Ejercicios de Análisis de la Varianza con R

Francesc Carmona
Departament d'Estadística

30 de noviembre de 2006

1. Introducción

En este documento se resuelven algunos de los problemas del libro *Problemas de Probabilidades y Estadística* vol. 2 de C.M. Cuadras[2] con el programa estadístico R. Los enunciados de los problemas se encuentran en dicho libro.
Para profundizar en la teoría subyacente al Análisis de la Varianza se puede consultar, entre otros, el libro de *Modelos lineales*[1]. Si se quiere aprender R desde el principio o practicar su utilización en la Estadística elemental un buen libro es el de J. Verzani[5]. Para estudiar modelos lineales avanzados con R se puede leer el libro de J.J. Faraway[3].

2. Diseño de un factor

Problema 10.1
Se trata de una comparación entre tres poblaciones.
En primer lugar procedemos a leer los datos

```
> Iris.setosa <- c(5.1, 4.9, 4.7, 4.6, 5, 5.4, 4.6, 5, 4.4, 4.9,
+     5.4, 4.8, 4.8, 4.3, 5.8)
> Iris.versicolor <- c(7, 6.4, 6.9, 5.5, 6.5, 5.7, 6.3, 4.9, 6.6,
+     5.2, 5, 5.9, 6, 6.1, 5.6)
> Iris.virginica <- c(6.3, 5.8, 7.1, 6.3, 6.5, 7.6, 4.9, 7.3, 6.7,
+     7.2, 6.5, 6.4, 6.8, 5.7, 5.8)
```

Pero ésta no es la forma adecuada para trabajar con un programa estadístico. Mejor ponemos los datos en un único vector y añadimos una variable cualitativa o factor que nos indique la población de cada dato.

```
> longitud <- c(Iris.setosa, Iris.versicolor, Iris.virginica)
> especie <- rep(1:3, each = 15)
> especie <- factor(especie, labels = c("Iris setosa", "Iris versicolor",
+     "Iris virginica"))
```

En R es imprescindible definir el vector `especie` como un factor, ya que en caso contrario se podría confundir con un vector numérico.
Una única instrucción realiza los dos pasos

```
> especie <- gl(3, 15, labels = c("Iris setosa", "Iris versicolor",
+     "Iris virginica"))
```

Con la instrucción `split` podemos separar los datos

```
> split(longitud, especie)
```

Ahora podemos realizar un resumen de los datos y el gráfico que puede verse en la figura 1.

```
> tapply(longitud, especie, summary)
> plot(longitud ~ especie)
```

Asumiendo que la variable *longitud* sigue una distribución normal con varianza común para las tres poblaciones, la tabla del análisis de la varianza es

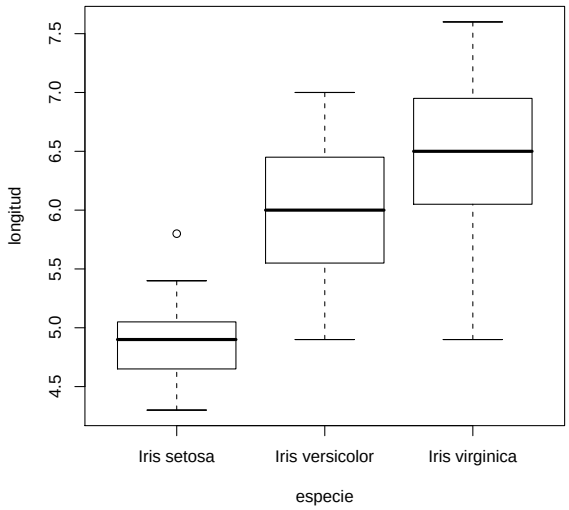

Figura 1: Gráficos de caja de las longitudes para las tres especies de flores

```
> p.aov <- aov(longitud ~ especie)
> summary(p.aov)

             Df  Sum Sq Mean Sq F value    Pr(>F)
especie       2 18.7631  9.3816  25.715 5.105e-08 ***
Residuals    42 15.3227  0.3648
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Otra posibilidad es definir el modelo lineal y obtener la tabla con la instrucción `anova`.

```
> g.lm <- lm(longitud ~ especie)
> anova(g.lm)
```

Como el p-valor es muy pequeño se concluye que hay diferencias muy significativas entre las tres especies. Las estimaciones de los parámetros se obtienen con

```
> model.tables(p.aov)

Tables of effects

 especie
especie
    Iris setosa Iris versicolor  Iris virginica
        -0.8689          0.1911          0.6778

> model.tables(p.aov, type = "mean")

Tables of means
Grand mean

5.782222

 especie
especie
    Iris setosa Iris versicolor  Iris virginica
          4.913           5.973           6.460
```

El modelo lineal contiene mucha información que se puede obtener con la instrucción `summary`.

```
> summary(g.lm)

Call:
lm(formula = longitud ~ especie)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56000 -0.31333 -0.01333  0.42667  1.14000

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.9133     0.1560  31.505  < 2e-16 ***
especieIris versicolor  1.0600     0.2206   4.806 1.99e-05 ***
especieIris virginica   1.5467     0.2206   7.013 1.39e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.604 on 42 degrees of freedom
Multiple R-Squared: 0.5505,        Adjusted R-squared: 0.5291
F-statistic: 25.72 on 2 and 42 DF,  p-value: 5.105e-08
```

Sin embargo, las estimaciones que se obtienen aquí corresponden al modelo con la restricción que el parámetro de la primera especie es cero. Esta es la opción por defecto para los modelos lineales en R. Se puede ver la matriz del diseño en esta situación:

```
> model.matrix(g.lm)
```

El error cuadrático medio o estimación insesgada de la varianza del modelo es

```
> ECM <- deviance(p.aov)/p.aov$df.residual
> ECM

[1] 0.3648254
```

Esta estimación también se obtiene directamente del modelo lineal

```
> summary(g.lm)$sigma^2

[1] 0.3648254
```

Además R contiene una base de datos con los famosos datos de Fisher o Anderson para 50 flores de cada una de las 3 especies.

