Análisis de la Covarianza con R

Francesc Carmona
Departamento de Genética, Microbiología y Estadística

4 de diciembre de 2018

1. Introducción

El Análisis de la Covarianza es una síntesis del Análisis de la Varianza y los métodos de Regresión. Combina por lo tanto, unas variables cualitativas con unas variables cuantitativas. Se trata estudiar las diferencias entre los niveles de un factor o contrastar la significación de algunos factores sobre una variable cuantitativa observable, cuando alguna o algunas variables regresoras, llamadas *concomitantes*, influyen también en la respuesta.
Desde el punto de vista de la Regresión, se trata de considerar, junto a las variables regresoras cuantitativas, variables predictoras cualitativas, como por ejemplo el sexo, que se califican de categóricas o, más técnicamente, como factores.
El Análisis de la Covarianza tiene las siguientes características (ver [2]):

- Tiene en cuenta la influencia de las variables concomitantes sobre la variable observable o respuesta.
- La variable concomitante es siempre cuantitativa. Cada réplica debe tener asociado un valor de la variable concomitante.
- Las variables concomitantes no se utilizan como variables de referencia para contrastar hipótesis. Lo que se pretende es eliminar su influencia sobre la variable observable.
- La varianza del diseño queda reducida al introducir una variable concomitante. Una consecuencia es el aumento de la precisión en las conclusiones.
- En general se logra simplificar el diseño, reduciendo el número de factores, lo que redundará en un número menor de réplicas.
- La interpretación del diseño es más fácil cuando los factores sólo influyen en la variable respuesta y no en las variables concomitantes.

En este documento se resuelven algunos ejemplos sencillos con el programa estadístico R.
Para profundizar en la teoría del Análisis de la Covarianza se puede consultar, entre otros, el libro clásico de Snedecor y Cochran[5]. Para estudiar modelos lineales avanzados con R se puede leer el libro de J.J. Faraway[3].

2. Un factor y una variable concomitante

Consideremos un factor con k niveles. Sea y_{ij}, $j = 1,\ldots,n_i$, la réplica j del nivel i. Supongamos que cada y_{ij} está relacionado con una observación concomitante x_{ij}. En esta situación podríamos considerar un Análisis de la Varianza únicamente con el factor o un Análisis de la Regresión con únicamente la variable concomitante como predictora, pero es mejor un modelo que combina ambos análisis:

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + \varepsilon_{ij} \qquad i = 1,\ldots,k; j = 1,\ldots,n_i$$

donde los errores ε_{ij} verifican todas las hipótesis de un modelo lineal normal.
Sin embargo en este modelo no se ha considerado una posible interacción no deseada entre el factor y la variable concomitante. Vamos a tratar este caso en un ejemplo muy sencillo y con un modelo alternativo.

2.1. Un factor con dos niveles

El ejemplo que explicamos a continuación se basa en un estudio realizado por Alan Pearson, veterinario del *Animal Health Laboratory*, Lincoln, Nueva Zelanda y se puede hallar en el libro de Saville y Wood[4]. El experimento tenía como objetivo determinar si el programa estándar de desparasitado por vía oral en 6 granjas de cabras era adecuado. Para ello se seleccionaron 40 cabras en cada granja. Veinte de ellas, elegidas completamente al azar, se desparasitaron con el programa estándar, mientras que las veinte restantes se desparasitaron con más frecuencia. Las cabras se pesaron al principio y al final del estudio que duró un año. Para nuestro ejemplo hemos tomado los datos de una única granja. Así pues, las variables consideradas son:

- Aumento de peso en vivo (kg.)
- Peso al inicio (kg.)
- Tratamiento: estándar o intensivo

Los datos se pueden descargar de internet:

```
> goats <- read.table("goats.data", skip=1)
> names(goats) <- c("treatment", "weightgain", "initial.wt")
> goats$treatment <- factor(goats$treatment,
+                           labels = c("standard","intensive"))
```

