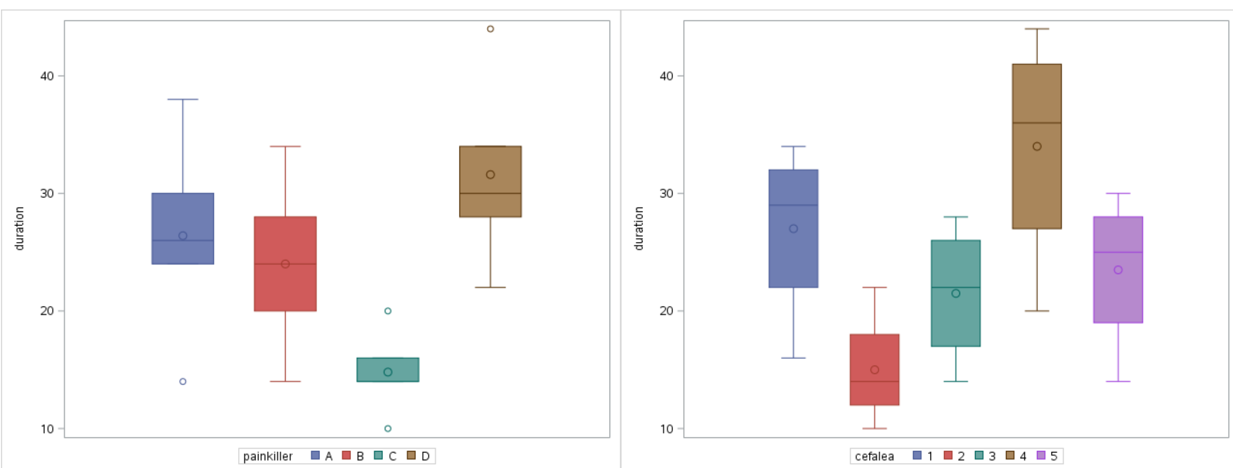
$$Y_{ij} = \mu + T_\alpha + B_\beta + \epsilon_{\alpha\beta k} \qquad \begin{aligned} &\alpha \in \{1, ..., n_\alpha\} \\ &\beta \in \{1, ..., n_\beta\} \\ &k \in \{1, ..., n_{\alpha\beta}\} \end{aligned} \tag{1}$$

Por dichas razones, utilizaremos el modelo de *1 factor + 1 bloque*, el cual se muestra en la ecuación (1). A este modelo se le ha añadido además la componente $\epsilon_i j$, que representa el error aleatorio y sigue una distribución $N(0, \sigma^2)$, asumiendo que σ^2 es la misma para todas las observaciones.

El contraste test de igualdad de medias se describe tal y como se indica en la ecuación (2), que tal y como se puede apreciar, se lleva a cabo teniendo en cuenta únicamente las medias del factor tratamiento, ya que el factor bloque es utilizado únicamente para reducir la variabilidad del modelo.

$$\begin{aligned} &H_0 : \forall i, j \ \mu_i = \mu_j \\ &H_1 : \exists i, j \ \mu_i \neq \mu_j \end{aligned} \qquad i, j \in \{1, ..., n_\alpha\}, \ i \neq j \tag{2}$$

En las figuras 1a y 1b se muestran los diagramas de caja de la variable respuesta `tiempo de remisión` agrupados por `analgésico` y `cefalea` respectivamente. En estos gráficos se puede apreciar la existencia de diferencias entre las distintas agrupaciones. Por tanto, tiene sentido el estudio del *análisis de la varianza* sobre estas para obtener conclusiones más claras de manera analítica.



(a) `analgésico` (b) `cefalea`

Figura 1: Diagramas de cajas

2.2. ¿Es adecuado usar las cefaleas como bloques?. ¿Existen diferencias significativas entre los tiempos de remisión de las cefaleas para los distintos analgésicos?

A partir del contexto del conjunto de datos se ha asumido que el experimento se ha realizado bloqueando el factor tipo de `cefalea`. Además, tras visualizar el diagrama de cajas de dicha variable en la figura 1b se puede apreciar la existencia de gran variabilidad referida a este factor.

