EL MODELO LINEAL

Clase 8

Pruebas y ANCOVA

Sergio Camiz

LIMA - Marzo-Mayo 2025

Modelo lineal y ANOVA

Teoricamente, la tabla *ANOVA* de dos vías se escribe así:

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				

Si $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ son ortogonales, las estimaciones y las sumas de cuadrados son independientes, así como la interacción.

Vemos como le encuentramos en las salidas del modelo lineal y en ellas de las varias *ANOVA*.

Asuntos de la clase 8

- Modelo lineal y *ANOVA*
- Test post-hoc
- Análisis de la covarianza
- Ejemplos

Hay tre tipos de *ANOVA*: *I, II* y *III*, dependiendo de como se calculan las sumas de cuadrados.

- Las sumas de cuadrados del *ANOVA* de tipo I corresponden a la suma de cuadrados de cada fuente después de las prévias: $SS(A)$, $SS(B|A)$, $SS(AB|A,B)$. Entonces si se cambia el orden estas cambian también: solo no cambian el residuo y el total.
- Las sumas de cuadrados del *ANOVA* de tipo II corresponden a la suma de cuadrados de cada fuente tirando las interacciones: $SS(A|AB)$, $SS(B|AB)$. Entonces cambiando el orden, no cambian.
- Las sumas de cuadrados del *ANOVA* de tipo III corresponden a la suma de cuadrados de cada fuente tirando todas las demás y las interacciones: $SS(A|B,AB), SS(B|A,AB)$.

En la salida del modelo lineal se encuentra el *Residual Standard Error* que es el desvío estándar de los residuos (raíz cuadrada de MSe).

Igualmente, el F que resulta corresponde en el anova (anova1) a la razón entre el promedio de los tres cuadrados promedios de los tres regresores (intercepta, *v1, v2*) (o sea $MSr = SSr/df$) y MSe.

Las probabilidades a lado de los beta están asociadas al t de student correspondiente y se refieren a la variación correspondiente a β_i con otros regresores fijos. Son las mismas a lado del F en el anova3 (Anova con type=3).

Test post hoc

Cuando se hace un test estadístico, se toma un riesgo de error (error de tipo I), que normalmente se asume ser 5 %. Esto significa que cada 20 test para rechazar una hipótesis nula, hay una buena probabilidad de rechazarla aún fuera verdadera.

All Pairwise Comparions Alpha = 0.05		
Groups	Comparisons	Experimentwise Error Rate
2	1	0.05
3	3	0.142625
4	6	0.264908109
5	10	0.401263061
6	15	0.53670877
7	21	0.659438374
8	28	0.762173115
9	36	0.842220785
10	45	0.900559743
11	55	0.940461445
12	66	0.966134464
13	78	0.981700416
14	91	0.990606054
15	105	0.995418807

Si los regresores están ortogonales, no hay variación en la suma de cuadrados en el *ANOVA I* debido a un intercambio. Igualmente por esta razón las sumas de cuadrados en *ANOVA II* son las mismas.

En `anova1` y `anova2` el total corresponde a la suma de cuadrados centrados $SStc$ del $\boldsymbol{y}$. En `anova3` se debería corresponder a la suma de cuadrados, ya que hay una suma de cuadrados del promedio arriba. Pero se resulta una diferencia.

Si los regresores no están ortogonales, si hay diferencias: se pueden organizar los regresores un orden hecho para ver el ingreso de cadauno después de los demás en el *ANOVA I*.

Si no hay interacción el *ANOVA* de tipo *II* resulta el más indicado.

Si hay interacción el *ANOVA* de tipo *III* se necesita.

El ventaje del *ANOVA* es que con solo un test F se prueba la hipótesis que no haya diferencia entre promedios. Pero para la identificación de cuales promedios son diferentes de los demás, se necesitan otros test, como un t de student entre cada par de promedios que si aumentan el riesgo de error de tipo I.

Los test post-hoc funcionan de manera diferente: se define el umbral π para el conjunto de test que se quieren y el método incluye este umbral ajustando los test individuales para que el conjunto no exceda el umbral fijado.

El test LSD de Fisher

Propuesto en 1935, va definir la mínima diferencia significativa entre promedios como

$$LSD = t_{n-k,\pi}\sqrt{MSe\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Como el test no corrige errores de tipo I, se prefiere corregir como Fisher-Hayter:

$$LSD = q_{k-1,\pi}\sqrt{\frac{MSe}{n}}$$

con q la distribución de *rango studentizado*.

En el paquete `agricolae` de R se encuentra la función `LSD.test`.