```
> data(iris)
> help(iris)
```

Se puede repetir el análisis con los datos de la variable `Sepal.length` y el factor `Species`.

Problema 10.2
Se trata de un análisis de la varianza con un único factor tratamiento y cuatro niveles (P,A,B,AB). La introducción de los datos es la siguiente:

```
> P <- c(10, 0, 15, -20, 0, 15, -5, NA, NA, NA)
> A <- c(20, 25, 33, 25, 30, 18, 27, 0, 35, 20)
> B <- c(15, 10, 25, 30, 15, 35, 25, 22, 11, 25)
> AB <- c(10, 5, -5, 15, 20, 20, 0, 10, NA, NA)
> descenso <- c(P, A, B, AB)
> tratam <- gl(4, 10, labels = c("placebo", "fármaco A", "fármaco B",
+     "asociación AB"))
```

Suponiendo normalidad y homogeneidad de las varianzas, planteamos el test sobre la igualdad de medias. Un resumen numérico y gráfico se puede obtener con las instrucciones

```
> mean(descenso, na.rm = TRUE)
> tapply(descenso, tratam, summary)
> stripchart(descenso ~ tratam, method = "stack")
```

El modelo lineal y la tabla del análisis de la varianza son

```
> g.lm <- lm(descenso ~ tratam)
> anova(g.lm)

Analysis of Variance Table

Response: descenso
          Df  Sum Sq Mean Sq F value    Pr(>F)
tratam     3 2492.61  830.87  8.5262 0.0002823 ***
Residuals 31 3020.93   97.45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor es inferior al nivel de significación propuesto (0.01) de modo que rechazamos la hipótesis nula de igualdad de medias y admitimos que hay diferencias entre los fármacos.
Para ver si hay diferencias entre los fármacos A y B calcularemos el intervalo de confianza para la diferencia de medias:

```
> mediaA <- mean(descenso[tratam == "fármaco A"])
> mediaB <- mean(descenso[tratam == "fármaco B"])
> dif <- mediaA - mediaB
> ee.dif <- summary(g.lm)$sigma * sqrt(1/10 + 1/10)
> c(dif - qt(0.995, 31) * ee.dif, dif + qt(0.995, 31) * ee.dif)

[1] -10.11422  14.11422
```

Como este intervalo contiene al cero, podemos pensar que las diferencias entre A y B no son significativas. Aunque puede comprobarse que ambos fármacos difieren significativamente del placebo, cuando se realiza más de una comparación necesitamos un método de comparaciones múltiples. En caso contrario el error de tipo I global no estaría controlado. En R se puede aplicar el método de la *diferencia significativa honesta de Tukey* con la función `TukeyHSD`. Otros métodos de comparación dos a dos se pueden hallar en el paquete `multcomp`.

3. Diseño de dos factores

Problema 10.3
Se trata de un diseño de bloques aleatorizados (cada finca es un bloque).
Introducimos los datos con las instrucciones:

```
> produc <- c(2.1, 2.2, 1.8, 2, 1.9, 2.2, 2.6, 2.7, 2.5, 2.8, 1.8,
+     1.9, 1.6, 2, 1.9, 2.1, 2, 2.2, 2.4, 2.1)
> fert <- gl(4, 5)
> finca <- factor(rep(1:5, 4))
> xtabs(produc ~ finca + fert)

     fert
finca   1   2   3   4
    1 2.1 2.2 1.8 2.1
    2 2.2 2.6 1.9 2.0
    3 1.8 2.7 1.6 2.2
    4 2.0 2.5 2.0 2.4
    5 1.9 2.8 1.9 2.1
```

Ahora podemos generar un resumen de los datos y los gráficos de la figura 2.

```
> tapply(produc, fert, summary)
> tapply(produc, finca, summary)
> stripchart(produc ~ fert, method = "stack")
> stripchart(produc ~ finca, method = "stack")
> interaction.plot(fert, finca, produc, legend = F)
> interaction.plot(finca, fert, produc, legend = F)
```

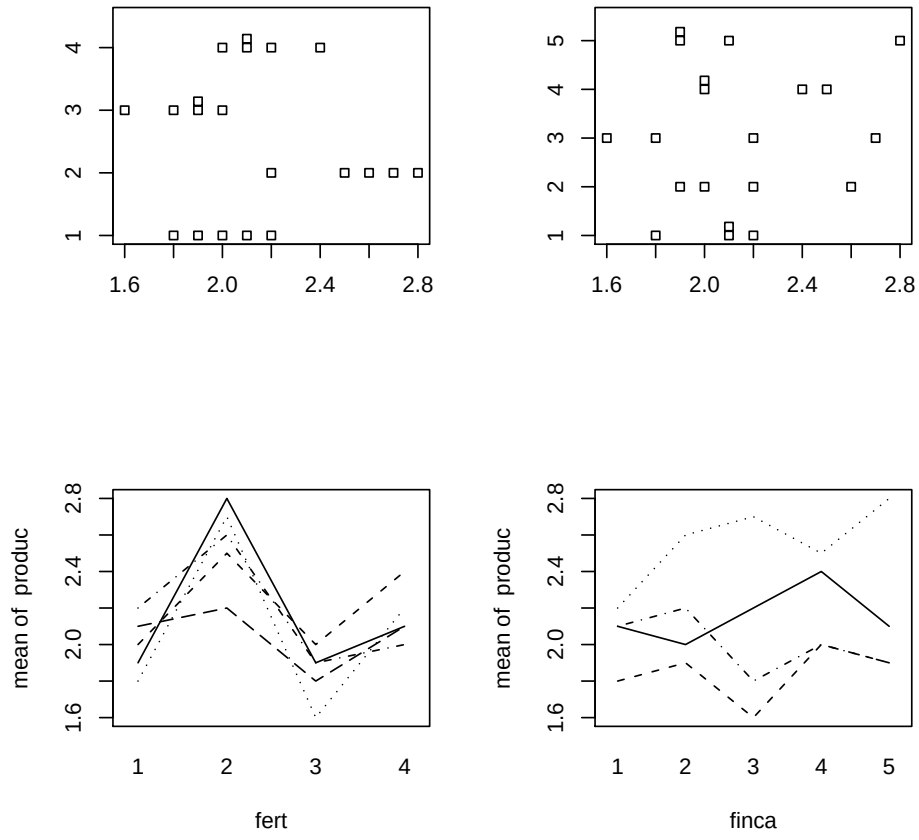

Figura 2: Gráficos de puntos y de interacciones con los datos de producción.

