Análisis de la covarianza

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

Enero 2002

El análisis de la covarianza es una técnica estadística que, utilizando un modelo de regresión lineal múltiple, busca comparar los resultados obtenidos en diferentes grupos de una variable cuantitativa, pero "corrigiendo" las posibles diferencias existentes entre los grupos en otras variables que puedieran afectar también al resultado (covariantes).

Supongamos que se está comparando la presión arterial sistólica de un grupo de mujeres diabéticas según su nivel de estudios, para lo que efectuamos un análisis de la varianza cuyos resultados se resumen a continuación:

Estadísita Descriptiva PAS

	Sin estudios	1º grado	2º y 3º grado
Media	141,16	140,93	131,27
Desv.Típ.	13,67	16,23	18,03
Tamaño	215	202	60

Análisis de la varianza

Fuente var.	Suma cuadrados	gl	Varianza	F	p	Nivel signif.
Factor	5020,04	2	2510,02	10,61	0,0000310	p < 0.001
Residual	112119,55	474	236,54			
Total	117139,59	476	246,09			

Vemos que hay diferencias estadísticamente significativas en cuanto a la media de la PAS entre los diferentes niveles de estudios, siendo inferior la media de PAS en el grupo de mujeres con estudios de 2º o 3º grado (del orden de 10 mmHg inferior). Ahora bien, sabemos que uno de los principales factores de riesgo en la hipertensión es la edad, por lo que nos podemos plantear que al tratarse de un estudio observacional, en el que las pacientes han sido seleccionadas de forma aleatoria entre las que acuden a la consulta, si éstas fueran representativas de la población, es de sospechar que las mujeres con mayor nivel de estudios sean en promedio más jóvenes, debido a que en el pasado las mujeres solían a menudo recibir como mucho una formación elemental.

Si para comprobarlo efectuamos un análisis de la varianza para la edad según el nivel de estudios, los resultados que obtenemos son

Estadística Descriptiva Edad

	Sin estudios	1º grado	2º y 3º grado
Media	69,75	64,80	54,25
Desv.Típ.	8,26	10,57	18,33
Tamaño	215	202	60

Análisis de la varianza

Fuente var.	Suma cuadrados	gl	Varianza	F	p	Nivel signif.
Factor	11563,46	2	5781,73	48,19	0,0000	p < 0.001
Residual	56869,86	474	119,98			
Total	68433,32	476	143,77			

donde, como nos temíamos, la edad media de las mujeres con estudios de 2º o 3º grado es inferior a la de los otros grupos, lo que por sí solo podría explicar las diferencias encontradas en cuanto a la media de PAS.

Utilizando el análisis de la covarianza nos planteamos la posiblidad de "*corregir*" o "*ajustar*" esa diferencia de edad, con el fin de hacer comparables los grupos. Para ello se construye un modelo de regresión entre la variable resultado PAS y la variable de confusión EDAD y la pregunta que nos hacemos es ¿explica la regresión por sí sola la diferencia de PAS media observada entre los grupos?.

Vamos pues a estimar una ecuación de regresión entre la PAS y la EDAD, pero ¿qué tipo de regresión?, porque tenemos tres posibilidades, que vamos a representar para el caso de que haya sólo dos grupos de estudio:

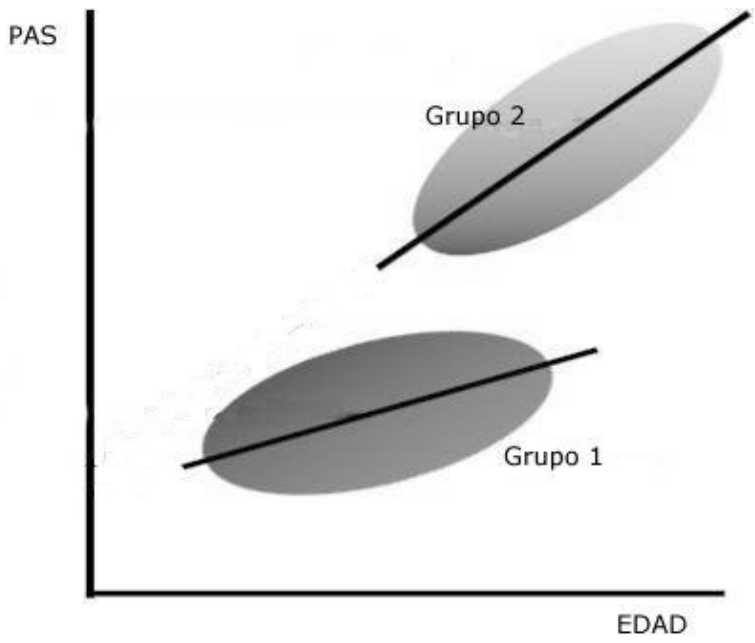

Fig.1 Pendiente de regresión diferente para cada grupo

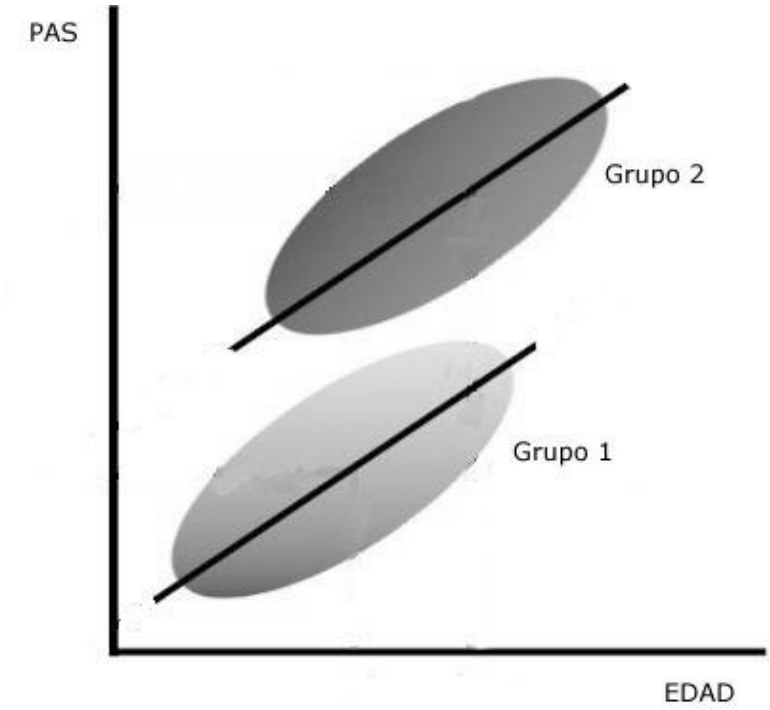

Fig.2 Igual pendiente para los grupos, a diferente altura

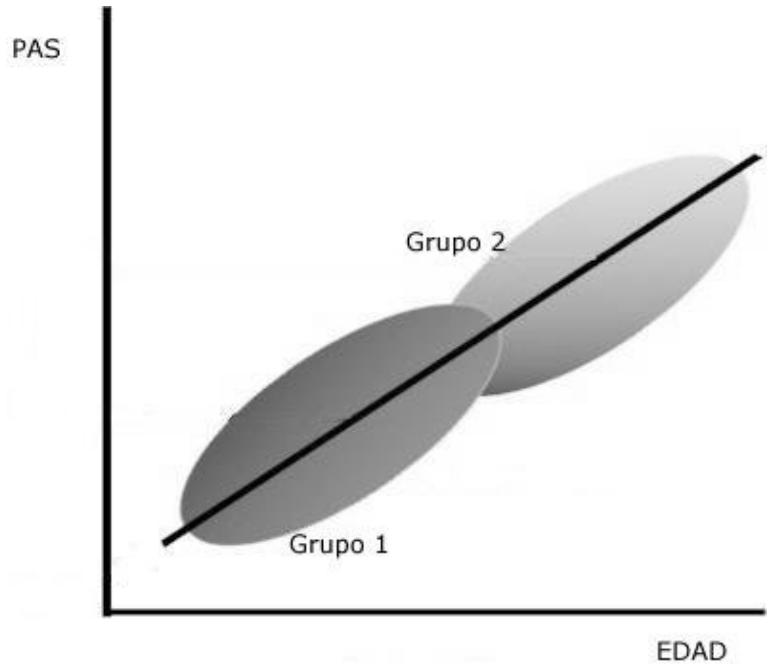

Fig.3 Igual pendiente, misma altura

Se trata pues de decidir, a la luz de nuestros datos, cuál de las tres posibilidades es más verosímil .

