Modelos Estadísticos

Javier Morales (Universidad Miguel Hernández de Elche)
Mª Asunción Martínez (Universidad Miguel Hernández de Elche)

2020-05-02

Contents

Introducción

```
library(tidyverse)
library(tidymodels)
library(stringr)
library(forcats)
library(lubridate)
library(magrittr)
library(broom)
library(datasets)
library(lmtest)
library(MASS)
library(kableExtra)
library(mosaic)
library(latex2exp)
library(pubh)
library(moonBook)
library(sjlabelled)
library(sjPlot)
library(nasaweather)
library(alr3)
library(reshape2)
library(olsrr)
library(ggfortify)
library(equatiomatic) # necesario solo para escritura markdown
```

Para instalar los paquetes necesarios para poder ejecutar la sintáxis mostrada en el libro se puede emplear el siguiente comando:

```
pkgs <- c('tidyverse', 'stringr', 'forcats', 'lubridate',
          'magrittr', 'broom', 'datasets', 'kableExtra',
          'mosaic', 'latex2exp', 'pubh', 'moonBook',
          'sjlabelled', 'sjPlot', 'nasaweather', 'tidymodels',
          'lmtest','alr3', 'reshape2', 'olsrr','ggfortify')
install.packages(setdiff(pkgs, installed.packages()[,"Package"]),
                 dependencies = TRUE)
```

Chapter 1

Modelos de Análisis de la covarianza (ANCOVA)

En este tema se presentan los modelos ANCOVA o modelos de análisis de la covarianza. Esto modelos surgen cuando entre las posibles variables predictoras de la respuesta (de tipo numérico) consideramos tanto variables numéricas como factores. El objetivo principal de este tipo de modelos es estudiar si la relación entre las predictoras numéricas y la respuesta viene condicionada por el factor o factores de clasificación considerados, es decir, si debemos construir:

- un único modelo entre la respuesta y las predictoras de tipo numérico,
- un único modelo entre la respuesta y las predictoras de tipo categórico (factores),
- un modelo diferente entre la respuesta y las predictoras numéricas para cada nivel o combinaciones de niveles de los factores.

Este tipo de modelos permiten una versatilidad que nos posibilita el estudio de situaciones experimentales más complejas. Sin embargo, no están exentos de dificultades sobre todo en lo que tiene que ver con el cumplimiento de las hipótesis del modelo. En función del modelo final las hipótesis de normalidad y homogeneidad varían en su aplicación. Además, el proceso de selección del mejor modelo requiere de un proceso de análisis más profundo debido a la inclusión de diferentes tipos de variables en el conjunto de posibles predictoras.

Para introducir los conceptos básicos de este tipo de modelos y mostrar todas sus posibilidades de análisis comenzaremos con el modelo ANCOVA más sencillo donde únicamente consideramos dos variables predictoras, una de tipo numérico y la otra un factor. La formulación presentada se puede generalizar rápidamente a situaciones más complejas donde el número de predictoras sea mayor.

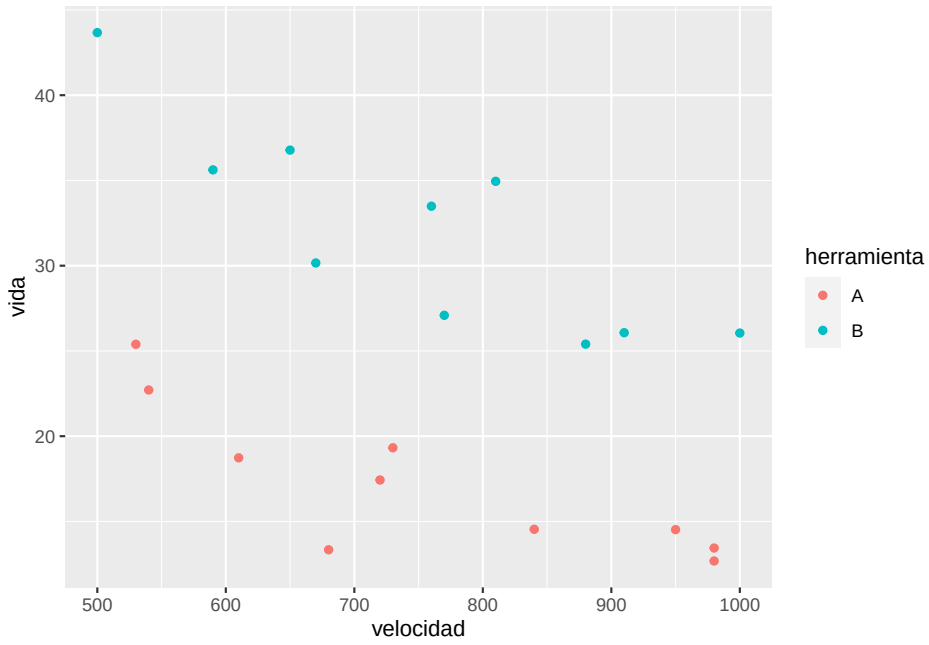
1.1 Bancos de datos

Veamos los diferentes bancos de datos que iremos analizando a lo largo de la unidad.

