Modelo de Análisis de la Covarianza. Introducción al modelo de Medidas Repetidas

Modelo de Análisis de la Covarianza

Introducción

El diseño por bloques se considera para eliminar el efecto de los factores de *ruido* que no son controlables. El análisis de la covarianza es otro método que se utiliza para un problema semejante: supongamos un experimento con una variable respuesta, y, donde existe otra variable, x, de modo que ambas están relacionadas linealmente. Supongamos, además, que x no es una variable controlable por el experimentador pero que puede ser observada junto con y. A la variable x se le denomina *covariable* o variable concomitante.

El análisis de la covarianza sirve para ajustar la variable respuesta por el efecto de la covariable. En caso de no hacerlo, la media de cuadrados del error puede aumentar mucho y hacer que las verdaderas diferencias en la respuesta debido a los tratamientos sean difíciles de detectar. El análisis de la covarianza resulta ser una combinación entre el ANOVA y el análisis de regresión.

Modelo

Suponemos un modelo con un solo factor y una covariable y asumimos una relación lineal entre la variable respuesta y la covariable:

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij}$$

para

$i = 1, \ldots, a$
$j = 1, \ldots, n$

En el modelo,

y_{ij} es la j-ésima observación bajo el i-ésimo nivel del tratamiento.
x_{ij} es la medida de la covariable que se hace para y_{ij}
$\bar{x}_{..}$ es la media de los valores de x_{ij}
μ es el valor medio global.
α_i es el efecto del nivel i-ésimo del tratamiento
β coeficiente de regresión que relaciona y_{ij} con la covariable x_{ij}
ε_{ij} error aleatorio

Se asume que $\varepsilon_{ij} \sim N(0, \sigma^2)$ son independientes entre sí, $\beta \neq 0$, $\sum_{i=1}^{a} \alpha_i = 0$ y la covariable x no está afectada por los tratamientos.

Se utiliza la siguiente notación:

$$\begin{aligned}
S_{yy} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{an} \\
S_{xx} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - \frac{x_{..}^2}{an} \\
S_{xy} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..}) = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}y_{ij} - \frac{(x_{..})(y_{..})}{an}
\end{aligned}$$

$$\begin{aligned}
T_{yy} &= n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 = \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{an} \\
T_{xx} &= n\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{x}_{..})^2 = \frac{1}{n}\sum_{i=1}^{a} x_{i.}^2 - \frac{x_{..}^2}{an} \\
T_{xy} &= n\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..}) = \frac{1}{n}\sum_{i=1}^{a} x_{i.}y_{i.} - \frac{x_{..}y_{..}}{an}
\end{aligned}$$

$$\begin{aligned}
E_{yy} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = S_{yy} - T_{yy} \\
E_{xx} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})^2 = S_{xx} - T_{xx} \\
E_{xy} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}) = S_{xy} - T_{xy}
\end{aligned}$$

En general, $S = T + E$ donde los símbolos S, T y E son las sumas de cuadrados y los dobles productos para el total, los tratamientos y el error respectivamente.

Los estimadores por mínimos cuadrados son

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \\ \hat{\beta} &= \frac{E_{xy}}{E_{xx}} \end{aligned}$$

La suma de cuadrados del error es

$$SCE = E_{yy} - \frac{E_{xy}^2}{E_{xx}},$$

con $a(n-1) - 1$ grados de libertad.

La varianza del error experimental es, así,

$$MCE = \frac{SCE}{a(n-1) - 1}.$$

Supongamos que no hay efecto del tratamiento, entonces el modelo queda reducido a

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij}$$

Los estimadores por mínimos cuadrados son

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\beta} &= \frac{E_{xy}}{E_{xx}}. \end{aligned}$$

La suma de cuadrados del error en el modelo reducido queda como

$$SCE^* = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

con $(an - 2)$ grados de libertad.

La cantidad $\frac{S_{xy}^2}{S_{xx}}$ es la reducción de la suma de cuadrados de y, obtenida por la regresión lineal de y en x. Además, $SCE < SCE^*$ porque el modelo completo incluye los parámetros

adicionales Así, $SCE^* - SCE$ es una reducción en la suma de cuadrados debida a los términos α_i.