Dicha variabilidad sería tratada como ruido en el modelo, por lo que se ha decidido recoger en un bloque. Los resultados del test de igualdad de medias sobre este modelo se muestran en la figura 2. Tal y como se puede apreciar, la variabilidad recogida por el bloque `cefalea` es muy elevada, además se rechaza la hipótesis nula por lo que el bloqueo ha sido positivo para el modelo.

The GLM Procedure
Dependent Variable: duration

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1525.200000	217.885714	35.33	<.0001
Error	12	74.000000	6.166667		
Corrected Total	19	1599.200000			

R-Square	Coeff Var	Root MSE	duration Mean
0.953727	10.26148	2.483277	24.20000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
painkiller	3	740.0000000	246.6666667	40.00	<.0001
cefalea	4	785.2000000	196.3000000	31.83	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
painkiller	3	740.0000000	246.6666667	40.00	<.0001
cefalea	4	785.2000000	196.3000000	31.83	<.0001

Figura 2: Resultados del test de igualdad de medias sobre el *ANOVA* de 1 factor y un bloque

En cuanto a las diferencias entre los distintos tratamientos del factor `analgésico`, se puede asegurar que estas son significativas con una confianza del 99 % debido al rechazo de la hipótesis nula de igualdad de medias.

2.3. Estudia gráficamente la existencia de interacción entre analgésico y cefalea.

En las figuras 3a y 3b se muestran los diagramas de interacción de las variables `analgésico` y `cefalea`. A partir de estos, se puede apreciar la no existencia de interacción entre ellos, ya que en ambos casos las rectas permanecen paralelas entre sí. A pesar de ello, estas son traslacciones verticales unas de otras, por lo que se puede asumir la existencia de diferencias entre las distintos tratamientos, pero no la interacción entre ellos.

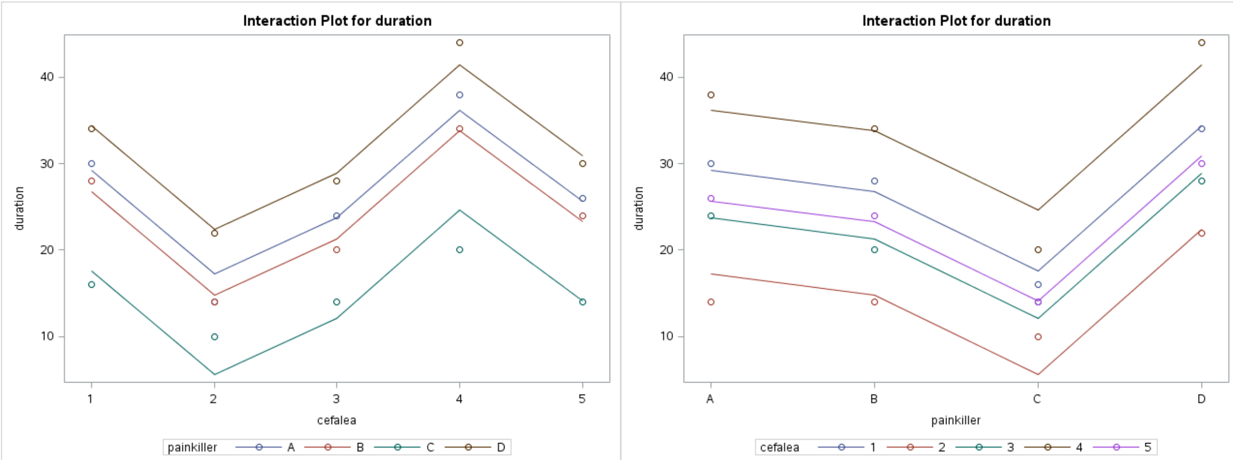
2.4. Determina mediante comparaciones múltiples cuál de los analgésicos es más eficaz.

Como método de comparaciones múltiples se ha escogido *Tukey* por su grado de potencia, con respecto a los requisitos necesarios para su utilización (número igual de observaciones en todos los tratamientos). Los resultados obtenidos se muestran en la figura **??**.

