EL MODELO LINEAL

Clase 6

Análisis de la varianza

Sergio Camiz

LIMA - Marzo-Mayo 2025

Asuntos de la clase 7

- Análisis de la varianza
- Análisis de la varianza de dos vías
- Ejemplos

Análisis de la varianza

A proposito de la estimacion de una variable respuesta cuantitativa a través de una cualitativa, se ha visto que considerando el promedio de la respuesta por cada modalidad cualitativa, se resulta que la suma de los cuadrados de los valores observados se puede descomponer en la suma de los cuadrados de los promedios en cada modalidad de la variable descriptiva, pesada para su frecuéncia, más la parte correspondiente al desvío cuadrático de los valores observados al promedio de su modalidad.

Con n observaciones en p modalidades con numerosidades $n_1 + n_2 + \cdots + n_p = n$, se resulta

$$SS_T = \sum_{i=1}^{n} y_i^2 = \sum_{k=1}^{p} n_k \bar{y}_k^2 + \sum_{k=1}^{p} \sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2 = SS_B + SS_W \tag{1}$$

con la tabla de análisis de varianza

Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)	F
Between	p	SS_B	$MS_B = SS_B/p$	MS_B/MS_W
Within	$n-p$	SS_W	$MS_W = SS_W/(n-p)$	
Total	n	SS_T		

Vemos como tratar este problema a través del modelo lineal.

Supongamos de tenere una $\boldsymbol{y}$ respuesta que queremos modelar a través de las modalidades de una $\boldsymbol{x}$ *cualitativa*. Claro que no podemos escribir el modelo como

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \varepsilon$$

ya que multiplicar β para un nivel (pelo rubio, ojos azules) no tiene sentido.

- Por otro lado, podemos pensar que nuestro interés es de poder estimar los promedios de las observaciones de cada clase o también su desvíos por respecto al promedio total (Ec. (1)).
- Ya hemos visto que efectivamente la estimación de mínimos cuadrados $E(\hat{\boldsymbol{\eta}}_x) = E(\boldsymbol{\eta}_x) = \bar{\boldsymbol{y}}_x$.
- Entonces, todo depende de como codificar a $\boldsymbol{x}$ para poder estimar $\boldsymbol{\eta}$, que será un vector de promedios estimados.
- Por esto se considera cada nivel de $\boldsymbol{x}$ como una variable separada, que solo asume valores 1 si la observación tiene la modalidad y 0 si no la tiene.
- A estas se le llaman *variables indicadoras*, y el conjunto $\boldsymbol{X}$ que se resulta corresponde a escribir $\boldsymbol{x}$ en *forma disjuntiva completa*.

Supongamos que $\boldsymbol{x}$ tiene p modalidades diferentes. El modelo sería entonces:

$$\boldsymbol{\eta} = \alpha + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p + \boldsymbol{\varepsilon}.$$

Como las columnas son todas de cero y un, es sencillo demonstrar que, con este modelo, α representaría el promedio general $\bar{\boldsymbol{y}}$ y que cada β_j representaría el desvío a $\bar{\boldsymbol{y}}$ del promedio de las observaciónes del nivel j, o sea.

$$\boldsymbol{\eta} = \bar{\boldsymbol{y}} + (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})\boldsymbol{x}_1 + \cdots + (\bar{\boldsymbol{y}}_p - \bar{\boldsymbol{y}})\boldsymbol{x}_p + \boldsymbol{\varepsilon}.$$

Entonces, esta es la formulación del modelo.

Del punto de vista del cálculo, considerando tres niveles la transformación es:

$$\boldsymbol{x} = \begin{pmatrix} \text{rojo} \\ \text{azul} \\ \text{azul} \\ \text{rojo} \\ \text{verde} \\ \text{azul} \\ \text{verde} \\ \text{azul} \\ \cdots \end{pmatrix} \longrightarrow \boldsymbol{X} = \begin{array}{c} \begin{array}{cccc} \alpha & \text{rojo} & \text{azul} & \text{verde} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} \end{array}$$

Pero noten que la suma de las tres columnas se corresponde a la columna α, así que la matriz $\boldsymbol{X}'\boldsymbol{X}$ no es invertible:


```
t(x)%*%x
     [,1] [,2] [,3] [,4]
[1,]    8    3    3    2
[2,]    3    3    0    0
[3,]    3    0    3    0
[4,]    2    0    0    2
> m <-t(x)%*%x
> solve(m)
Error in solve.default(m) :
system is computationally singular:
reciprocal condition number = 3.46945e-18
```

De hecho se reconoce que la primeras filas y columna son la suma de las demás.

