Bloques aleatorizados

José Gabriel Palomo Sánchez
gabriel.palomo@upm.es

E.U.A.T.
U.P.M.

Julio de 2011

Índice

1. Introducción
   1. Definición de variable bloque
   2. Organización de los datos
2. Inferencia en el modelo de bloques aleatorizados
   1. Hipótesis del modelo
   2. Consecuencias de las hipótesis del modelo
   3. Estimación de los parámetros del modelo
   4. El Análisis de la varianza en el modelo de bloques aleatorizados
   5. El test de la F
   6. La tabla ADEVA
   7. Diagnosis y validación del modelo
   8. Inferencia sobre los parámetros del modelo

Bloques aleatorizados I

Definición

Una variable bloque es una variable cuyo efecto sobre la respuesta no interesa específicamente al investigador, pero cuya inclusión en el modelo puede disminuir la variabilidad experimental y, en consecuencia, facilitar el análisis del efecto de los factores de interés sobre la variable dependiente.

Ejemplo I

Supóngase que se desea analizar la diferencia de la bondad de cuatro tipos de semillas de trigo distintas. Para ello se eligen cuatro parcelas de igual área y, de forma aleatoria, en cada una de ellas se planta un tipo de semilla. Posteriormente se comparan los rendimentos en cada una de las parcelas, (variable respuesta).

Ejemplo I

Si pueden existir diferentes fertilidades en las distintas parcelas, y se encuentran en el análisis diferencias entre los rendimientos de las distintas semillas, ¿cómo se sabe que estas diferencias no están sobrevaloradas a causa de las diferentes fertilidades de las parcelas?

Ejemplo I

Una alternativa consiste en incluir, como factor (bloque), el tipo de parcela en el análisis de los resultados. Con ello se consigue:

1. Distinguir el efecto sobre la respuesta de la fertilidad de la parcela del de la bondad de la semilla.
2. Eliminar la variabilidad producida por las distintas parcelas de la variabilidad experimental, por lo que se consigue mayor sensibilidad para detectar posibles diferencias entre las semillas.

Ejemplo I

- De esta manera, cada parcela se dividiría, por ejemplo, en cuatro subparcelas y, de forma aleatoria, se adjudicaría cada una de estas subparcelas a un tipo de semilla. Los datos del experimento se podrían resumir en una tabla como la siguiente:

		SEMILLA			
		1	2	3	4
PARCELA	1	y_{11}	y_{21}	y_{31}	y_{41}
	2	y_{12}	y_{22}	y_{32}	y_{42}
	3	y_{13}	y_{23}	y_{33}	y_{43}
	4	y_{14}	y_{24}	y_{34}	y_{44}

Bloques aleatorizados II

- En general, si existe un factor de interés con K niveles y un factor bloque con J valores distintos, el conjunto de datos de la experimentación se resume en una tabla del tipo:

		FACTOR			
		1	2	$\cdots$	K
BLOQUE	1	y_{11}	y_{21}	$\cdots$	y_{K1}
	2	y_{12}	y_{22}	$\cdots$	y_{K2}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
	J	y_{1J}	y_{2J}	$\cdots$	y_{KJ}

- El valor y_{ij} representa la observación realizada de la variable respuesta, en el nivel i del factor, y en el valor j del bloque.

Bloques aleatorizados III

Observación

Debe observarse que, a diferencia de lo expuesto en el modelo del análisis de la varianza con un factor, en el modelo de bloques aleatorizados en cada condición experimental, ij, definida por el nivel i del factor y por el nivel j del bloque, se obtendrá un único dato experimental.

Hipótesis del modelo I

La significatividad del factor se analiza bajo las siguientes hipótesis:

- La variable respuesta se puede descomponer según el modelo:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

donde:

  - μ representa la media general.
  - α_i es la desviación de la media general debida a que la observación se realiza en el nivel i del factor.
  - β_j es la desviación de la media debida a que la observación se realiza en el valor j del bloque.
  - e_{ij} representa el error aleatorio.