De esta manera, a partir de $SCE^* - SCE$ se tiene una suma de cuadrados con $(a-1)$ grados de libertad para contrastar la hipótesis de que no hay efectos de los tratamientos.

Para contrastar la hipótesis

$$H_0 \equiv \alpha_i = 0, \forall i$$

se calcula

$$F_0 = \frac{\frac{SCE^* - SCE}{a-1}}{\frac{SCE}{a(n-1)-1}}$$

que, si la hipótesis nula es cierta, se distribuye como una F de Snedecor:

$$F_{a-1,a(n-1)-1},$$

de modo que se rechaza H_0 al nivel α si

$$F_0 > F_{a-1,a(n-1)-1;\alpha}$$

Se obtiene la siguiente tabla:

		Sumas de productos		
F. Variación	g. l.	x	xy	y
Tratamientos	$a-1$	T_{xx}	T_{xy}	T_{yy}
Error	$a(n-1)$	E_{xx}	E_{xy}	E_{yy}
Total	$an-1$	S_{xx}	S_{xy}	S_{yy}

	Ajuste por Regresión		
F. Variación	y	g. l.	Cuadrados medios
Tratamientos	$SCE = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$		
Error	$SCE^* = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$	$a(n-1)-1$	$MCE = \frac{SCE}{a(n-1)-1}$
Total		$an-2$	
Tratamientos Ajustados	$SCE^* - SCE$	$a-1$	$\frac{SCE^*-SCE}{a-1}$

También es importante efectuar un contraste también sobre el termino de regresión β.

Se calculan las medias ajustadas por la regresión:

$$\text{Ajust\_}\bar{y}_{i\cdot} = \bar{y}_{i\cdot} - \hat{\beta}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})$$

para $i = 1, 2, \ldots, a$, donde

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}}.$$

Esta media ajustada es el estimador de mínimos cuadrados de $\mu + \alpha_i$, donde $i = 1, \ldots, a$.

Por otro lado, el error estándar de la media ajustada de cada tratamiento es

$$S_{Ajust\_\bar{y}_{i\cdot}} = \left[MCE \left(\frac{1}{n} + \frac{(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2}{E_{xx}} \right) \right]^{\frac{1}{2}}.$$

Finalmente, con respecto al coeficiente de regresión se puede contrastar la hipótesis $H_0 \equiv \beta = 0$ mediante el estadístico

$$F_0 = \frac{\frac{E_{xy}^2}{E_{xx}}}{MCE},$$

que, si la hipótesis nula es cierta, se distribuye como una F de Snedecor, $F_{1,a(n-1)-1}$.

De este modo, se rechaza $H_0 \equiv \beta = 0$, a nivel α si

$$F_0 > F_{1,a(n-1)-1,\alpha}.$$

Ejemplo. Se considera un estudio para determinar si existen diferencias en la resistencia de una fibra producida por tres máquinas diferentes. Se piensa que el grosor de las fibras (x) influye también, obteniéndose los siguientes valores

Máquina 1		**Máquina 2**		**Máquina 3**	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15
207	126	216	130	180	106

Se trata de eliminar el efecto del grosor (x) en la resistencia de las fibras para poder comparar el efecto de las tres máquinas.

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{an} = 36^2 + 41^2 + \cdots + 32^2 - \frac{603^2}{3\cdot 5} = 346{,}4 \\ S_{xx} &= \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - \frac{x_{..}^2}{an} = 20^2 + 25^2 + \cdots + 15^2 - \frac{362^2}{3\cdot 5} = 261{,}73 \\ S_{xy} &= \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}y_{ij} - \frac{(x..)(y..)}{an} = 20\cdot 36 + \cdots + 15\cdot 32 - \frac{362\cdot 603}{3\cdot 5} = 282{,}6 \end{aligned}$$