Tal y como se puede apreciar en los resultados, el **analgésico C** es el más eficaz para la reducción del tiempo de cefalea, mientras que los analgésicos A y C podrían ser considerados igual de eficaces puesto que no se rechaza la hipótesis de igualdad de medias entre ellos.

2.5. Suponiendo que el analgésico A es un placebo, realiza el test de Dunnett

[TODO]



(a) `analgésico` (b) `cefalea`

Figura 3: Diagramas de Interacción

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

painkiller	duration LSMEAN	LSMEAN Number
A	26.4000000	1
B	24.0000000	2
C	14.8000000	3
D	31.6000000	4

Least Squares Means for effect painkiller
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: duration

i/j	1	2	3	4
1		0.4520	<.0001	0.0276
2	0.4520		0.0004	0.0020
3	<.0001	0.0004		<.0001
4	0.0276	0.0020	<.0001	

Figura 4: Resultados obtenidos en el método de *Tukey*

2.6. Haz un análisis gráfico de los residuos.

[TODO]

2.7. Si los analgésicos se hubieran elegido al azar entre todos los existentes, plantea el modelo adecuado y estima las componentes de la varianza.

En este caso se pide tratar el factor `analgésico` como un factor aleatorio. Por tanto, el modelo a utilizar se define de manera diferente. Este se muestra en la ecuación (3). A pesar de ser equivalente al anterior a nivel de notación, tiene una interpretación diferente tal y como se verá a continuación.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

painkiller	duration LSMEAN	H0:LSMean=Control Pr > |t|
A	26.4000000	
B	24.0000000	0.3320
C	14.8000000	<.0001
D	31.6000000	0.0162

Figura 5: Resultados obtenidos en el método de *Dunnett*

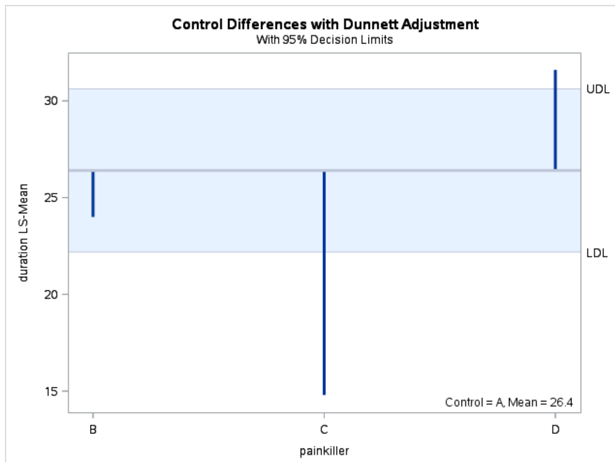
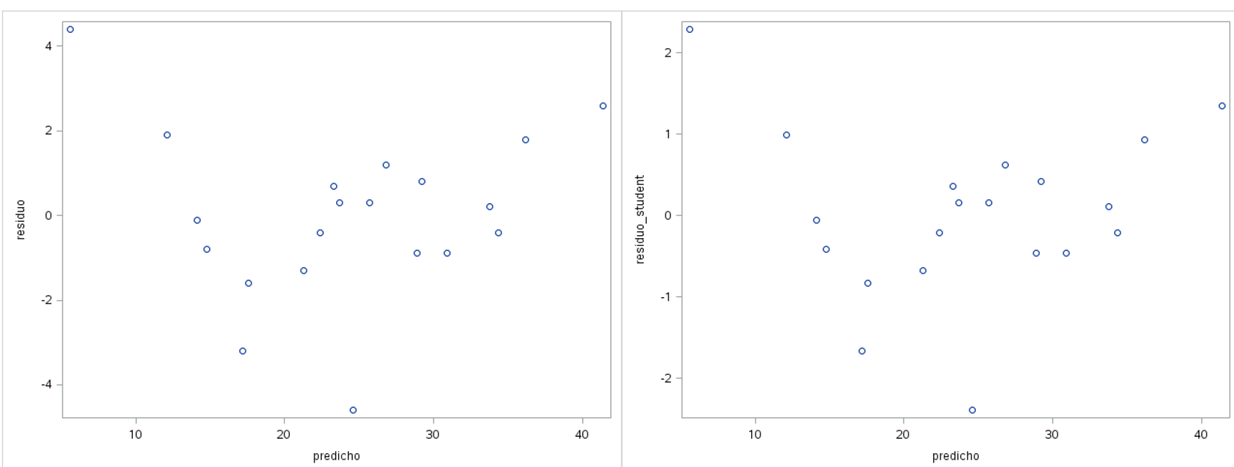