Al contrario, tirando la columna **1**,

$$\boldsymbol{x} = \begin{pmatrix} \text{rojo} \\ \text{azul} \\ \text{azul} \\ \text{rojo} \\ \text{verde} \\ \text{azul} \\ \text{verde} \\ \text{azul} \\ \dots \end{pmatrix} \longrightarrow \boldsymbol{X} = \begin{pmatrix} \text{rojo} & \text{azul} & \text{verde} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \dots & \dots & \dots \end{pmatrix} \tag{2}$$

donde se resulta

$$\boldsymbol{X'X} = \begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & n_p \end{pmatrix} \text{ así que } (\boldsymbol{X'X})^{-1} = \begin{pmatrix} 1/n_1 & 0 & \dots & 0 \\ 0 & 1/n_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/n_p \end{pmatrix}$$

y $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{Xy} = (\bar{y}_1, \dots, \bar{y}_p)'$ los promedios de las clases.

Además, $\hat{\boldsymbol{\eta}}_x = \bar{\boldsymbol{y}}_x$.

Bajo lo que sabemos, $E(SS_B) = p\sigma^2 + \Sigma_k n_k \beta_k^2$, así que se resulta la tabla de análisis de varianza:

Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)	*Esperanza de cuadrados promedios*	F	$p - value$
Mean	1	$n\bar{\boldsymbol{y}}^2$				
Between	$p-1$	SS_B	$MS_B = SS_B/(p-1)$	$\sigma^2 + \sum_k n_k\beta_k^2/(p-1)$	MS_B/MS_W	π
Within	$n-p$	SS_W	$MS_W = SS_W/(n-p)$	σ^2		
Total	n	SS_T				

Los resultados habitualmente son proporcionados con una modalidad i cuyo promedio se corresponde a la intercepta y los otros β_j son desvíos de los j promedios al promedio de i:

$$\boldsymbol{\eta} = \bar{\boldsymbol{y}}_1 + (\bar{\boldsymbol{y}}_2 - \bar{\boldsymbol{y}}_1)\boldsymbol{x}_1 + \dots + (\bar{\boldsymbol{y}}_p - \bar{\boldsymbol{y}}_1)\boldsymbol{x}_p + \boldsymbol{\varepsilon}.$$

Claro que todas las estadísticas que se resultan por el modelo lineal, se aplican a este caso.

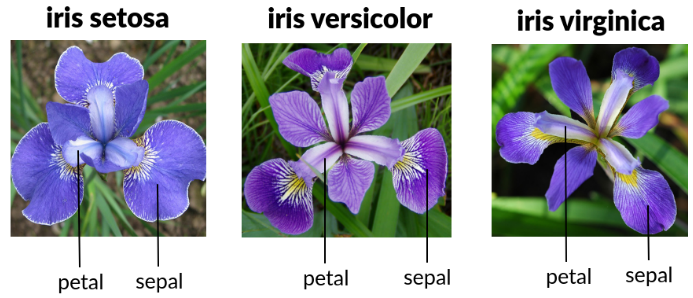
Hay que acuerdar que, para un correcto empleo del *ANOVA*, hay que cumplir con los presupuestos del modelo lineal, oportunamente traducidos:

$$\begin{cases} y_{ij} = \bar{y}_j + \varepsilon_{ij} \\ E(y_{ij}|\boldsymbol{x}_j) = \bar{y}_j \text{ (correcto)} \\ V(y_{ij}|\boldsymbol{x}_j) = \sigma^2 \text{ (homoscedasticidad)} \\ y_{ij} \text{ y } y_{hk} \text{ independientes } \forall i, j, h, k \\ y_{ij} \sim N(\bar{y}_j, \sigma^2) \text{ (normalidad)} \end{cases} \quad \begin{cases} y_{ij} = \bar{y}_j + \varepsilon_{ij} \\ E(\varepsilon_{ij}|\boldsymbol{x}_j) = 0 \\ V(\varepsilon_{ij}|\boldsymbol{x}_j) = \sigma^2 \\ \varepsilon_{ij} \text{ y } \varepsilon_{hk} \text{ independientes } \forall i, j, h \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

Esto significa que siempre se necesitan test para homoscedasticidad y normalidad *antes* de efectuar un *ANOVA*.

