

Comparación de rectas de regresión con R

Francesc Carmona

Regresión, modelos y métodos

Máster universitario de Bioinformática y Bioestadística

25 de marzo de 2019

1. Introducción

Supongamos la situación en la que disponemos de datos de dos poblaciones o dos tratamientos y sus correspondientes rectas de regresión. El objetivo de este artículo es presentar diversas formas de comparar dichas rectas y su inferencia. Pongamos un ejemplo.

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[5]. 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)

```

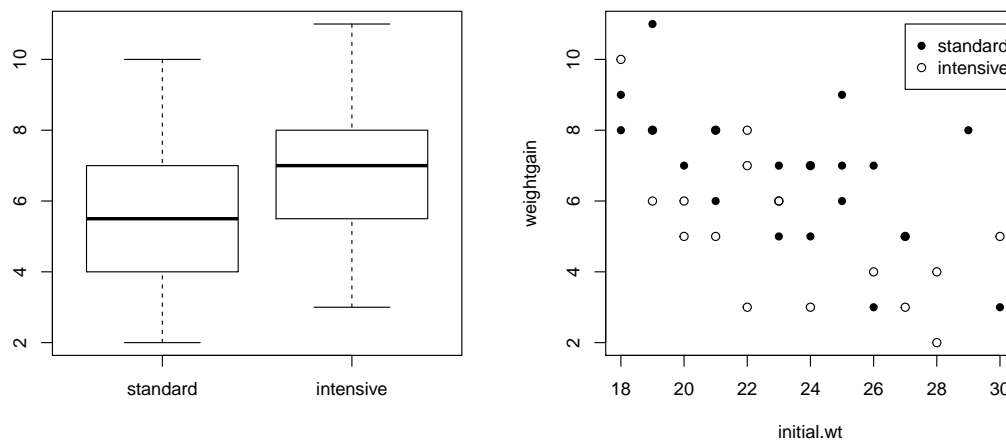


Figura 1: Comparación del aumento de peso.

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

$t = -2.0322$, $df = 37.936$, $p\text{-value} = 0.04918$

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 permite investigar la verdadera influencia del factor y la llamada variable *concomitante* (el peso inicial) en el aumento de peso.

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 ejemplo planteado y otros ejemplos sencillos se resuelven en el documento de F. Carmona[1] con el Análisis de la Covarianza y 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[6]. Para estudiar modelos lineales avanzados con R se puede leer el libro de J.J. Faraway[4].

Sin embargo, en este artículo vamos a tratar este ejemplo como un problema de comparación de rectas de regresión. Es decir, las preguntas sobre los tratamientos son del tipo: ¿las rectas de regresión son paralelas? ¿coinciden? ¿se cortan en un punto determinado?

En el apartado 6.7 del libro de F. Carmona[2] se describen los contrastes, los tests F y sus fórmulas explícitas para la comparación de dos o más rectas de regresión. Realmente resultan un conjunto de fórmulas y tests complejos. Ahora bien, si el problema lo tratamos desde el punto de vista de un contraste de modelos, la solución en R es más sencilla o eso parece.

2. Contraste de paralelismo

En primer lugar consideremos las dos rectas por separado. El gráfico 2 nos muestra la situación.

La primera pregunta que nos podemos plantear es si las dos rectas son paralelas. Para resolver el contraste debemos escribir los modelos de las dos rectas, en principio por separado.

Los dos modelos asociados a cada una de las rectas son:

$$\begin{aligned} y_{1i} &= \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} & i &= 1, \dots, n_1 \\ y_{2i} &= \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} & i &= 1, \dots, n_2 \end{aligned}$$

donde los errores ϵ_{1i} y ϵ_{2i} verifican todas las hipótesis de un modelo lineal normal. Entonces la pregunta se puede plantear en términos paramétricos como un contraste con las pendientes de las rectas:

$$\begin{aligned} H_0 : & \beta_1 = \beta_2 \\ H_1 : & \beta_1 \neq \beta_2 \end{aligned}$$

Ahora bien, para poder acometer este contraste hace falta considerar un modelo general único con los cuatro parámetros $\alpha_1, \beta_1, \alpha_2, \beta_2$ que verifique las condiciones de Gauss-Markov, en particular la homocedasticidad (igualdad de las varianzas de los errores).

Este modelo general tendrá este aspecto:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

La concreción a nuestro ejemplo con R es la siguiente:

```

> plot(weightgain ~ initial.wt, pch=ifelse(treatment=="standard",1,16), data=goats)
> legend(27, 11, levels(goats$treatment), pch=c(16,1))
> intensive.lm <- lm(weightgain ~ initial.wt, data=goats, sub= treatment=="intensive")
> standard.lm <- lm(weightgain ~ initial.wt, data=goats, sub= treatment=="standard")
> abline(intensive.lm, lty=1)
> abline(standard.lm, lty=2)

```

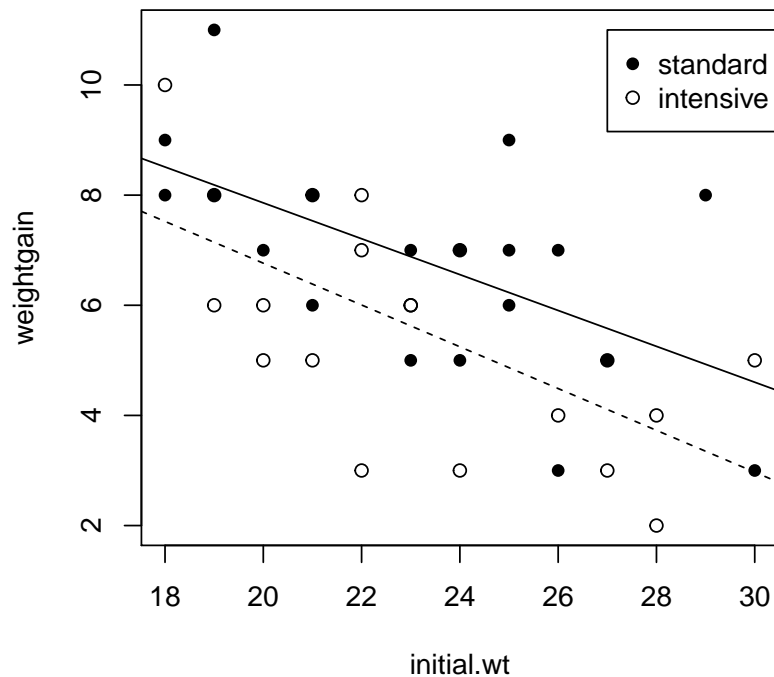


Figura 2: Comparación de rectas de regresión.