Figura 6



(a) Resodios Estándar

(b) Residuos Studentizados

Figura 7: Diagramas de Residuos

$$Y_{ij} = \mu + \mathcal{T}_\alpha + B_\beta + \epsilon_{\alpha\beta k} \qquad \begin{array}{l} \alpha \in \{1, ..., n_\alpha\} \\ \beta \in \{1, ..., n_\beta\} \\ k \in \{1, ..., n_{\alpha\beta}\} \end{array} \tag{3}$$

En este caso el factor tratamiento T_α pasa a denotarse como $\mathcal{T}_\alpha$ y en lugar de representar la diferencia respecto del fator α como un valor escalar, ahora se refiere a una variable aleatoria $\mathcal{T} \sim N(0, \sigma^2_{\mathcal{T}})$. Gracias a esta diferencia, el test *ANOVA* se formula de una manera diferente, en este caso se estudia el efecto de la varianza $\sigma^2_{\mathcal{T}}$, tal y como se indica en la ecuación (4).

$$\begin{aligned} H_0 &: \sigma^2_{\mathcal{T}} = 0 \\ H_1 &: \sigma^2_{\mathcal{T}} > 0 \end{aligned} \tag{4}$$

Desde esta perspectiva, ya no tiene sentido estudiar si un tratamiento genera mejores resultados que otro, sino que únicamente se analiza desde el punto de vista de la homogeneidad de la población.

Tras realizar el *ANOVA* utilizando este modelo, los resultados obtenidos se muestran en las figuras 8 y 9. A partir de dichos resultados se obtiene la conclusión de que se trata de una población heterogénea, es decir, con distintas subpoblaciónes, dado que la hipótesis nula es rechazada con un $p - valor$ muy próximo a cero.

The GLM Procedure

Source	Type III Expected Mean Square
painkiller	Var(Error) + 5 Var(painkiller)
cefalea	Var(Error) + Q(cefalea)

Figura 8: Resultados de descomposición de la varianza de *ANOVA* aleatorizado

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance
Dependent Variable: duration

Source	DF	Type III SS	Mean Square	F Value	Pr > F
painkiller	3	740.000000	246.666667	40.00	<.0001
cefalea	4	785.200000	196.300000	31.83	<.0001
Error: MS(Error)	12	74.000000	6.166667		

Figura 9: Resultados del test de igualdad de medias sobre el *ANOVA* aleatorizado

A partir de estos resultados puede obtenerse una estimación acerca de la varianza del factor `analgésico`, la cual se realiza a continuación

$$\begin{aligned} Var(\text{Error}) + 5Var(\text{analgésico}) &= 246.\widehat{6} \\ Var(\text{Error}) &= 6{,}1\widehat{6} \\ \implies Var(\text{analgésico}) &= \frac{246.\widehat{6} - 6{,}1\widehat{6}}{5} \approx 48{,}1 \end{aligned}$$

2.8. Plantea el modelo como diseño unifactorial completamente aleatorizado y compara los resultados.

En esta sección se realiza el test de igualdad de medias asumiendo la selección complemtamente aleatoria de las observaciones respecto del factor `analgésico`, de tal manera que se ignora la agrupación por tipos de `cefalea`. Por tango, el modelo descrito en la ecuación (1) puede reducirse al que se muestra en la ecuación (5).