$$\begin{aligned} T_{yy} &= \frac{1}{n}\sum_{i=1}^{a} y_{i\cdot}^2 - \frac{y_{..}^2}{an} = \frac{1}{5}(207^2 + 216^2 + 180^2) - \frac{603^2}{3\cdot 5} = 140{,}4 \\ T_{xx} &= \frac{1}{n}\sum_{i=1}^{a} x_{i\cdot}^2 - \frac{x_{..}^2}{an} = \frac{1}{5}(126^2 + 130^2 + 106^2) - \frac{362^2}{3\cdot 5} = 66{,}13 \\ T_{xy} &= \frac{1}{n}\sum_{i=1}^{a} x_{i\cdot}y_{i\cdot} - \frac{x..y..}{an} = \frac{1}{5}(207\cdot 126 + 216\cdot 130 + 180\cdot 106) - \frac{362\cdot 603}{3\cdot 5} = 96 \end{aligned}$$

$$\begin{aligned} E_{yy} &= S_{yy} - T_{yy} = 206 \\ E_{xx} &= S_{xx} - T_{xx} = 195{,}6 \\ E_{xy} &= S_{xy} - T_{xy} = 186{,}6 \end{aligned}$$

Por otro lado

$$SCE^* = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 346{,}4 - \frac{282{,}6^2}{261{,}73} = 41{,}27$$

con $(an-2) = 3\cdot 5 - 2 = 13$ grados de libertad, y

$$SCE = E_{yy} - \frac{E_{xy}^2}{E_{xx}} = 206 - \frac{186{,}6^2}{195{,}6} = 27{,}99$$

con $a(n-1) - 1 = 3(5-1) - 1 = 11$ grados de libertad.

La suma de cuadrados para contrastar $H_0 \equiv \alpha_i = 0$, para $i = 1, 2, 3$ es

$$SCE^* - SCE = 41{,}27 - 27{,}99 = 13{,}28$$

con $a - 1 = 3 - 1 = 2$ grados de libertad.

Así, se calcula

$$F_0 = \frac{\frac{SCE^* - SCE}{a-1}}{\frac{SCE}{a(n-1)-1}} = \frac{\frac{13{,}28}{2}}{\frac{27{,}99}{2{,}54}} = 2{,}61.$$

Como, para $\alpha = 0{,}10$,

$$F_0 = 2{,}61 < F_{a-1,a(n-1)-1;\alpha} = F_{2,11;0'10} = 2{,}86$$

se acepta la hipótesis nula.$H_0 \equiv \alpha_i = 0$, para $i = 1, 2, 3$.

Se estima el coeficiente de regresión

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}} = \frac{186{,}6}{195{,}6} = 0{,}954.$$

Para contrastar la hipótesis $H_0 \equiv \beta = 0$ se usa el estadístico

$$F_0 = \frac{\frac{E_{xy}^2}{E_{xx}}}{MCE} = \frac{\frac{186{,}6^2}{195{,}6}}{2{,}54} = 70{,}08$$

Como $F_{1,a(n-1)-1} = F_{1,11;0'10} = 9{,}65$, se rechaza $H_0 \equiv \beta = 0$, a nivel 0,10. Con lo cual la corrección mediante el análisis de la covarianza es necesario.

Las medias de los tratamientos ajustadas son

$$\begin{aligned}
\text{Ajust_}\bar{y}_{i\cdot} &= \bar{y}_{i\cdot} - \hat{\beta}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}) \implies \\
\text{Ajust_}\bar{y}_{1\cdot} &= 41{,}4 - 0{,}954 \cdot (25{,}2 - 24{,}13) = 40{,}38 \\
\text{Ajust_}\bar{y}_{2\cdot} &= 43{,}2 - 0{,}954 \cdot (26. - 24{,}13) = 41{,}42 \\
\text{Ajust_}\bar{y}_{3\cdot} &= 36 - 0{,}954 \cdot (21{,}2. - 24{,}13) = 38{,}8
\end{aligned}$$

Si se comparan las medias de los tratamientos sin ajustar con las ajustadas, se observa que estas últimas están mucho más próximas entre sí, lo cual es otra indicación de la necesidad de hacer un análisis de la covarianza.