Hipótesis del modelo II

Además se supondrá que:

- Para todos los valores de i y j, se tiene que:

$$e_{ij} \approx N(0, \sigma).$$

  El hecho de que el valor de σ no dependa de los valores de i ni de j se conoce con el nombre de **homocedasticidad**

- Todos los e_{ij} son independientes entre sí.

- Por último, para evitar problemas de indeterminación en la estimación del modelo se impondrá la siguiente condición:

$$\sum_{i=1}^{K} \alpha_i = \sum_{j=1}^{J} \beta_j = 0.$$

Consecuencias de las hipótesis del modelo I

Como consecuencia de las hipótesis se cumple que:

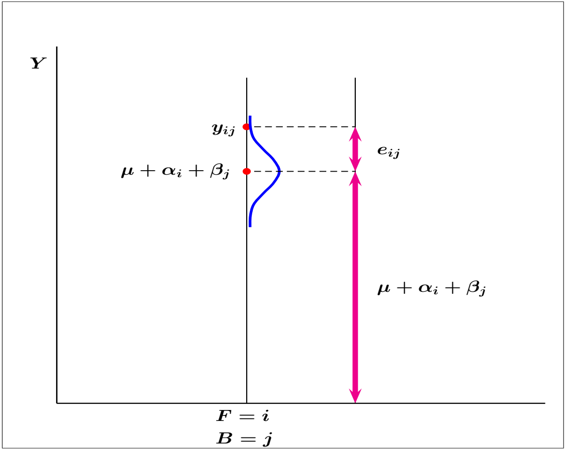
- La variable respuesta en los individuos sometidos al nivel i del factor, y al valor j del bloque, sigue una distribución:

$$(Y|F = i \wedge B = j) \approx N(\mu + \alpha_i + \beta_j, \sigma).$$

- Todos los y_{ij} son independientes entre sí.

Consecuencias de las hipótesis del modelo II

Gráficamente,



Observaciones

1. Debe notarse que la diferencia entre dos observaciones obtenidas con distinto nivel del factor, pero dentro del mismo bloque, no depende de la contribución del bloque. En efecto:

$$y_{ij}-y_{lj} = (\mu+\alpha_i+\beta_j+e_{ij})-(\mu+\alpha_l+\beta_j+e_{lj}) = (\alpha_i-\alpha_l)+(e_{ij}-e_{lj}).$$

2. El objetivo fundamental del estudio es analizar si existen diferencias significativas entre los valores de α_i.

Estimación de los parámetros del modelo I

- Para estimar la media global μ, se utiliza la media de todas las observaciones:

$$\bar{y}_{\bullet\bullet} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{J} y_{ij}}{n}$$

- El valor de α_i, desviación de la media global atribuible a que la observación se realiza en el nivel i del factor, se estima por:

$$\hat{\alpha}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet},$$

donde:

$$\bar{y}_{i\bullet} = \frac{\sum_{j=1}^{J} y_{ij}}{J}$$

Estimación de los parámetros del modelo II. Observación

Debe observarse que $\hat{\alpha}_i$ se calcula como la diferencia entre la media de todas las observaciones realizadas en el nivel i del factor y la media global, lo que es coherente con el significado de α_i en el modelo.

Estimación de los parámetros del modelo III

- El valor de β_j, desviación de la media global atribuible a que la observación se realiza en el valor j del bloque, se estima por:

$$\hat{\beta}_j = \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet},$$

donde:

$$\bar{y}_{\bullet j} = \frac{\sum_{i=1}^{K} y_{ij}}{K}$$

- La varianza del error experimental, σ^2, se estima por la varianza residual:

$$\hat{s}_R^2 = \frac{\sum\sum e_{ij}^2}{(K-1)(J-1)}.$$

Estimación de los parámetros del modelo IV. Observación

De manera similar a como se hizo en el caso de $\hat{\alpha}_i$, debe observarse que $\hat{\beta}_j$ se calcula como la diferencia entre la media de todas las observaciones realizadas en el nivel j del bloque y la media global, lo que es coherente con el significado de β_j en el modelo.