En la primera figura vemos que hay **interacción** entre la variable para la que ajustamos, covariante, y el grupo, de tal manera que en uno de los grupos la relación entre la PAS y la edad es más acusada, aumenta más rápidamente al aumentar la edad.

Cuando existe interacción la interpretación es complicada ya que puede incluso ocurrir que en uno de los grupos esa relación se invierta y que al aumentar el covariante X el valor de Y disminuya (pendiente negativa).

En el análisis de la covarianza en primer lugar nos planteamos si es razonable creer que la regresión tiene pendientes diferentes en cada grupo o si por el contrario es verosímil pensar que la pendiente se mantiene, pudiendo entonces considerar una pendiente común para todos los grupos. Solo en el caso de que aceptemos esta última situación tiene sentido decidir entre la segunda y tercera alternativa: plantearnos si la diferencia observada entre los grupos se explica sólo por la regresión (figura 3) o por algo más.

Una vez aceptada la hipótesis de igual pendiente en todos los grupos, el razonamiento a seguir se explica de forma gráfica en la figura 4, aunque un tanto exagerado

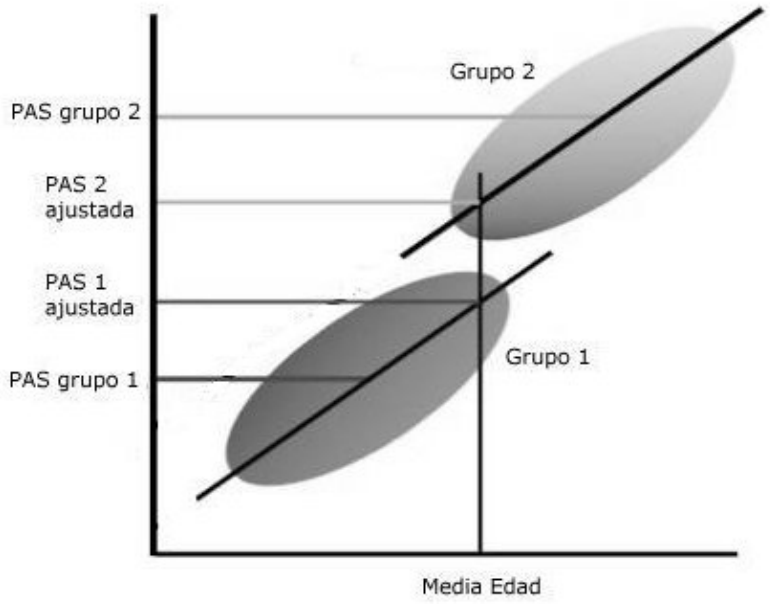

Fig. 4 Comparación de medias ajustadas

Ahora, dado que no hay razón para pensar que la pendiente de la ecuación sea diferente en cada grupo, calculamos cuál sería el valor de la PAS previsto por la ecuación de regresión para la media global de la edad (media calculada combinando ambos grupos), y determinamos el valor de la PAS estimado a partir de la ecuación de regresión en cada grupo, este valor es lo que denominamos **medias ajustadas** de la PAS: aquellas que obtendríamos si ambos grupos hubiesen tenido la misma media de edad. Vemos claramente en el dibujo como la diferencia de medias de PAS ajustadas ha disminuido con respecto a la diferencia de medias sin ajustar, y será tanto menor cuanto más nos acerquemos a la situación reflejada en la figura 3, cuanto menor sea la separación de alturas entre las dos rectas de regresión.

En la siguiente figura vemos ilustrado un caso en el que la media sin ajustar de la PAS para el grupo 2 es inferior a la del grupo 1 (representado por las líneas rectas en la figura), mientras que si efectúa la corrección para la edad estaremos en la misma situación de la figura anterior: media ajustada del grupo 2 superior a la del grupo 1; situación debida a que en el grupo 2 tenemos edades más bajas que en el grupo 1.