Los datos de Iris

- El conjunto de datos del iris de Fisher es un conjunto de datos multivariante introducido por Sir Ronald Fisher (1936, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7(II): 179–188) como ejemplo de análisis discriminante.
- Edgar Anderson (1936, The species problem in Iris, *Annals of the Missouri Botanical Garden*, 23(3):457–509) recogió los datos para cuantificar la variación morfológica de flores del iris de tres especies relacionadas.
- Por su história, véase Unwin y Kleinman (2021), The Iris Data Set: In Search of the Source of Virginica, *Significance*, 18(6):26–29.



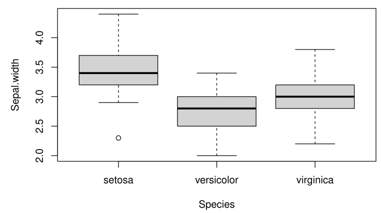
Los datos se encuentran como `iris` en *R* mismo: se trata de 50 muestras de cada una de las tres especies de Iris (*Iris setosa, Iris versicolor* y *Iris virginica*). Se midieron cuatro características de cada muestra: la longitud y la anchura de los sépalos y pétalos, en centímetros.

Aquí calculamos las estadísticas para cada modo y total:

```
v                   <- iris$Sepal.width                     # ancho sepalos
S                   <- iris$Species                         # Species
# se construye la tabla de estadísticas por cada modo
s_s                 <- tapply(v,S,length)                   # frecuencias
s_m                 <- tapply(v,S,mean)                     # promedios
s_v                 <- tapply(v,S,var)                      # varianzas
s_sd                <- tapply(v,S,sd)                       # desv.estándar
s_cv                <- s_sd/s_m                             # coef.variación
s_t                 <- cbind(length(v),mean(v),var(v),sd(v),sd(v)/mean(v))
v_brk               <- cbind(rbind(s_s,s_m,s_v,s_sd,s_cv),t(s_t))
dimnames(v_brk)     <- list(statistics=c("count","mean","variance",
                            "st.dev.","variation coef."),
                             Species=c(names(table(S)),"Total"))
cat("\n Breakdown statistics of Sepal.length by Species \n")
v_brk
boxplot(v~S)
```

donde se resulta la siguiente tabla:


```
 Breakdown statistics of Sepal.length by Species
Species
statistics          setosa  versicolor  virginica       Total
count           50.0000000 50.00000000 50.0000000 150.0000000
mean             3.4280000  2.77000000  2.9740000   3.0573333
variance         0.1436898  0.09846939  0.1040041   0.1899794
st.dev.          0.3790644  0.31379832  0.3224966   0.4358663
variation coef.  0.1105789  0.11328459  0.1084387   0.1425642
```



Con estos comandos se estima todo:

```
leveneTest(v~Species)                            # test de homocedasticidad
shapiro.test(v); tapply(v,Species,shapiro.test)  # test de normalidad
# estimación desvíos y anova
lm1                <- lm(Sepal.width~Species, data<-Iris)
anova(lm1)
l1                 <- summary(lm1)               # guarda el summary
c1                 <- confint(lm1)               # intervalos de confianza
l1$coefficients    <- cbind(l1$coefficients[,1], # incluidos en el summary
                            c1,l1$coefficients[,2:4])
l1
```