Estimación de los parámetros del modelo V

Observaciones:

- Solamente es necesario estimar $K-1$ valores de $\hat{\alpha}_i$, ya que:

$$\sum_{i=1}^{K} \hat{\alpha}_i = 0.$$

- Por la misma razón, sólo hay que estimar $J-1$ de los $\hat{\beta}_j$, pues:

$$\sum_{j=1}^{J} \hat{\beta}_j = 0.$$

- El residuo e_{ij} representa la diferencia entre el valor observado y el previsto por el modelo:

$$e_{ij} = y_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}$$

Estimación de los parámetros del modelo VI

La siguiente tabla resume la estimación de los efectos del factor:

		FACTOR			
		1	2	$\cdots$	K
BLOQUE	1	y_{11}	y_{21}	$\cdots$	y_{K1}
	2	y_{12}	y_{22}	$\cdots$	y_{K2}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	J	y_{1J}	y_{2J}	$\cdots$	y_{KJ}
	$\hat{\alpha}$	$\hat{\alpha}_1 = (\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})$	$\hat{\alpha}_2 = (\bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet})$	$\cdots$	$\hat{\alpha}_K = (\bar{y}_{K\bullet} - \bar{y}_{\bullet\bullet})$

Estimación de los parámetros del modelo VII

La siguiente tabla resume la estimación de los efectos del bloque:

FACTOR

		1	2	$\cdots$	K	$\hat{\beta}$
BLOQUE	1	y_{11}	y_{21}	$\cdots$	y_{K1}	$\hat{\beta}_1 = (\bar{y}_{\bullet 1} - \bar{y}_{\bullet\bullet})$
	2	y_{12}	y_{22}	$\cdots$	y_{K2}	$\hat{\beta}_2 = (\bar{y}_{\bullet 2} - \bar{y}_{\bullet\bullet})$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	J	y_{1J}	y_{2J}	$\cdots$	y_{KJ}	$\hat{\beta}_J = (\bar{y}_{\bullet J} - \bar{y}_{\bullet\bullet})$

Análisis de la varianza I

Como en el caso del Análisis de la varianza con un factor, el análisis de la existencia de diferencias en la variable respuesta debidas al nivel del factor, o al valor del bloque, se puede realizar comparando la variabilidad explicada por cada uno de los factores con la variabilidad total.

Así:

$$(y_{ij} - \bar{y}_{\bullet\bullet}) = \underbrace{(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})}_{\hat{\alpha}_i} + \underbrace{(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})}_{\hat{\beta}_j} + \underbrace{(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})}_{e_{ij}}.$$

Análisis de la varianza II

Llamando variabilidad total y variabilidad explicada por el factor, respectivamente, a los términos:

$$VT = \sum_{i=1}^{K} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{\bullet\bullet})^2$$

y

$$VE(\alpha) = \sum_{i=1}^{K} \sum_{j=1}^{J} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = J \sum_{i=1}^{K} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = J \sum_{i=1}^{K} \hat{\alpha}_i^2.$$

Análisis de la varianza III

Y, de forma análoga, llamando variabilidad explicada por el bloque y variabilidad no explicada, respectivamente, a los términos:

$$VE(\beta) = \sum_{i=1}^{K}\sum_{j=1}^{J}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2 = K\sum_{j=1}^{J}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2 = K\sum_{j=1}^{J}\hat{\beta}_j^2,$$

y

$$VNE = \sum_{i=1}^{K}\sum_{j=1}^{J} e_{ij}^2.$$

Análisis de la varianza IV

Se puede demostrar que:

$$VT = VE(\alpha) + VE(\beta) + VNE.$$

Por otro lado,

- El término $VE(\alpha)$ depende solo de la variabilidad entre los distintos valores de $\hat{\alpha}_i$.
- El término $VE(\beta)$ depende solo de la variabilida entre los distintos de $\hat{\beta}_j$.
- El término VNE es una medida de la variabilidad de los residuos.