A la vista de los gráficos, concluimos que no hay datos atípicos, asimetrías o heterocedasticidad. Tampoco parece haber interacciones.
El modelo lineal y la tabla del análisis de la varianza son:

```
> g.lm <- lm(produc ~ finca + fert)
> anova(g.lm)

Analysis of Variance Table

Response: produc
          Df  Sum Sq Mean Sq F value    Pr(>F)
finca      4 0.08800 0.02200  0.6471 0.6395716
fert       3 1.43200 0.47733 14.0392 0.0003137 ***
Residuals 12 0.40800 0.03400
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No hay diferencias entre las fincas, pero sí las hay entre los fertilizantes.
Los efectos y las medias son:

```
> p.aov <- aov(produc ~ finca + fert)
> efectos <- model.tables(p.aov)
> efectos

Tables of effects

 finca
finca
     1      2      3      4      5
-0.090  0.035 -0.065  0.085  0.035

 fert
fert
    1     2     3     4
-0.14  0.42 -0.30  0.02

> medias <- model.tables(p.aov, type = "means")
> medias

Tables of means
Grand mean

2.14

 finca
finca
    1     2     3     4     5
2.050 2.175 2.075 2.225 2.175

 fert
fert
   1    2    3    4
2.00 2.56 1.84 2.16
```

En este caso el modelo es balanceado, de forma que el diseño es ortogonal y el orden de los factores en la instrucción `anova` no es importante. En este sentido hay que señalar que la tabla ANOVA de R corresponde a un contraste secuencial de modelos:

```
y ~ 1
y ~ finca
y ~ finca + fert
```

El primer p-valor corresponde a la comparación de los dos primeros modelos de la lista, mientras que el segundo p-valor corresponde a la comparación de los dos últimos. El denominador de ambos contrastes F es el error cuadrático medio del modelo completo, aquí 0,034.
Cuando el diseño no es ortogonal, por ejemplo si falta una observación, para contrastar el efecto del tratamiento es mejor el modelo con el efecto bloque en primer lugar. Una forma de contrastar todos los términos de un modelo respecto a dicho modelo completo es:

```
> drop1(g.lm, test = "F")

Single term deletions

Model:
produc ~ finca + fert
       Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>               0.408 -61.844
finca   4     0.088   0.496 -65.938  0.6471 0.6395716
fert    3     1.432   1.840 -37.719 14.0392 0.0003137 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problema 10.6
Los datos son:

```
> frec <- c(22, 21, 17, 20, 16, 21, 25, 19, 23, 31, 35, 35, 24,
+     18, 26, 25, 23, 23, 11, 16, 17, 24, 24, 20)
> mes <- factor(rep(1:4, each = 6), labels = c("Enero", "Marzo",
+     "Mayo", "Julio"))
> hora <- factor(rep(1:6, 4), labels = as.character(9:14))
```

Las medias son

```
> tapply(frec, mes, mean)

   Enero    Marzo     Mayo    Julio
19.50000 28.00000 23.16667 18.66667

> tapply(frec, hora, mean)

    9    10    11    12    13    14
20.50 18.50 20.75 25.00 24.50 24.75
```

La tabla del análisis de la varianza es

```
> p.aov <- aov(frec ~ hora + mes)
> summary(p.aov)

            Df Sum Sq Mean Sq F value   Pr(>F)
hora         5 152.83   30.57  1.7313 0.188155
mes          3 325.67  108.56  6.1485 0.006155 **
Residuals   15 264.83   17.66
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No existen diferencias entre las horas.
Existen diferencias entre los meses.

4. Diseño de dos factores con interacción

Problema 10.4
Vamos a introducir los datos:

```
> huevos <- c(93, 94, 93, 90, 93, 86, 95.5, 83.5, 92, 92.5, 82,
+     82.5, 92, 91, 90, 95, 84, 78, 83.3, 87.6, 81.9, 80.1, 79.6,
+     49.4, 84, 84.4, 77, 67, 69.1, 88.4, 85.3, 89.4, 85.4, 87.4,
+     52, 77)
> genotipo <- rep(rep(1:3, each = 6), 2)
> siembra <- rep(1:2, each = 18)
> genotipo <- factor(genotipo, labels = c("++", "+-", "--"))
> siembra <- factor(siembra, labels = c("100", "800"))
```

El número de huevos eclosionados por casilla sigue la distribución binomial con $n = 100$ o $n = 800$. Para normalizar la muestra se aplica la transformación

```
> y <- asin(sqrt(huevos/100))
> y <- y * 180/pi
```

de donde resulta la tabla:

```
> split(round(y, 2), genotipo)
```

```
$`++`
 [1] 74.66 75.82 74.66 71.57 74.66 68.03 65.88 69.38 64.82 63.51 63.15 44.66

$`+-`
 [1] 77.75 66.03 73.57 74.11 64.90 65.27 66.42 66.74 61.34 54.94 56.23 70.09

$`--`
 [1] 73.57 72.54 71.57 77.08 66.42 62.03 67.46 71.00 67.54 69.21 46.15 61.34
```

Aunque no es absolutamente necesario, vamos a poner los datos en forma de `data.frame` o base de datos de R.

```
> problema <- data.frame(y, siembra, genotipo)
> rm(y, siembra, genotipo)
> attach(problema)
```

Algunos de los siguientes gráficos pueden verse en la figura 3.

```
> boxplot(y ~ siembra)
> boxplot(y ~ genotipo)
> plot.design(problema, fun = "mean")
> plot.design(problema, fun = "median")
> interaction.plot(genotipo, siembra, y)
> interaction.plot(siembra, genotipo, y)
```

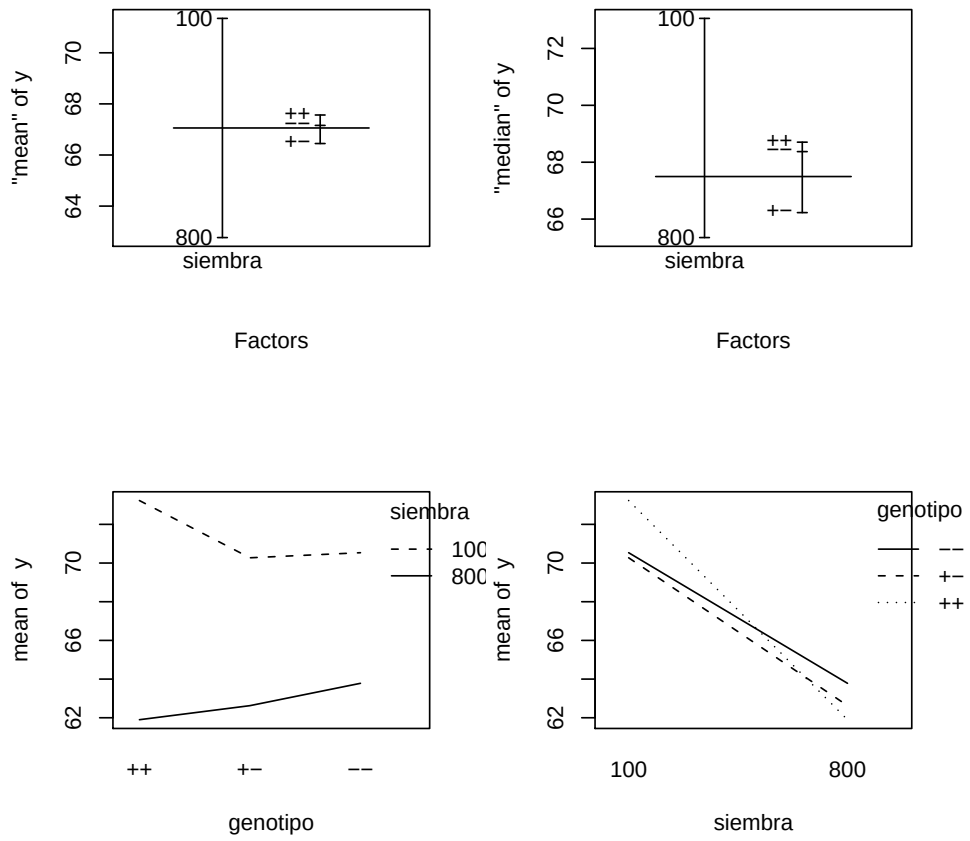

Figura 3: Gráficos de medias, medianas e interacciones con los datos transformados del problema de los huevos.

La tabla del análisis de la varianza para un diseño de dos factores con interacción es