Sin embargo, en este caso, toda la variabilidad procedente del bloqueo del factor tipo de `cefalea` se convierte en error no explicado por el modelo. Es decir, ahora lo recoge la variable $\epsilon_{\alpha j}$.

$$Y_{ij} = \mu + T_\alpha + \epsilon_{\alpha j} \qquad \begin{array}{l} i \in \{1, ..., n_\alpha\} \\ j \in \{1, ..., n_i\} \end{array} \tag{5}$$

El test para la igualdad de medias, en este caso se plante de manera equivalente, tal y como se indica en la ecuación (6). Puesto que la variabilidad procedente del tipo de cefalea ahora es recogido por el error no explicado por el modelo ($\epsilon_{\alpha j}$), las conclusiones que se puedan obtener a partir del estudio del análisis de la varianza deberán ser tomadas con mayor prudencia.

$$\begin{array}{l} H_0 : \forall i, j \;\; \mu_i = \mu_j \\ H_1 : \exists i, j \;\; \mu_i \neq \mu_j \end{array} \qquad i, j \in \{1, ..., n_\alpha\}, \;\; i \neq j \tag{6}$$

En la figura 10 se muestran los resultados obtenidos tras las realización del estudio *ANOVA* de un factor. Tal y como se puede apreciar, en este caso también se debe rechazar la hipótesis de igualdad de medias, puesto que el *p-valor* del test es muy bajo (0,0167). Por tanto, se puede asegurar con un nivel de confianza del 98 % que existen diferencias significativas entre los resultados de los distintos analgésicos.

The GLM Procedure
Dependent Variable: duration

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	740.000000	246.666667	4.59	0.0167
Error	16	859.200000	53.700000		
Corrected Total	19	1599.200000			

R-Square	Coeff Var	Root MSE	duration Mean
0.462731	30.28111	7.328028	24.20000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
painkiller	3	740.0000000	246.6666667	4.59	0.0167

Source	DF	Type III SS	Mean Square	F Value	Pr > F
painkiller	3	740.0000000	246.6666667	4.59	0.0167

Figura 10: Resultados del teste de igualdad de medias sobre el *ANOVA* de un factor

Sin embargo, cabe destacar que estos resultados deben ser tomados con mucha prudencia, debido a que en este caso la variabilidad del error es mucho mayor que en el modelo de *1 factor + 1 bloque*, variando el coeficiente de determinación R^2 del valor 0,95 a 0,46. Es decir, con este modelo se explica aproximadamente la mitad de la variabilidad, por tanto no es un modelo acertado para este problema.

3. Código Fuente

En esta sección se incluyen los distintos procedimientos de código *SAS* utilizados para la realización de este trabajo.

```
data painkillers;
  input duration painkiller$ cefalea;
  datalines;
  30 A 1
  28 B 1
  16 C 1
  34 D 1
  14 A 2
  14 B 2
  10 C 2
  22 D 2
  24 A 3
  20 B 3
  14 C 3
  28 D 3
  38 A 4
  34 B 4
  20 C 4
  44 D 4
  26 A 5
  24 B 5
  14 C 5
  30 D 5
;
run;
proc print data=painkillers;
run;
```

Figura 11: *Código SAS:* Lectura del conjunto de datos.

```
proc sgplot data=painkillers;
  vbox duration / group=painkiller;
run;

proc sgplot data=painkillers;
  vbox duration /group=cefalea;
run;
```

Figura 12: *Código SAS:* Generación de los diagramas de cajas.

```
proc glm data=painkillers;
  *class cefalea painkiller;
  class painkiller cefalea;
  model duration=painkiller cefalea ;
  lsmeans painkiller / adjust=tukey;
  lsmeans painkiller / adjust=dunnett;
  random painkiller / test;
  output out=soluc P=predicho R=residuo student=residuo_student;
run;
```

Figura 13: *Código SAS:* Realización de ANOVA por bloques.

```
proc sgplot data=soluc;
  scatter y=residuo x=predicho;
run;

proc sgplot data=soluc;
  scatter y=residuo_student x=predicho;
run;
```

Figura 14: *Código SAS:* Generación de diagramas de residuos.

```
proc glm data=painkillers;
  class painkiller;
  model duration=painkiller;
run;
```

Figura 15: *Código SAS:* Realización de ANOVA completamente aleatorizado.

Referencias

[1] Barba Escribá, L. Regresión y ANOVA, 2017/18. Facultad de Ciencias: Departamento de Estadística.

[2] SAS® Software Institute. Sas. https://www.sas.com/.