ANÁLISIS DE LA VARIANZA V

Observaciones:

- En la medida en que $VE(\alpha)$ sea grande en relación con VNE, habrá evidencia de la existencia de diferencias significativas en la respuesta producidas por los distintos niveles del factor.
- En la medida en que $VE(\beta)$ sea grande en relación con VNE, habrá evidencia de la existencia de diferencias significativas en la respuesta producidas por los distintos niveles del bloque.

Análisis de la varianza VI

- Para discutir la magnitud de $VE(\alpha)$ y $VE(\beta)$, respectivamente, es necesario analizar sus distribuciones de probabilidad.

Teorema I

1. Si se verifica la hipótesis:

$$\alpha_1 = \alpha_2 = \cdots = \alpha_K = 0,$$

la variable $VE(\alpha)/\sigma^2$ se distribuye como una χ^2_{K-1}.

2. Si se verifica la hipótesis:

$$\beta_1 = \beta_2 = \cdots = \beta_J = 0,$$

la variable $VE(\beta)/\sigma^2$ se distribuye como una χ^2_{J-1}.

Análisis de la varianza VII

Teorema II

La variable VNE/σ^2 se distribuye, en cualquier caso, como una $\chi^2_{(K-1)(J-1)}$ y es independiente con las anteriores.

Análisis de la varianza VIII

Consecuencia I

Si se verifica la hipótesis:

$$\alpha_1 = \alpha_2 = \cdots = \alpha_K = 0$$

la variable

$$\frac{\frac{VE(\alpha)}{\sigma^2(K-1)}}{\frac{VNE}{\sigma^2(K-1)(J-1)}} \longrightarrow F_{(K-1;(K-1)(J-1))}$$

Análisis de la varianza IX

Consecuencia II

Si se verifica las hipótesis:

$$\beta_1 = \beta_2 = \cdots = \beta_J = 0$$

la variable

$$\frac{\frac{VE(\beta)}{\sigma^2(J-1)}}{\frac{VNE}{\sigma^2(K-1)(J-1)}} \longrightarrow F_{(J-1;(K-1)(J-1))}$$

Análisis de la varianza X

- Llamando $\hat{s}_e^2(\alpha)$ al valor de $VE(\alpha)$/(K-1),
- $\hat{s}_e^2(\beta)$ al valor de $VE(\beta)$/(J-1),
- y $\hat{s}_R^2$ al valor de VNE/((K-1)(J-1)).

Se tiene que cuando $\alpha_1 = \alpha_2 = \cdots = \alpha_K = 0$:

$$\frac{\hat{s}_e^2(\alpha)}{\hat{s}_R^2} \longrightarrow F_{(K-1);((K-1)(J-1))}.$$

Y cuando $\beta_1 = \beta_2 = \cdots = \beta_J = 0$:

$$\frac{\hat{s}_e^2(\beta)}{\hat{s}_R^2} \longrightarrow F_{(J-1);((K-1)(J-1))}.$$

Análisis de la varianza XI. El test de la F I

- Empleando los resultados anteriores, para discutir el contraste:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0,$$

frente a

$$H_1 : \text{Existe al menos un } \alpha_i \text{ tal que } \alpha_i \neq 0,$$

basta con analizar el valor del estadístico

$$F = \frac{\hat{s}_e^2(\alpha)}{\hat{s}_R^2}$$

Análisis de la varianza XII. El test de la F II

- De manera que, si se denomina F_α al valor tal que

$$P(F_{(K-1;(K-1)(J-1))} > F_\alpha) = \alpha,$$

cuando

$$F = \frac{\hat{s}_e^2(\alpha)}{\hat{s}_R^2} < F_\alpha$$

se aceptará la hipótesis nula, ($\alpha_i = 0$, para todo i), que se rechazará en caso contrario.

Observación: Nótese que el test de la F es un contraste unilateral, en coherencia con la hipótesis que se contrasta.