Modelo de Medidas Repetidas

En numerosas ocasiones, los sujetos que se estudian son sujetos que pueden presentar numerosas diferencias entre ellos ante el mismo tratamiento, introduciéndose, entonces, una mayor fuente de error experimental. Aumenta, así, la media de cuadrados de los errores haciendo difícil distinguir las diferencias entre los tratamientos. Una manera de controlar esta variabilidad entre los sujetos, consiste en aplicar a cada uno de ellos los a tratamientos. Este diseño se denomina de *medidas repetidas*. Equivale a un diseño por bloques completos, donde la variable bloque son los sujetos, siendo ésta de efectos aleatorios.

Supongamos un experimento donde aparece un factor con a tratamientos, y que cada tratamiento se aplica exactamente sobre cada uno de los n individuos:

	Sujetos				
Tratamientos	1	2	$\cdots$	n	*Totales*
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a\cdot}$
Totales	$y_{\cdot 1}$	$y_{\cdot 2}$	$\cdots$	$y_{\cdot n}$	$y_{\cdot\cdot}$

La observación y_{ij} es la respuesta del sujeto j al tratamiento i y sólo se usan n sujetos. El modelo se escribe como

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

donde α_i es el efecto del $i-$ésimo tratamiento y β_j el efecto del j-ésimo sujeto.

Se supone que

$$\sum_{i=1}^{a} \alpha_i = 0,$$

y que los individuos son una muestra aleatoria de una población dada, de modo que los individuos actúan como un efecto aleatorio

$$\beta_j \sim N(0, \sigma_\beta^2).$$

Dado que β_j es común a todos los a tratamientos medidos sobre el mismo sujeto j, la covarianza entre y_{ij} e $y_{i'j}$ es, en general, diferente de 0, asumiéndose que es constante sobre los tratamientos y los sujetos.

La suma total de cuadrados se parte en dos sumatorios:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = a\sum_{j=1}^{n}(\bar{y}_{.j}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{.j})^2$$

El primer término de la suma de cuadrados recoge la diferencia *entre* sujetos y el segundo término la diferencia *dentro* de sujetos, esto es,

$$SCT = SC_{entre} + SC_{dentro}$$

de modo que ambas sumas de cuadrados son independientes entre sí con

$$an - 1 = (n-1) + n(a-1)$$

grados de libertad.

Las diferencias dentro de los sujetos dependen tanto de las diferencias en los efectos de los tratamientos como del ruido. Así, se descompone la suma de cuadrados dentro de sujetos de la siguiente forma:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{.j})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2.$$

El primer término mide la contribución de las diferencias entre las medias de los tratamientos a la suma de cuadrados dentro de los sujetos, y el segundo término es la variación residual debido al error. Ambos términos son independientes. Así,

$$SC_{dentro} = SCTra + SCE$$

con

$$n(a-1) = (a-1) + (a-1)(n-1)$$

grados de libertad respectivamente.

Se contrasta

$$\begin{aligned} H_0 &\equiv \alpha_i = 0, \qquad i = 1, \ldots, a \\ H_1 &\equiv \alpha_i \neq 0, \qquad \text{para algún } i \end{aligned}$$

usándose el cociente

$$F_0 = \frac{\dfrac{SCTra}{a-1}}{\dfrac{SCE}{(a-1)(n-1)}} = \frac{MCTra}{MCE}$$

que, cuando H_0 es cierta, se distribuye como una F de Snedecor $F_{a-1,(a-1)(n-1)}$. Se rechaza H_0 a nivel α cuando

$$F_0 > F_{a-1,(a-1)(n-1)}.$$

La tabla de análisis de la varianza es

Fuentes Variación	**Suma de Cuadrados**	**grados libertad**	**F**
Entre Sujetos	$\sum_{j=1}^{n} \frac{y_{\cdot j}^2}{a} - \frac{y_{\cdot\cdot}^2}{an}$	$n-1$	
Dentro Sujetos	$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \sum_{j=1}^{n} \frac{y_{\cdot j}^2}{a}$	$n(a-1)$	
Tratamientos	$SCTra = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{a} - \frac{y_{\cdot\cdot}^2}{an}$	$a-1$	$F_0 = \frac{\frac{SCTra}{a-1}}{\frac{SCE}{(a-1)(n-1)}}$
Residual	$SCE = (2) - (3)$	$(a-1)(n-1)$	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{an}$	$an-1$	