y se resultan las estimaciones del promedio de un modo y de los desvíos de los demás a esto.


```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group   2  0.5902 0.5555

Shapiro-Wilk normality test data:  Sepal.width
W = 0.98492, p-value = 0.1012

Analysis of Variance Table  Response: Sepal.width
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 11.345  5.6725   49.16 < 2.2e-16  ***
Residuals 147 16.962  0.1154

             Coefficients:    2.5 %   97.5 %  Std. Error t value Pr(>|t|)
(Intercept)        3.42800  3.33306  3.52294     0.04804  71.359  < 2e-16  ***
Speciesversicolor -0.65800 -0.79226 -0.52374     0.06794  -9.685  < 2e-16  ***
Speciesvirginica  -0.45400 -0.58826 -0.31974     0.06794  -6.683 4.54e-10  ***

Residual standard error: 0.3397 on 147 degrees of freedom
Multiple R-squared:  0.4008, Adjusted R-squared:  0.3926
F-statistic: 49.16 on 2 and 147 DF,  p-value: < 2.2e-16
```


Nóte-se que tirando el alfa del modelo:

```
lm2                <- lm(Sepal.width~Species-1, data<-Iris)
```

se resultan estimaciones del promedio de cada modo:

```
Analysis of Variance Table Response: Sepal.width
           Df  Sum Sq Mean Sq F value    Pr(>F)
Species     3 1413.44  471.15  4083.2 < 2.2e-16
Residuals 147   16.96    0.12

             Coefficients:   2.5%   97.5% Std.Error t value Pr(>|t|)
Speciessetosa     3.42800 3.33306 3.52294   0.04804   71.36   <2e-16 ***
Speciesversicolor 2.77000 2.67506 2.86494   0.04804   57.66   <2e-16 ***
Speciesvirginica  2.97400 2.87906 3.06894   0.04804   61.91   <2e-16 ***

Residual standard error: 0.3397 on 147 degrees of freedom
Multiple R-squared:  0.9881, Adjusted R-squared:  0.9879
F-statistic:  4083 on 3 and 147 DF,  p-value: < 2.2e-16
```

con un grado de libertad más.

Análisis de la varianza de dos vías

Supongamos de tener una variable respuesta cuantitativa $\boldsymbol{y}$ que piensamos depender de dos variables explicativas cualitativas, $\boldsymbol{x}_1, \boldsymbol{x}_2$. Se puede proceder como en el análisis con solo una variable, con el modelo:

$$\boldsymbol{y} = \alpha + \beta_i \boldsymbol{x}_{1i} + \gamma_j \boldsymbol{x}_{2j} + \boldsymbol{\varepsilon}$$

si $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ son ortogonales, se resulta la tabla $ANOVA$:

Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)	*Esperanza de cuadrados promedios*	F	$p-value$
Mean	1	$n\bar{\boldsymbol{y}}^2$				
$\boldsymbol{x}_1$	p_1-1	SS_{x_1}	$MS_{x_1} = SS_{x_1}/(p_1-1)$	$\sigma^2+\sum_{j1} n_{j1}\beta_{j1}^2/(p_1-1)$	MS_{x_1}/MS_W	π
$\boldsymbol{x}_2$	p_2-1	SS_{x_2}	$MS_{x_2} = SS_{x_2}/(p_2-1)$	$\sigma^2+\sum_{j2} n_{j2}\beta_{j2}^2/(p_2-1)$	MS_{x_1}/MS_W	π
Within	$n-(p_1+p_2-1)$	SS_W	$MS_W = SS_W/n-(p_1+p_2-1)$	σ^2		
Total	n	SS_T				

donde ambos $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ son testados contra los residuos.

Aquí se tienen observaciones de producción agrícola en dependencia de tres fertilizantes y dos densidades de cultivo. Se resultan los resultados siguientes:

```
Analysis of Variance Table Response: Yield
Df  Sum Sq Mean Sq F value    Pr(>F)
Density     1  5.1217  5.1217 15.3162 0.0001741 ***
Fertilizer  2  6.0680  3.0340  9.0731 0.0002533 ***
Residuals  92 30.7645  0.3344

Coefficients:    2.5 %   97.5 % Std. Error  t value Pr(>|t|)
(Intercept) 176.5261 176.2916 176.7605     0.1180 1495.490  < 2e-16 ***
Density2      0.4620   0.2275   0.6964     0.1180    3.914 0.000174 ***
Fertilizer2   0.1762  -0.1110   0.4633     0.1446    1.219 0.226115
Fertilizer3   0.5991   0.3120   0.8862     0.1446    4.144 7.57e-05 ***

Residual standard error: 0.5783 on 92 degrees of freedom
Multiple R-squared:  0.2667, Adjusted R-squared:  0.2428
F-statistic: 11.15 on 3 and 92 DF,  p-value: 2.601e-06
```