Vamos a echar un vistazo a los datos:

```
> by(goats, goats$treatment, summary)
```

```
goats$treatment: standard
     treatment    weightgain      initial.wt
 standard :20   Min.   : 2.00   Min.   :18.00
 intensive: 0   1st Qu.: 4.00   1st Qu.:20.75
                Median : 5.50   Median :22.50
                Mean   : 5.55   Mean   :23.20
                3rd Qu.: 7.00   3rd Qu.:26.25
                Max.   :10.00   Max.   :30.00
------------------------------------------------------------
goats$treatment: intensive
     treatment    weightgain      initial.wt
 standard : 0   Min.   : 3.00   Min.   :18.00
 intensive:20   1st Qu.: 5.75   1st Qu.:19.75
                Median : 7.00   Median :23.50
                Mean   : 6.85   Mean   :23.10
                3rd Qu.: 8.00   3rd Qu.:25.25
                Max.   :11.00   Max.   :30.00
```

Ahora dibujamos un par de gráficos:

```
> op<-par(mfrow = c(1,2),pty="s")
> boxplot(weightgain ~ treatment, goats)
> plot(weightgain ~ initial.wt, pch=ifelse(treatment=="standard",1,16), data=goats)
> par(op)
```

Vemos en la figura 1 que las cabras que recibieron un tratamiento intensivo tienen mayor aumento de peso. Como el factor tratamiento tiene sólo dos niveles podemos hacer un contraste con la t de Student:

```
> t.test(weightgain ~ treatment, data=goats)
```

```
        Welch Two Sample t-test

data:  weightgain by treatment
t = -2.0322, df = 37.936, p-value = 0.04918
```

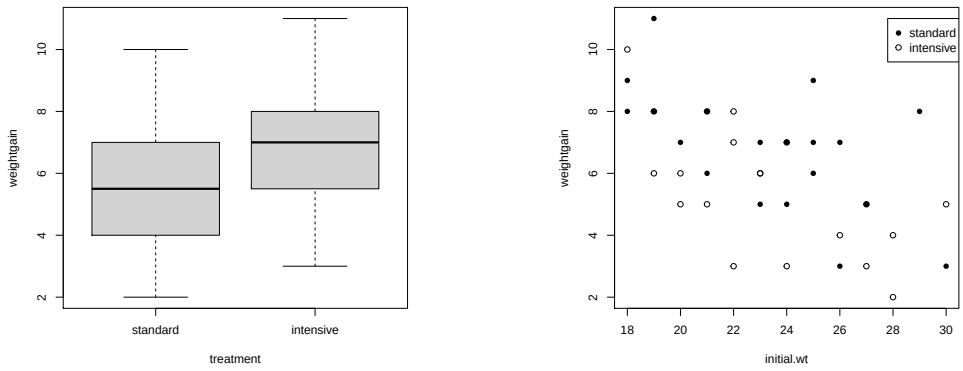

Figura 1: Comparación del aumento de peso.

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.595068186 -0.004931814
sample estimates:
 mean in group standard mean in group intensive
                   5.55                    6.85
```

y la diferencia es significativa.
Sin embargo, en el gráfico de dispersión de la derecha observamos una correlación negativa entre el aumento de peso y el peso inicial y en el summary anterior vimos que el peso inicial de los dos grupos parece equilibrado. El Análisis de la Covarianza nos permitirá investigar la verdadera influencia del factor y la variable concomitante en el aumento de peso.
Antes de considerar algunos modelos para el Análisis de la Covarianza, debemos observar el orden de los niveles del factor. A veces, al capturar los datos de un factor, los niveles tienen el orden alfabético. No es nuestro caso, pero si así fuera, debemos situar como primer nivel el tratamiento normal establecido o basal que aquí es el estándar. La modificación en R se haría así:

```
> goats$treatment <- relevel(goats$treatment, ref="standard")
```

Para incorporar un factor al análisis de la regresión vamos a considerar una variable dicotómica (*dummy variable*) de la siguiente forma

$$d_{ij} = \begin{cases} 0 & \text{si } i = 1, \text{ tratamiento estándar} \\ 1 & \text{si } i = 2, \text{ tratamiento intensivo} \end{cases}$$

Los posibles modelos de regresión que podemos considerar son:

1. La misma recta de regresión para los dos grupos $y = \beta_0 + \beta_1 x + \varepsilon$ que en R se escribe `y ~ x`.
2. Una recta de regresión para cada grupo con la misma pendiente: $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$ que en R se escribe `y ~ x + d`.
3. Rectas de regresión separadas para cada grupo con diferentes pendientes:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x \cdot d + \varepsilon$$

que en R se escribe `y ~ x + d + x:d` o también `y ~ x * d`.

Vamos a ajustar este último modelo de regresión con los datos de nuestro ejemplo:

```
> g1 <- lm(weightgain ~ initial.wt * treatment, data=goats)
> summary(g1)
```

```
Call:
lm(formula = weightgain ~ initial.wt * treatment, data = goats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0053 -1.2038 -0.0339  0.9175  3.0714

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  14.35211    2.54662   5.636 2.13e-06 ***
initial.wt                   -0.37940    0.10863  -3.493  0.00128 **
treatmentintensive            0.02077    3.52029   0.006  0.99533
initial.wt:treatmentintensive 0.05374    0.15040   0.357  0.72296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.637 on 36 degrees of freedom
Multiple R-squared:  0.4402,        Adjusted R-squared:  0.3935
F-statistic: 9.435 on 3 and 36 DF,  p-value: 9.765e-05
```

Como `treatment` no es numérica, R la trata automáticamente como cualitativa y procede a asignarle el código como vemos en la matriz de diseño:

```
> model.matrix(g1)[1:3, ]

  (Intercept) initial.wt treatmentintensive initial.wt:treatmentintensive
1           1         21                  0                             0
2           1         24                  0                             0
3           1         21                  0                             0


> model.matrix(g1)[38:40, ]

   (Intercept) initial.wt treatmentintensive initial.wt:treatmentintensive
38           1         27                  1                            27
39           1         30                  1                            30
40           1         29                  1                            29
```

Como se ha fijado, con la instrucción `relevel()` si fuera necesario, el nivel de referencia es el tratamiento estándar y su código es 0, mientras que el nivel de tratamiento intensivo tiene código 1. El término de interacción se representa en la cuarta columna de la matriz y es el producto de la segunda y la tercera columnas. Vemos que el modelo se puede simplificar ya que el término de interacción no es significativo. Así el modelo es:

```
> g2 <- lm(weightgain ~ initial.wt + treatment, data=goats)
> summary(g2)


Call:
lm(formula = weightgain ~ initial.wt + treatment, data = goats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9716 -1.2419 -0.0338  0.9878  3.2231

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.70175    1.75987   7.786 2.61e-09 ***
initial.wt        -0.35137    0.07424  -4.733 3.21e-05 ***
treatmentintensive  1.26486    0.51169   2.472   0.0182 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.618 on 37 degrees of freedom
Multiple R-squared:  0.4382,        Adjusted R-squared:  0.4078
F-statistic: 14.43 on 2 and 37 DF,  p-value: 2.331e-05
```

El contraste entre el modelo simple `g2` y el modelo `g1` se conoce como el contraste de paralelismo y es equivalente al contraste de significación de la interacción. El modelo `g2` ya no se puede simplificar más puesto que las variables predictoras restantes son significativas. En el gráfico de la figura 2 se observan las dos rectas paralelas. La pendiente de ambas rectas es $-0{,}35137$, pero la recta del tratamiento estándar

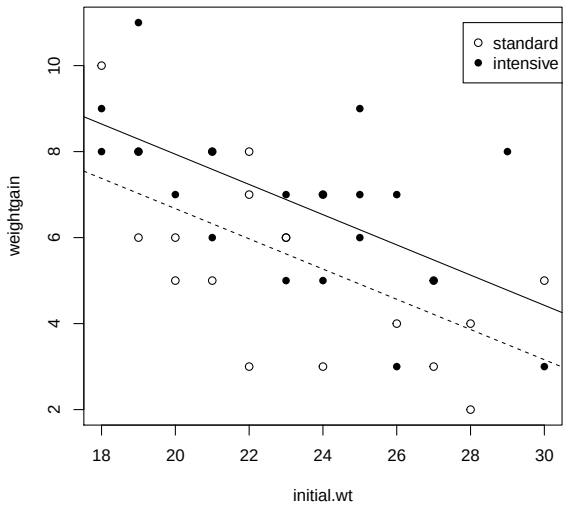

Figura 2: Dos rectas de regresión paralelas.

está 1,26486 por debajo frente al tratamiento intensivo. Del contraste t de Student anterior resultaba una diferencia de $6{,}85 - 5{,}55 = 1{,}3$ en el aumento de peso medio por efecto del tratamiento, luego al ajustar teniendo en cuenta el peso inicial el estimador del efecto del tratamiento se ha reducido ligeramente. También podemos comparar el intervalo de confianza del efecto del tratamiento