```
> p.aov <- aov(y ~ siembra * genotipo, data = problema)
> summary(p.aov)

                Df  Sum Sq Mean Sq F value    Pr(>F)
siembra          1  662.09  662.09 14.8329 0.0005736 ***
genotipo         2    7.66    3.83  0.0859 0.9179521
siembra:genotipo 2   35.35   17.68  0.3960 0.6764562
Residuals       30 1339.09   44.64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aunque las sumas de cuadrados son ligeramente distintas a las del libro de Cuadras, por la mayor precisión utilizada, los resultados son evidentemente los mismos. No es significativa la diferencia entre los genotipos ni la interacción, pero sí existen diferencias significativas sembrando 100 u 800 huevos, siendo el porcentaje de eclosiones mayor en el primer caso (al haber menos huevos, las larvas disponen de más alimento).

```
> medias <- model.tables(p.aov, type = "means")
> medias$tables$siembra

siembra
     100      800
71.34575 62.76873

> detach(problema)
```

5. Diseño de cuadrados latinos

Problema 10.5
Vamos a introducir los datos:

```
> produc <- c(12, 17, 24, 12, 18, 22, 14, 15, 15, 13, 20, 31, 20,
+     14, 12, 18)
> fila <- factor(rep(1:4, 4))
> columna <- factor(rep(1:4, each = 4))
> variedad <- c("A", "C", "D", "B", "B", "D", "C", "A", "C", "A",
+     "B", "D", "D", "B", "A", "C")
> problema <- data.frame(fila, columna, variedad, produc)
> rm(fila, columna, variedad, produc)
> attach(problema)
```

Efectivamente, se trata de un diseño de cuadrados latinos:

```
> matrix(problema$variedad, 4, 4)

     [,1] [,2] [,3] [,4]
[1,] "A"  "B"  "C"  "D"
[2,] "C"  "D"  "A"  "B"
[3,] "D"  "C"  "B"  "A"
[4,] "B"  "A"  "D"  "C"
```

La tabla del análisis de la varianza para este diseño es

```
> p.aov <- aov(produc ~ fila + columna + variedad, data = problema)
> summary(p.aov)

            Df  Sum Sq Mean Sq F value  Pr(>F)
fila         3  18.688   6.229  0.5273 0.67964
columna      3  35.188  11.729  0.9929 0.45737
variedad     3 280.688  93.563  7.9206 0.01651 *
Residuals    6  70.875  11.813
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

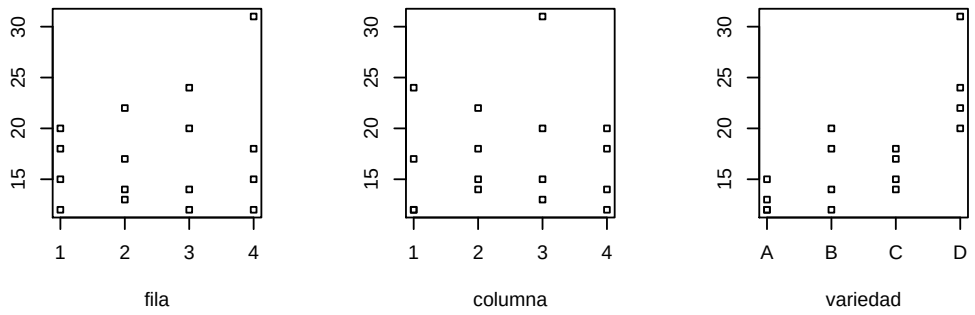

Figura 4: Gráficos de puntos con los datos de producción del problema 10.5.

```
> detach(problema)
```

No hay diferencias significativas entre filas ni entre columnas. En cambio sí hay diferencias entre variedades.

6. Diseño multifactorial

Problema 11.5
Existen dos causas de variabilidad (tiempo y dosis). Como además los individuos son los mismos en cada casilla, debemos añadir un efecto bloque que recoja los efectos individuales. Admitiremos que los datos se ajustan a un diseño de dos factores con interacción en bloques aleatorizados.
Los datos son

```
> glucemia <- c(82, 83, 85, 75, 81, 88, 87, 91, 79, 85, 83, 85,
+     85, 79, 81, 86, 87, 90, 80, 83, 90, 91, 94, 83, 88, 96, 97,
+     99, 88, 93, 108, 109, 112, 89, 103, 110, 110, 117, 90, 109,
+     118, 120, 125, 119, 114)
> tiempo <- factor(rep(rep(1:3, each = 5), 3), labels = c("0'",
+     "15'", "30'"))
> dosis <- factor(rep(1:3, each = 15), labels = c("0 mg", "5 mg",
+     "10 mg"))
> bloque <- factor(rep(1:5, 9))
```

La tabla del análisis de la varianza para este diseño es

```
> p.aov <- aov(glucemia ~ tiempo + dosis + tiempo:dosis + bloque)
> summary(p.aov)

             Df Sum Sq Mean Sq  F value    Pr(>F)
tiempo        2  556.0   278.0  27.3502 1.186e-07 ***
dosis         2 5939.9  2970.0 292.1667 < 2.2e-16 ***
bloque        4  841.9   210.5  20.7056 1.661e-08 ***
tiempo:dosis  4  357.4    89.4   8.7903 6.615e-05 ***
Residuals    32  325.3    10.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observemos que el residuo queda determinado por la definición del modelo en la fórmula.
Tanto el factor tiempo, como el factor dosis y la interacción son muy significativos. Incluso es también significativo el efecto bloque. La principal causa de variabilidad es el factor dosis.
En este caso, una posibilidad interesante sería explotar el hecho de que los factores tienen escala ordinal como en el ejemplo de la sección 15.2 de [3].