Con dos vías puede ser interesante también considerar la interacción. Efectivamente, en este caso, no solo tenemos que estimar los efectos de p_1 niveles de $\boldsymbol{x}_1$ y de p_2 niveles de $\boldsymbol{x}_2$, pero también los efectos de las $p_1 \times p_2$ casillas que se consiguen cruzando los niveles de las dos variables. Por lo tanto, el modelo será:

$$\boldsymbol{y} = \alpha + \beta_i \boldsymbol{x}_{1i} + \gamma_j \boldsymbol{x}_{2j} + \delta_{ij} \boldsymbol{x}_{1i} \boldsymbol{x}_{2j} + \boldsymbol{\varepsilon}$$

Para intender esto, pensamos a la siguiente tabla de análisis de la varianza:

Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)	*Esperanza de cuadrados promedios*	F	$p-value$
Mean	1	$n\bar{\boldsymbol{y}}^2$				
$\boldsymbol{x}_1$	$p_1 - 1$	SS_{x_1}	$MS_{x_1} = SS_{x_1}/(p_1-1)$	$\sigma^2 + \sum_{j1} n_{j1}\beta_{j1}^2/(p_1-1)$	MS_{x_1}/MS_W	π
$\boldsymbol{x}_2$	$p_2 - 1$	SS_{x_2}	$MS_{x_2} = SS_{x_2}/(p_2-1)$	$\sigma^2 + \sum_{j2} n_{j2}\beta_{j2}^2/(p_2-1)$	MS_{x_1}/MS_W	π
Interaction	$(p_1-1)(p_2-1)$	SS_{Int}	$MS_{Int} = SS_{Int}/(p_1-1)(p_2-1)$	$\sigma^2 + \sum_j n_j\beta_j^2/(p_1-1)(p_2-1)$	MS_{Int}/MS_W	π
Within	$n - p_1p_2 - 1$	SS_W	$MS_W = SS_W/(n-p_1p_2-1)$	σ^2		
Total	n	SS_T				

El interese de esta tabla es que las tres fuentes de variación, o sea $\boldsymbol{x}_1, \boldsymbol{x}_2$ y la interacción generan espacios ortogonales, así que todos se pueden comparar independientemente con los residuos (within) usando testes F.

Normalmente se empieza con testar a la interacción, ya que si hay, hay que considerar también los efectos simples, aún unos de estos podrían no ser significativos.

Introduciendo la interacción en la producción agrícola se encuentran los resultados siguientes:


```
Analysis of Variance Table  Response: Yield
                   Df  Sum Sq Mean Sq F value    Pr(>F)
Density             1  5.1217  5.1217 15.1945 0.0001864 ***
Fertilizer          2  6.0680  3.0340  9.0011 0.0002732 ***
Density:Fertilizer  2  0.4278  0.2139  0.6346 0.5325001
Residuals          90 30.3367  0.3371

Coefficients:   2.5 %    97.5 % Std. Error  t value Pr(>|t|)
(Intercept) 176.43960 176.15124 176.72795    0.14515 1215.607  < 2e-16 ***
Density2      0.63489   0.22710   1.04269    0.20527    3.093  0.00264 **
Fertilizer2   0.33869  -0.06911   0.74649    0.20527    1.650  0.10243
Fertilizer3   0.69601   0.28821   1.10381    0.20527    3.391  0.00104 **
Den2:Fert2   -0.32504  -0.90176   0.25167    0.29029   -1.120  0.26581
Den2:Fert3   -0.19377  -0.77048   0.38294    0.29029   -0.668  0.50616

Residual standard error: 0.5806 on 90 degrees of freedom
Multiple R-squared:  0.2769, Adjusted R-squared:  0.2367
F-statistic: 6.893 on 5 and 90 DF,  p-value: 1.728e-05
```

Se resulta que la interacción en este caso no es significativa.