El test Student-Neuman-Kuels

El método se ha introducido para Newman in 1939 y luego desarrollado para Keuls in 1952.

Este test se puede aplicar también a grupos no iguales, ya que la estadística q se calcula como

$$q = \frac{\bar{x}_{max} - \bar{x}_{min}}{\frac{S}{2}(\frac{1}{n_{min}}+\frac{1}{n_{max}})}$$

El método es más poderoso del *HSD*: empeza buscando si hay diferencias entre promedios extremos del número total de grupos, luego va reduciendo este número progresivamente.

En el paquete `agricolae` de R se encuentra la función `SNK.test`.

La distribución de rango studentizado

q la distribución de *rango studentizado* depende de:

1. el umbral π de error de tipo I eligido ($5\,\%, 1\,\%$),
2. el número k de promedios,
3. los grados de libertad $n-k$ con n número de observaciones.

El método produce intervalos de confianza a nivel $1-\pi$ por cada comparación entre parejas.

En R se encuentran las funciones

- `ptukey`, que dado un valor de q devuelve la probabilidad asociada,
- `qtukey`, que dada una probabilidad, devuelve el valor de q correspondiente.

HSD de Tukey

El test *HSD, Honestly Significant Difference* de Tukey (1949) solo se aplica al caso de grupos de igual tamaño. Se calcula

$$q = \frac{\bar{x}_{max} - \bar{x}_{min}}{MSE\sqrt{2/n}}.$$

En R hay la función `TukeyHSD`, que se aplica depués de correr `lm` seguida de `aov`, y en el paquete `agricolae` de R se encuentra la función `HSD.test`.

El método de Scheffé (1959)

Se basa sobre contrastes, o sea combinaciones lineales de promedios $\hat{C} = c_1\bar{x}_1 + \ldots + c_k\bar{x}_k$ con $\Sigma\, c_i = 0$.

Como la varianza de un contraste resulta $S_{\hat{C}}{}^{\prime}2 = MSe\, \Sigma_i \frac{c_i}{n_i}$ con n_i el tamaño de cada grupo, se construyen intervales de confianza.

En el paquete `agricolae` de *R* se encuentra la función `scheffe.test`.

Es fácil de intender esto, ya que los $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ son variables indicadoras, que valen cero para todos los niveles, sino uno. 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	$\bar{\boldsymbol{y}}^2$				
$\boldsymbol{x}_1$	2	SS_r	$MSr = SS_r/2$	$\sigma^2 + \boldsymbol{\eta}'\boldsymbol{\eta}/2$	MS_r/MS_e	π
$\boldsymbol{x}_2$	$p-1$	SS_{x_2}	$MS_{x_2} = SS_{x_2}/(p-1)$	$\sigma^2 + \sum_j n_j\beta_j^2/(p-1)$	MS_{x_2}/MS_W	π
Interaction	$p-1$	SS_{Int}	$MS_{Int} = SS_{Int}/(p-1)$	$\sigma^2 + \sum_j \boldsymbol{\eta}_j'\boldsymbol{\eta}_j/(p-1)$	MS_{Int}/MS_W	π
Within	$n-2p-1$	SS_W	$MS_W = SS_W/(n-2p-1)$	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 que se pueden ortogonalizar, así que todos se pueden comparar independientemente con los residuos (within) usando testes F.

Análisis de la covarianza

Supongamos de querer estimar una variable respuesta cuantitativa a través de una cualitativa, y una cuantitativa. Claro que la cualitativa se transforma en una matríz disjunctiva completa, mientras la cuantitativa sigue estar como siempre.

Podemos imaginar dos modelos posibles

1. $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x}_1 + \gamma_2\boldsymbol{x}_2 + \cdots + \gamma_p\boldsymbol{x}_p + \varepsilon$
   las $(\boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ siendo variables indicadoras de las modalidades. En este caso, los γ representan desvíos al promedio del nivel eliminado.
2. $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x}_1 + \gamma_2\boldsymbol{x}_2 + \cdots + \gamma_p\boldsymbol{x}_p + \delta_2\boldsymbol{x}_1\boldsymbol{x}_2 + \cdots + \delta_p\boldsymbol{x}_1\boldsymbol{x}_p + \varepsilon$
   de esta manera incluyendo la interacción, los δ representando el desvío al pendiente.

Normalmente se empeza con testar a la interacción, ya que si hay, hay que considerar también los efectos simples, aún unos de estos podrían no ser significativos. Al contrario, si no hay interacción, la fila se puede borrar, corrigendo oportunamente la tabla.