Problema 11.6
Se trata de un diseño con 3 factores a 3,2 y 3 niveles, con 3 réplicas por casilla. Los datos son

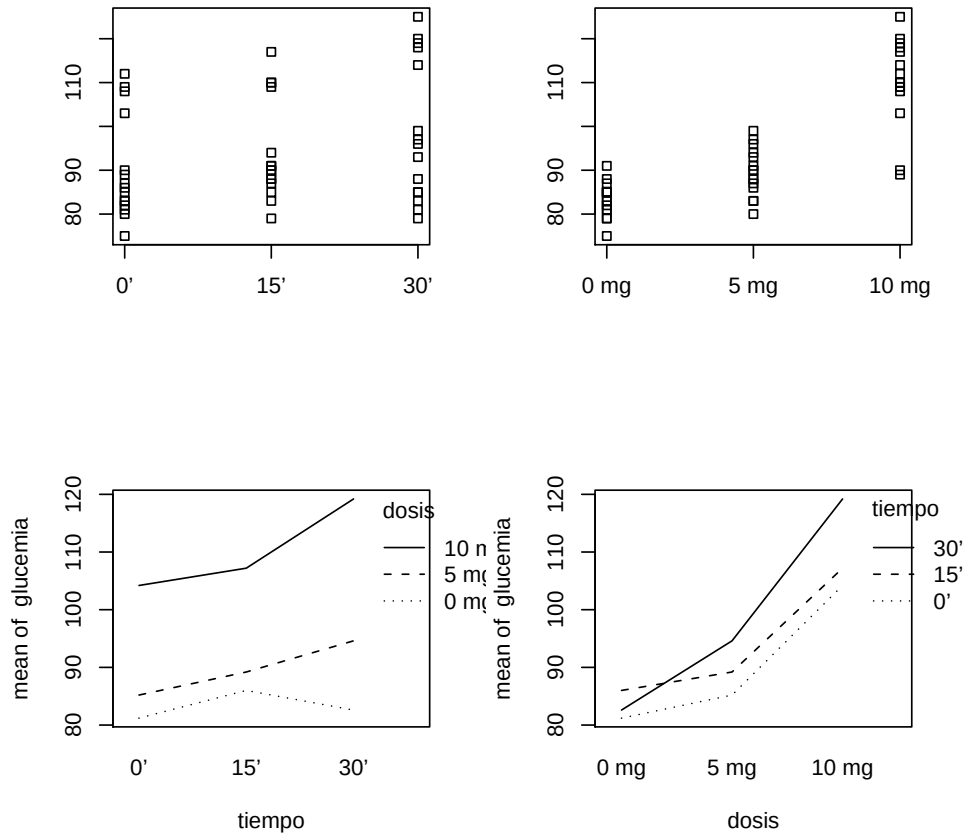

Figura 5: Gráficos de puntos y de interacción con los datos de glucemia del problema 11.5.

```
> horas <- c(7.3, 7.5, 7.1, 7.1, 7.3, 6.9, 8.1, 8.2, 8, 7.6, 7.4,
+     7.2, 6.8, 7.3, 7.2, 8.3, 8.2, 8.1, 8.5, 8.3, 8.4, 7.5, 7.2,
+     7.2, 8.9, 8.4, 8.1, 8.3, 8.7, 7.9, 7.6, 7.4, 7.2, 9, 8.5,
+     8, 6.7, 6.5, 6.3, 6.7, 6.3, 6.2, 6.8, 6.2, 6.2, 6.1, 6.2,
+     6.9, 6.4, 6.9, 6.8, 6, 6.1, 6.2)
> tratam <- factor(rep(1:3, each = 18), labels = c("A1", "A2",
+     "A3"))
> sexo <- factor(rep(rep(1:2, each = 9), 3), labels = c("B1", "B2"))
> forma <- factor(rep(rep(1:3, each = 3), 6), labels = c("C1",
+     "C2", "C3"))
```

La tabla del análisis de la varianza para este diseño es

```
> p.aov <- aov(horas ~ tratam * sexo * forma)
> summary(p.aov)
```

```
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
tratam             2 25.3781 12.6891 167.5330 < 2.2e-16 ***
sexo               1  0.0030  0.0030   0.0391    0.8443
forma              2  3.6048  1.8024  23.7971 2.596e-07 ***
tratam:sexo        2  0.0226  0.0113   0.1491    0.8620
tratam:forma       4  4.8896  1.2224  16.1394 1.172e-07 ***
sexo:forma         2  0.0959  0.0480   0.6333    0.5367
tratam:sexo:forma  4  0.2252  0.0563   0.7433    0.5689
Residuals         36  2.7267  0.0757
```

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

Son significativos los efectos principales tratamiento y forma, y su interacción. No es significativo el efecto sexo, ni las demás interacciones.
Aunque algunos autores discrepan, Cuadras añade los efectos no significativos al residuo y calcula de nuevo la tabla del análisis de la varianza para el diseño reducido:

```
> p.aov <- aov(horas ~ tratam * forma)
> summary(p.aov)

             Df  Sum Sq Mean Sq F value    Pr(>F)
tratam        2 25.3781 12.6891 185.794 < 2.2e-16 ***
forma         2  3.6048  1.8024  26.391 2.609e-08 ***
tratam:forma  4  4.8896  1.2224  17.899 7.375e-09 ***
Residuals    45  3.0733  0.0683
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como la interacción es significativa, no deberíamos contrastar los efectos principales. La estimación de los efectos principales y su significación dependen de la codificación cuando las interacciones están presentes en el modelo.

7. Análisis de la covarianza

Problema 12.4
Se trata de un diseño unifactorial con una variable concomitante.
En primer lugar introducimos los datos:

```
> tiempo <- c(570, 710, 630, 633, 640, 552, 620, 585, 593, 710,
+     698, 560, 842, 940, 898, 730, 872, 855)
> neurot <- c(12, 26, 18, 19, 19, 10, 19, 19, 6, 18, 20, 3, 18,
+     29, 25, 8, 22, 18)
> farmaco <- rep(c("A", "B", "C"), each = 6)
> datos <- data.frame(tiempo, neurot, farmaco)
> rm(tiempo, neurot, farmaco)
> attach(datos)
```

En el `data.frame` el vector `farmaco` (no es numérico) se convierte automáticamente en un `factor`.
Un resumen de los datos se puede obtener con la instrucción

```
> by(datos, farmaco, summary)

farmaco: A
     tiempo          neurot      farmaco
 Min.   :552.0   Min.   :10.00   A:6
 1st Qu.:585.0   1st Qu.:13.50   B:0
 Median :631.5   Median :18.50   C:0
 Mean   :622.5   Mean   :17.33
 3rd Qu.:638.3   3rd Qu.:19.00
 Max.   :710.0   Max.   :26.00
------------------------------------------------------------
farmaco: B
     tiempo          neurot      farmaco
 Min.   :560.0   Min.   : 3.00   A:0
 1st Qu.:587.0   1st Qu.: 9.00   B:6
 Median :606.5   Median :18.50   C:0
 Mean   :627.7   Mean   :14.17
 3rd Qu.:678.5   3rd Qu.:19.00
 Max.   :710.0   Max.   :20.00
```