```
> confint(g2)[3,]

    2.5 %    97.5 %
0.2280746 2.3016517
```

que comparado con $(0{,}005; 2{,}595)$ es más estrecho. En general, el diseño de experimentos con variables concomitantes mejora la precisión de los estimadores de un efecto.
Para finalizar debemos realizar el imprescindible análisis de los residuos para estudiar las hipótesis del modelo y que en este ejemplo no muestra ninguna patología.

```
> plot(g2)
```

2.2. Un factor multinivel

En el libro de Faraway[3] pág. 174 podemos seguir un ejemplo con un factor multinivel que se resuelve de forma similar al anterior.

```
> library(faraway)
> data(fruitfly)
> g <- lm(longevity ~ thorax * activity, fruitfly)
> summary(g)
```

```
Call:
lm(formula = longevity ~ thorax * activity, data = fruitfly)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9509  -6.7296  -0.9103   6.1854  30.3071

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -50.2420    21.8012  -2.305    0.023 *
thorax              136.1268    25.9517   5.245 7.27e-07 ***
activityone           6.5172    33.8708   0.192    0.848
activitylow          -7.7501    33.9690  -0.228    0.820
activitymany         -1.1394    32.5298  -0.035    0.972
activityhigh        -11.0380    31.2866  -0.353    0.725
thorax:activityone   -4.6771    40.6518  -0.115    0.909
thorax:activitylow    0.8743    40.4253   0.022    0.983
thorax:activitymany   6.5478    39.3600   0.166    0.868
thorax:activityhigh -11.1268    38.1200  -0.292    0.771
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.71 on 114 degrees of freedom
Multiple R-squared:  0.6534,        Adjusted R-squared:  0.626
F-statistic: 23.88 on 9 and 114 DF,  p-value: < 2.2e-16
```

```
> model.matrix(g)[1:3, ]

  (Intercept) thorax activityone activitylow activitymany activityhigh
1           1   0.68           0           0            1            0
2           1   0.68           0           0            1            0
3           1   0.72           0           0            1            0
  thorax:activityone thorax:activitylow thorax:activitymany thorax:activityhigh
1                  0                  0                0.68                   0
2                  0                  0                0.68                   0
3                  0                  0                0.72                   0
```

```
> anova(g) # ANOVA secuencial

Analysis of Variance Table

Response: longevity
                Df  Sum Sq Mean Sq F value    Pr(>F)
thorax           1 15003.3 15003.3 130.733 < 2.2e-16 ***
activity         4  9634.6  2408.6  20.988 5.503e-13 ***
thorax:activity  4    24.3     6.1   0.053    0.9947
Residuals      114 13083.0   114.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Referencias

[1] F. Carmona, *Modelos lineales*, Publicacions UB, 2005.

[2] C.M. Cuadras, *Problemas de Probabilidades y Estadística*. Vol.2:Inferencia Estadística. EUB, 2000.

[3] J.J. Faraway, *Linear Models with R*, Chapman & Hall/CRC, 2014.

[4] D.J. Saville y G.R. Wood, *Statistical Methods: The Geometric Approach*, Cap. 17, New York:Springer, 1991.

[5] G.W. Snedecor y W.G. Cochran, *Statistical Methods*, Iowa State University Press, 1989.

[6] J. Verzani, *Using R for Introductory Statistics*. Chapman & Hall/CRC, 2004.