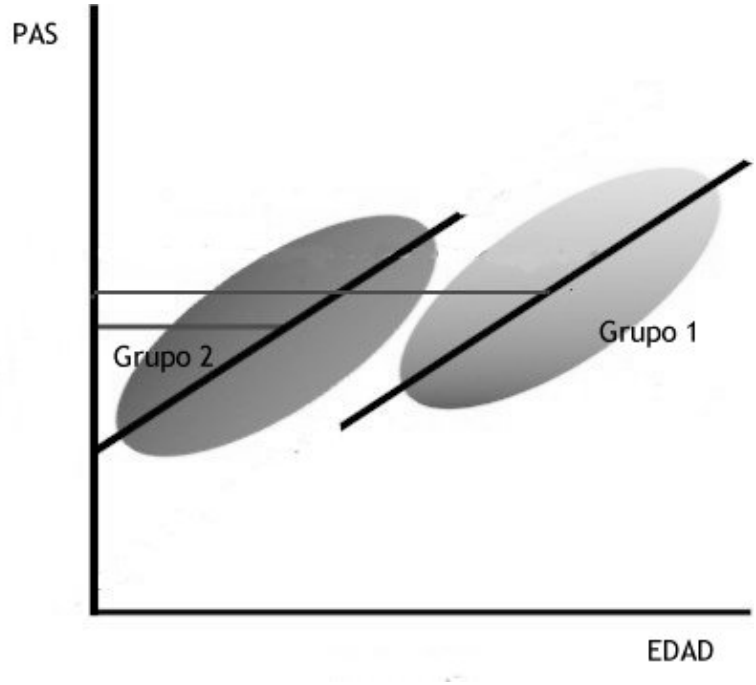

Fig. 5

Veamos cuales son nuestros resultados cuando efectuamos un análisis de la covarianza en el ejemplo planteado:

Análisis de la covarianza PAS

Fuente var.	Suma cuadrados	gl	Varianza	F	p	Nivel signif.
Pendientes iguales	596,28	2	298,14	1,306	0,2718	NO
Error	107501	471	228,24			
Igualdad medias ajustadas	1994,7	2	997,34	4,364	0,0132	p < 0.05
Error	108097	473	228,53			

Medias PAS

Grupo	Media	Media ajustada	Nº de casos
Sin estudios	141,16	140,08	215
1º grado	140,93	141,17	202
2º y 3º grado	131,27	134,32	60
Total	139,82		477

Medias covariante EDAD según ESTUDIOS

	Sin estudios	1º grado	2º o 3º grado
EDAD	69,75	64,80	54,25

Los contrastes en el análisis de la covarianza se efectuan utilizando el valor de la suma de cuadrados medios residual de Y (la variable respuesta estudiada) en cada una de las tres situaciones, por lo que el parámetro obtenido se distribuye según una F.

En primer lugar efectuamos un contraste para ver si es razonable suponer pendientes iguales (p=0,27) hipótesis que no llegamos a rechazar. Una vez aceptada esa premisa tiene sentido plantearnos el comprobar si son iguales las medias ajustadas, hipótesis que en este ejemplo rechazamos (p=0.013), aunque ahora la diferencia de medias no es tan acusada como antes de ajustar.

El modelo planteado se puede extender a más de un covariante, en ese caso el ajuste se realiza de tal manera que los cálculos se efectúan como si todos los grupos hubiesen tenido la misma distribución de covariantes.

Enlaces de interés

- Calculadora on line para el análisis de la covarianza
- Covariance designs
- A New View of Statistics. Analysis of Covariance (ANCOVA)
- PROPHET StatGuide: Analysis of Covariance (ANCOVA)−−Comparing simple linear regression lines

Bibliografía seleccionada

- **Applied regression analysis and other multivariate methods**
  Kleinbaum, Kupper, Muller, Nizam
  Ed. Duxbury Press 1998
- **The analysis of covariance and alternatives**
  Huitema, B
  Ed. Wiley 1980

Indice de artículos

Principio de la página