```
------------------------------------------------------------
farmaco: C
     tiempo          neurot      farmaco
 Min.   :730.0   Min.   : 8.00   A:0
 1st Qu.:845.3   1st Qu.:18.00   B:0
 Median :863.5   Median :20.00   C:6
 Mean   :856.2   Mean   :20.00
 3rd Qu.:891.5   3rd Qu.:24.25
 Max.   :940.0   Max.   :29.00
```

Si prescindimos de la información que el neuroticismo puede influir en el tiempo de frenado, el modelo lineal es

```
> g0 <- lm(tiempo ~ farmaco, datos)
> anova(g0)

Analysis of Variance Table

Response: tiempo
          Df Sum Sq Mean Sq F value    Pr(>F)
farmaco    2 213678  106839  26.537 1.184e-05 ***
Residuals 15  60390    4026
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pero en el segundo gráfico de la figura 6 se observa la influencia de la variable concomitante:

```
> plot(tiempo ~ neurot, pch = as.character(farmaco), datos)
```

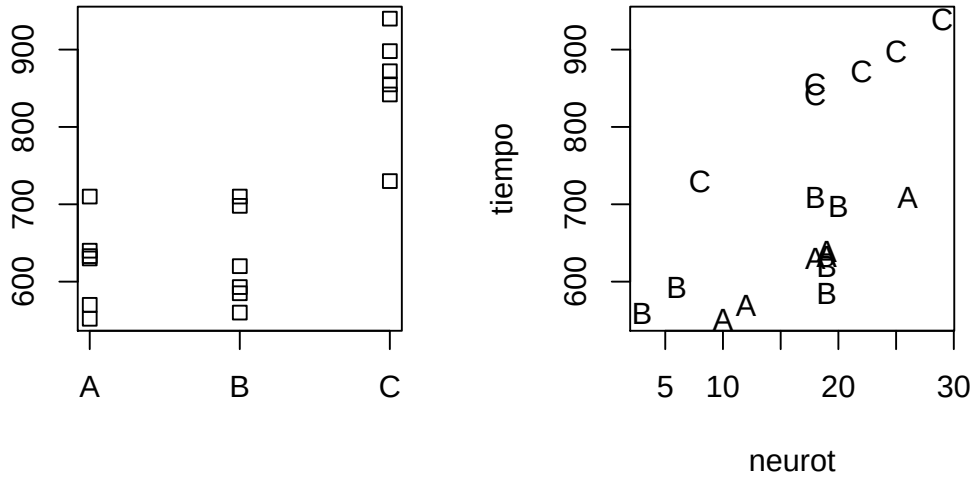

Figura 6: Gráficos de puntos y de dispersión con los datos del problema 12.4.

Se comprueba que la interacción `fármaco:neurot` no es significativa:

```
> g1 <- lm(tiempo ~ neurot + farmaco + neurot:farmaco, datos)
> model.matrix(g1)

   (Intercept) neurot farmacoB farmacoC neurot:farmacoB neurot:farmacoC
1            1     12        0        0               0               0
2            1     26        0        0               0               0
3            1     18        0        0               0               0
4            1     19        0        0               0               0
5            1     19        0        0               0               0
6            1     10        0        0               0               0
```

```
7            1     19        1        0              19              0
8            1     19        1        0              19              0
9            1      6        1        0               6              0
10           1     18        1        0              18              0
11           1     20        1        0              20              0
12           1      3        1        0               3              0
13           1     18        0        1               0             18
14           1     29        0        1               0             29
15           1     25        0        1               0             25
16           1      8        0        1               0              8
17           1     22        0        1               0             22
18           1     18        0        1               0             18
attr(,"assign")
[1] 0 1 2 2 3 3
attr(,"contrasts")
attr(,"contrasts")$farmaco
[1] "contr.treatment"
```

```
> anova(g1)

Analysis of Variance Table

Response: tiempo
              Df Sum Sq Mean Sq  F value    Pr(>F)
neurot         1 111919  111919 112.9986 1.843e-07 ***
farmaco        2 146875   73438  74.1460 1.760e-07 ***
neurot:farmaco 2   3388    1694   1.7106     0.222
Residuals     12  11885     990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Así que consideramos el modelo con la variable concomitante

```
> g <- lm(tiempo ~ neurot + farmaco, datos)
> summary(g)

Call:
lm(formula = tiempo ~ neurot + farmaco, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-81.137  -8.229   2.115  17.578  51.823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  484.539     25.339  19.122 1.97e-11 ***
neurot         7.959      1.238   6.431 1.57e-05 ***
farmacoB      30.371     19.469   1.560    0.141
farmacoC     212.442     19.353  10.977 2.91e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.03 on 14 degrees of freedom
Multiple R-Squared: 0.9443,        Adjusted R-squared: 0.9323
F-statistic: 79.07 on 3 and 14 DF,  p-value: 5.117e-09

> sqrt(41.35)

[1] 6.430397
```

La estimación del parámetro de regresión γ es 7.959 y su significación resulta muy clara.
El `summary` de un `anova` contrasta los modelos secuencialmente, para ver la diferencia entre fármacos (eliminando la influencia del neuroticismo) debemos utilizar la instrucción `drop1`:

```
> drop1(g, test = "F")

Single term deletions

Model:
tiempo ~ neurot + farmaco
        Df Sum of Sq    RSS    AIC F value     Pr(F)
<none>                15274    129
neurot   1     45116  60390    152  41.353 1.572e-05 ***
farmaco  2    146875 162149    168  67.313 6.580e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observemos que este resultado contrasta la significación de la variable concomitante (dada la presencia del factor) y la del factor (con la presencia de la variable concomitante). La diferencia entre fármacos es muy significativa.