Análisis de la varianza XIII. El test de la F III

- Análogamente, para discutir el contraste:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0,$$

frente a

$$H_1 : \text{Existe al menos un } \beta_j \text{ tal que } \beta_j \neq 0,$$

basta con analizar el valor del estadístico

$$F = \frac{\hat{s}_e^2(\beta)}{\hat{s}_R^2}$$

Análisis de la varianza XIV. El test de la F IV

- De manera que, si se denomina F_α al valor tal que

$$P(F_{(J-1;(K-1)(J-1))} > F_\alpha) = \alpha,$$

cuando

$$F = \frac{\hat{s}_e^2(\beta)}{\hat{s}_R^2} < F_\alpha$$

se aceptará la hipótesis nula, ($\beta_j = 0$, para todo j), que se rechazará en caso contrario.

Observación: Nótese que el test de la F es un contraste unilateral, en coherencia con la hipótesis que se contrasta.

Análisis de la varianza XV. La tabla ADEVA

Los resultados de los tests de la F se resumen en la

Tabla ADEVA

Fuentes de variación	**Suma de cuadrados**	**Grados de libertad**	**Varianzas**	F	p-v.
Factor	$\sum J(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$K-1$	$\hat{s}_e^2(\alpha)$	$\frac{\hat{s}_e^2(\alpha)}{\hat{s}_R^2}$	$p(\alpha)$
Bloque	$\sum K(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2$	$J-1$	$\hat{s}_e^2(\beta)$	$\frac{\hat{s}_e^2(\beta)}{\hat{s}_R^2}$	$p(\beta)$
Residual	$\sum\sum(e_{ij})^2$	$(K-1)(J-1)$	$\hat{s}_R^2$		
Total	$\sum\sum(y_{ij} - \bar{y}_{\bullet\bullet})^2$	$n-1$	$\hat{s}_Y^2$		

Análisis de la varianza XVI. Diagnosis y validación del modelo

- Como en el Análisis de la varianza con un factor, una vez realizado el análisis de la varianza, antes de emplear las conclusiones allí extraídas, es necesario verificar las hipótesis del modelo.
- Esta verificación se lleva a cabo por medio del análisis de los residuos.
  - La discusión de la normalidad se realiza a través del papel probabilístico normal.
  - La comprobación de la homocedasticidad requiere gráficos de los residuos frente a los distintos valores del factor, del bloque y frente a los valores previstos.

Inferencia sobre los parámetros del modelo I

Una vez realizada la diagnosis del modelo, puede ser necesario hacer inferencia respecto de los parámetros del mismo.

- La inferencia respecto del valor de α_i se puede hacer teniendo en cuenta que:

$$\frac{\bar{y}_{i\bullet} - \alpha_i}{\hat{s}_R/\sqrt{J}} \longrightarrow t_{(K-1)(J-1)}$$

- La comparación de las desviaciones de la media general, α_i y α_j, provocadas por dos valores distintos del factor, se puede realizar si se tiene en cuenta que:

$$\frac{(\bar{y}_{i\bullet} - \bar{y}_{j\bullet}) - (\alpha_i - \alpha_j)}{\hat{s}_R\sqrt{\frac{2}{J}}} \longrightarrow t_{(K-1)(J-1)}$$

Inferencia sobre los parámetros del modelo II

Análogamente,

- La inferencia respecto del valor de β_j se puede hacer teniendo en cuenta que:

$$\frac{\bar{y}_{\bullet j} - \beta_j}{\hat{s}_R/\sqrt{K}} \longrightarrow t_{(K-1)(J-1)}$$

- La comparación de las desviaciones de la media general, β_i y β_j, provocadas por dos valores distintos del bloque, se puede realizar si se tiene en cuenta que:

$$\frac{(\bar{y}_{\bullet i} - \bar{y}_{\bullet j}) - (\beta_i - \beta_j)}{\hat{s}_R\sqrt{\frac{2}{K}}} \longrightarrow t_{(K-1)(J-1)}$$

Inferencia sobre los parámetros del modelo III

- La inferencia respecto de σ^2 se realiza teniendo en cuenta la siguiente distribución:

$$\frac{(n-K)\hat{s}_R^2}{\sigma^2} \longrightarrow \chi^2_{(K-1)(J-1)}$$