```

> attach(goats)
> n1 <- sum(treatment=="intensive")
> n2 <- sum(treatment=="standard")
> y <- c(weightgain[treatment=="intensive"], weightgain[treatment=="standard"])
> x <- matrix(numeric(4*(n1+n2)), ncol=4)
> intensive.lm <- lm(weightgain ~ initial.wt, data=goats, sub= treatment=="intensive")
> standard.lm <- lm(weightgain ~ initial.wt, data=goats, sub= treatment=="standard")
> x[1:n1,1:2] <- model.matrix(intensive.lm)
> x[(n1+1):(n1+n2),3:4] <- model.matrix(standard.lm)
> general.lm <- lm(y ~ 0 + x)
> model.matrix(general.lm)[c(1:5,36:40), ]

```

	x1	x2	x3	x4
1	1	18	0	0
2	1	18	0	0

```

3   1 19 0 0
4   1 19 0 0
5   1 19 0 0
36  0 0 1 27
37  0 0 1 19
38  0 0 1 20
39  0 0 1 19
40  0 0 1 22

```

La última instrucción nos permite ver las cinco primeras y las cinco últimas filas de la matriz de diseño del modelo construido.

El siguiente paso es definir el modelo bajo la hipótesis nula. Si las rectas son paralelas, las pendientes de ambas serán iguales. Sea β esa pendiente común. Entonces el modelo que representa a la hipótesis nula es:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_{11} \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{1n_1} \\ 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

Este modelo se puede escribir en R de la siguiente forma:

```

> x0 <- matrix(numeric(3*(n1+n2)), ncol=3)
> x0[,1] <- x[,1]
> x0[,2] <- x[,3]
> x0[,3] <- x[,2] + x[,4]
> h0.lm <- lm(y ~ 0 + x0)

```

y el test F que resuelve el contraste es

```

> anova(h0.lm,general.lm)

Analysis of Variance Table

Model 1: y ~ 0 + x0
Model 2: y ~ 0 + x
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     37 96.857
2     36 96.514  1    0.34225 0.1277  0.723

```

de forma que aceptamos la hipótesis de paralelismo.

Un tema importante es que “a posteriori” debemos comprobar las hipótesis de Gauss-Markov para estos modelos. Si suponemos que las rectas de regresión las verificaban, sólo nos quedaría estudiar la igualdad de varianzas de los errores de las dos rectas. La independencia de los errores entre las dos rectas se supone al ser muestras independientes. Para contrastar la igualdad de varianzas se puede hacer con un test como el de Bartlett entre los residuos de las dos rectas o con otro test apropiado.

Se deja para el lector el contraste de igualdad de los términos de intercepción de las dos rectas y el contraste de coincidencia total (cuando se acepta el paralelismo).

3. Análisis de la Covarianza

Como se explica en el documento de F. Carmona[1] para resolver el contraste de paralelismo y otros con el Análisis de la Covarianza hay que considerar un modelo con la variable concomitante `initial.wt`, pero también con el factor `treatment` y su interacción.

El modelo en R es

```
> 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

El contraste de significación de la interacción equivale al contraste de paralelismo como se puede comprobar así:

```
> g0 <- lm(weightgain ~ initial.wt + treatment, data=goats)
> anova(g0,g1)
```

Analysis of Variance Table

Model 1: weightgain ~ initial.wt + treatment
Model 2: weightgain ~ initial.wt * treatment

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	96.857				
2	36	96.514	1	0.34225	0.1277	0.723

El siguiente contraste secuencial permite contrastar (de abajo arriba) la hipótesis de paralelismo, la de coincidencia y la significación de la variable concomitante.

```
> anova(g1)
```

Analysis of Variance Table

Response: weightgain

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
initial.wt	1	59.548	59.548	22.2115	3.6e-05 ***
treatment	1	15.995	15.995	5.9663	0.01962 *
initial.wt:treatment	1	0.342	0.342	0.1277	0.72296
Residuals	36	96.514	2.681		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Referencias

- [1] F. Carmona, *Análisis de la Covarianza con R*, 2013.
<http://erre-que-erre-paco.blogspot.com.es/2013/12/analisis-de-la-covarianza-con-r.html>
- [2] F. Carmona, *Modelos lineales*, Publicacions UB, 2005.
- [3] C.M. Cuadras, *Problemas de Probabilidades y Estadística*. Vol.2:Inferencia Estadística. EUB, 2000.
- [4] J.J. Faraway, *Linear Models with R*, Chapman & Hall/CRC, 2015.
- [5] D.J. Saville y G.R. Wood, *Statistical Methods: The Geometric Approach*, Cap. 17, New York:Springer, 1991.
- [6] G.W. Snedecor y W.G. Cochran, *Statistical Methods*, Iowa State University Press, 1989.
- [7] J. Verzani, *Using R for Introductory Statistics*. Chapman & Hall/CRC, 2004.