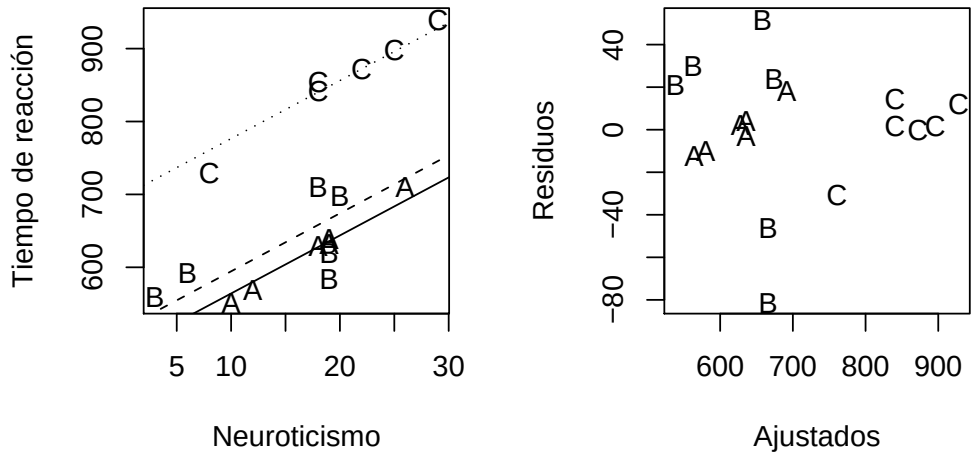

Figura 7: Gráfico con las rectas ajustadas y gráfico de residuos con los datos del problema 12.4.

Con las estimaciones obtenidas por el `summary(g)` podemos dibujar las rectas de regresión del gráfico 7.

```
> plot(tiempo ~ neurot, pch = as.character(farmaco), xlab = "Neuroticismo",
+     ylab = "Tiempo de reacción")
> abline(484.539, 7.959)
> abline(484.539 + 30.371, 7.959, lty = 2)
> abline(484.539 + 212.442, 7.959, lty = 3)
> plot(fitted(g), residuals(g), pch = as.character(farmaco), xlab = "Ajustados",
+     ylab = "Residuos")
> detach(datos)
```

Problema 12.6
En primer lugar introducimos los datos del peso inicial y el engorde semanal de cerdos, clasificados por sexos y corrales, con tres tipos de alimentación.

```
> engorde <- c(9.94, 9.52, 9.48, 8.21, 9.32, 9.32, 10.98, 10.56,
+     8.82, 10.42, 10, 8.51, 9.24, 9.95, 9.34, 8.43, 9.68, 8.86,
+     9.67, 9.2, 9.75, 9.11, 8.66, 8.5, 7.63, 8.9, 10.37, 9.51,
+     8.57, 8.76)
> peso <- c(48, 38, 32, 35, 35, 41, 46, 48, 32, 43, 48, 39, 32,
+     38, 41, 46, 46, 40, 37, 40, 48, 48, 28, 37, 33, 42, 50, 42,
```

```
+     30, 40)
> sexo <- rep(c("M", "H"), 15)
> corral <- factor(rep(rep(1:5, each = 2), 3))
> aliment <- rep(c("A", "B", "C"), each = 10)
> datos <- data.frame(engorde, peso, sexo, corral, aliment)
> rm(engorde, peso, sexo, corral, aliment)
> attach(datos)
```

El peso inicial es la variable concomitante y sólo se toma la interacción del tipo de alimentación con el sexo y se ignoran las demás interacciones.
Si inicialmente prescindimos de la variable concomitante, el modelo es

```
> g <- lm(engorde ~ aliment + corral + sexo + aliment:sexo, datos)
> anova(g)

Analysis of Variance Table

Response: engorde
             Df Sum Sq Mean Sq F value  Pr(>F)
aliment       2 2.3242  1.1621  2.7657 0.08701 .
corral        4 4.9607  1.2402  2.9515 0.04554 *
sexo          1 0.4539  0.4539  1.0802 0.31107
aliment:sexo  2 0.4642  0.2321  0.5523 0.58413
Residuals    20 8.4038  0.4202
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

donde se observa que no hay diferencias entre tipos de alimentación ni entre sexos. Tampoco es significativa la interacción `aliment:sexo`. Sin embargo, hay diferencias entre corrales. La matriz de diseño del modelo se obtiene con la instrucción `model.matrix(g)`.
La estimación y el contraste del parámetro de regresión de la variable peso se consigue así:

```
> gp <- lm(engorde ~ peso + aliment + corral + sexo + aliment:sexo,
+     datos)
> summary(gp)

Call:
lm(formula = engorde ~ peso + aliment + corral + sexo + aliment:sexo,
    data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13435 -0.28296  0.07649  0.25202  0.91814

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.68812    1.13880   4.995 8.04e-05 ***
peso           0.08927    0.02407   3.709  0.00149 **
alimentB      -0.58029    0.32048  -1.811  0.08603 .
alimentC      -0.72141    0.32091  -2.248  0.03664 *
corral2        0.53183    0.39715   1.339  0.19634
corral3       -0.18711    0.31776  -0.589  0.56290
corral4        0.47703    0.29267   1.630  0.11958
corral5        0.46760    0.34792   1.344  0.19478
sexoM          0.31624    0.32550   0.972  0.34347
alimentB:sexoM  0.26190    0.45732   0.573  0.57357
alimentC:sexoM  0.08083    0.45465   0.178  0.86078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.5065 on 19 degrees of freedom
Multiple R-Squared: 0.7065,        Adjusted R-squared: 0.552
F-statistic: 4.574 on 10 and 19 DF,  p-value: 0.002179
```

```
> sqrt(13.76)

[1] 3.709447
```

La estimación es $\hat{\gamma}$ =0.08927 y el estadístico de contraste t =3.709 (significativo, p-valor=0.00149).
Si no hay diferencias entre los tipos de alimentación el modelo lineal es:

```
> ga <- lm(engorde ~ peso + corral + sexo, datos)
> summary(ga)

Call:
lm(formula = engorde ~ peso + corral + sexo, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80818 -0.38845 -0.04262  0.35904  0.88094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.21131    1.22237   4.263 0.000292 ***
peso         0.09020    0.02607   3.460 0.002127 **
corral2      0.54228    0.43756   1.239 0.227724
corral3     -0.18228    0.35332  -0.516 0.610845
corral4      0.47656    0.32689   1.458 0.158396
corral5      0.47493    0.38523   1.233 0.230085
sexoM        0.43242    0.21349   2.025 0.054575 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5657 on 23 degrees of freedom
Multiple R-Squared: 0.5567,        Adjusted R-squared: 0.4411
F-statistic: 4.814 on 6 and 23 DF,  p-value: 0.002566
```

De modo que la estimación del parámetro de regresión para el peso es 0.0902 y es significativo. El contraste de los modelos es

```
> anova(ga, gp)

Analysis of Variance Table

Model 1: engorde ~ peso + corral + sexo
Model 2: engorde ~ peso + aliment + corral + sexo + aliment:sexo
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1     23 7.3614
2     19 4.8741  4    2.4873 2.424 0.08378 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esta tabla nos dice que no hay diferencias entre alimentaciones.

Sin embargo, en el libro de Cuadras[2] se propone considerar las interacciones entre alimentos y sexo de modo que el modelo lineal es un poco más complejo.

```
> aliAsexoM <- c(rep(c(1,-1), 5), rep(0, 10), rep(c(-1,1), 5))
> aliBsexoH <- c(rep(0, 10), rep(c(-1,1), 5), rep(c(1,-1), 5))
> ga0 <- lm(engorde ~ peso + corral + sexo + aliAsexoM + aliBsexoH)
> summary(ga0)
```

```
Call:
lm(formula = engorde ~ peso + corral + sexo + aliAsexoM + aliBsexoH)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80354 -0.34457 -0.02659  0.38268  0.85203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.78574    1.11954   5.168 4.04e-05 ***
peso         0.08781    0.02779   3.160  0.00472 **
corral1     -0.25098    0.25193  -0.996  0.33048
corral2      0.26459    0.27951   0.947  0.35460
corral3     -0.44561    0.21520  -2.071  0.05090 .
corral4      0.22678    0.25945   0.874  0.39196
sexo1       -0.21374    0.11120  -1.922  0.06827 .
aliAsexoM   -0.05736    0.15200  -0.377  0.70966
aliBsexoH   -0.07548    0.15516  -0.486  0.63167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5884 on 21 degrees of freedom
Multiple R-Squared: 0.5622,     Adjusted R-squared: 0.3954
F-statistic:  3.37 on 8 and 21 DF,  p-value: 0.01211
```

```
> anova(ga0, gp)

Analysis of Variance Table

Model 1: engorde ~ peso + corral + sexo + aliAsexoM + aliBsexoH
Model 2: engorde ~ peso + aliment + corral + sexo + aliment:sexo
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     21 7.2710
2     19 4.8741  2    2.3968 4.6716 0.02238 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora la estimación del parámetro de regresión para el peso es 0.08781 y también es significativo. La tabla del análisis de la varianza nos da un estadístico F =4.6716 que es significativo (p-valor=0.02238), de modo que hay diferencias entre los tipos de alimentación. Estos resultados coinciden con los del libro de Cuadras[2] y con los de Wishart(1938) y Rao(1965).

Análogamente, si no hay diferencias entre los corrales el modelo es

```
> gb <- lm(engorde ~ peso + aliment + sexo + aliment:sexo, datos)
> summary(gb)

Call:
lm(formula = engorde ~ peso + aliment + sexo + aliment:sexo,
    data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00049 -0.29798 -0.06508  0.36105  1.15798

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.48393    0.75555   8.582 1.26e-08 ***
peso           0.07615    0.01738   4.382 0.000217 ***
alimentB      -0.58554    0.35596  -1.645 0.113581
alimentC      -0.71092    0.35617  -1.996 0.057905 .
```

```
sexoM           0.28476    0.35833   0.795 0.434929
alimentB:sexoM  0.29602    0.50534   0.586 0.563729
alimentC:sexoM  0.05984    0.50408   0.119 0.906540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5627 on 23 degrees of freedom
Multiple R-Squared: 0.5614,        Adjusted R-squared: 0.447
F-statistic: 4.907 on 6 and 23 DF,  p-value: 0.002302
```

```
> anova(gb, gp)

Analysis of Variance Table

Model 1: engorde ~ peso + aliment + sexo + aliment:sexo
Model 2: engorde ~ peso + aliment + corral + sexo + aliment:sexo
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     23 7.2831
2     19 4.8741  4    2.4089 2.3476 0.09126 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La estimación del parámetro de regresión para el peso es 0.07615 y también es significativo. La tabla del análisis de la varianza nos da un estadístico F =2.3476 que no es significativo (p-valor=0.09126).

La varianza estimada de los errores en el modelo sin variable concomitante es

```
> summary(g)$sigma^2

[1] 0.4201907
```

En cambio, la varianza estimada de los errores en el modelo con variable concomitante es

```
> summary(gp)$sigma^2

[1] 0.2565335
```

Casi la mitad de la anterior.

8. Análisis de los residuos

En todos los modelos deberíamos hacer un diagnóstico mediante un análisis de los residuos más o menos sofisticado.
En general y como mínimo, un par de gráficos nos pueden servir. Por ejemplo, en el modelo lineal del problema 10.1 podemos representar los residuos como se puede ver en la figura 8.

```
> g.lm <- lm(longitud ~ especie)
> plot(especie, residuals(g.lm), ylab = "residuos")
> abline(h = 0)
> qqnorm(residuals(g.lm))
> qqline(residuals(g.lm))
```

Para contrastar la igualdad de las varianzas en las tres especies podemos realizar el contraste de Levene.

```
> y <- longitud
> med <- tapply(y, especie, median)
> med

    Iris setosa Iris versicolor  Iris virginica
            4.9             6.0             6.5

> aresid <- abs(y - med[especie])
> anova(lm(aresid ~ especie))
```

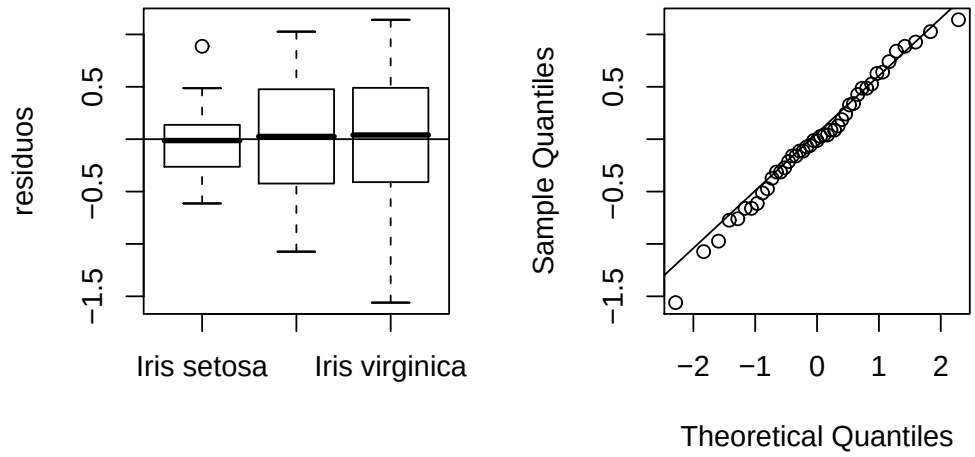

Figura 8: Gráficos de diagnóstico con los datos del problema 10.1.

```
Analysis of Variance Table

Response: aresid
          Df Sum Sq Mean Sq F value Pr(>F)
especie    2 0.5760  0.2880   2.185 0.1251
Residuals 42 5.5360  0.1318
```

Se considera que hay heterocedasticidad si el p-valor es menor que 0,01. En este caso no hay razón para dudar de la homocedasticidad.

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, 2004.

[4] P. Murrell, *R Graphics*, Chapman & Hall/CRC, 2005.

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