Ejemplo 1.1 (Tiempo de vida)**.** Se desea estudiar el tiempo de vida de una pieza (`vida`) cortadora de dos tipos, A y B (`herramienta`), en función de la velocidad del torno (`velocidad`) en el que está integrada (en revoluciones por segundo). El objetivo del análisis es describir la relación entre el tiempo de vida de la pieza y la velocidad del torno, teniendo en cuenta de qué tipo es la pieza.

Cargamos los datos y realizamos el gráfico descriptivo:

```
# Carga de datos
velocidad <- c(610,950,720,840,980,530,680,540,980,730,670,
               770,880,1000,760,590,910,650,810,500)
vida <- c(18.73,14.52,17.43,14.54,13.44,25.39,13.34,22.71,
          12.68,19.32,30.16,27.09,25.40,26.05,33.49,35.62,
          26.07,36.78,34.95,43.67)
herramienta <- gl(2,10,20,labels=c("A","B"))
tiempovida <- data.frame(velocidad,vida,herramienta)
# Gráfico
ggplot(tiempovida, aes(x = velocidad, y = vida, color = herramienta)) +
  geom_point()
```



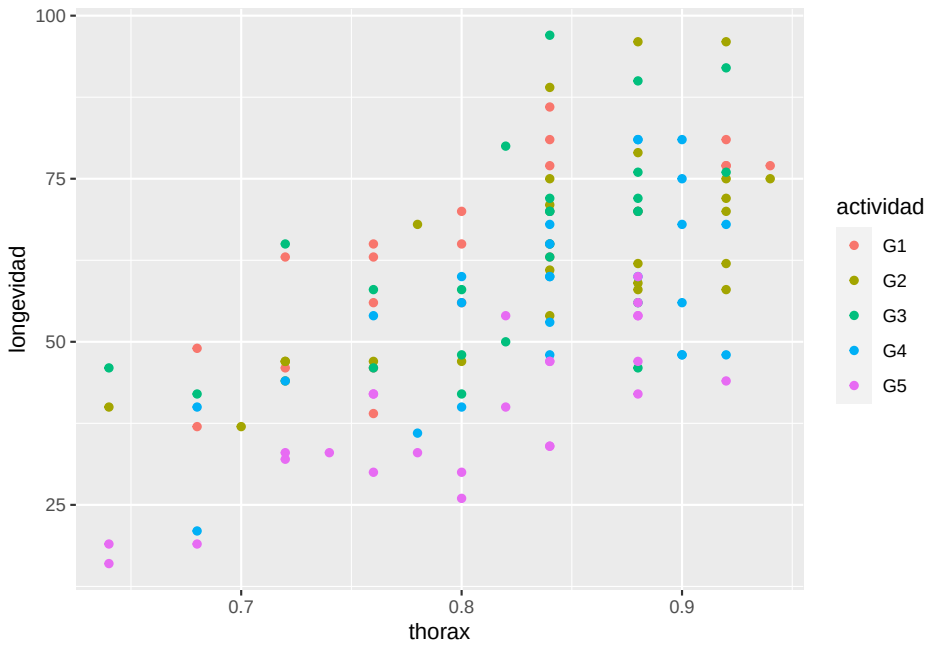
A la vista de la Figura anterior se puede apreciar que el comportamiento del tiempo de vida con repecto a la velocidad disminuye al aumentar esta última,

y además ese decenso es distinto en función de la herramienta utilizada. A la misma velocidad la herramienta A proporciona un tiempo de vida inferior que la herramienta B. Para ambas máquinas se aprecía una tendencia lineal bastante clara que deberá ser objeto de estudio. Esto implicaría que nuestro modelo se desdoblaría en dos rectas de regresión lineal simple en función de la herramienta utilizada.

Ejemplo 1.2 (Longevidad)**.** Partridge y Farquhar realizan un experimento para relacionar la vida útil (`longevidad`) de las moscas de la fruta con su actividad sexual (`actividad`). La información recogida es la longevidad en días de 125 moscas macho, divididas en cinco grupos bajo diferentes condiciones ambientales para medir su actividad sexual. Asimismo, se recoge la longitud del tórax (`thorax`) ya que se sospecha que afecta directamente a la longevidad de las moscas.

Cargamos los datos y realizamos el gráfico descriptivo:

```
# Carga de datos
thorax <- c(0.68, 0.68, 0.72, 0.72, 0.76, 0.76,
            0.76, 0.76, 0.76, 0.8, 0.8, 0.8, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.88, 0.88, 0.92
longevidad <- c(37, 49,  46,  63,  39,  46,  56,  63,  65,  56,  65,  70,  63,  65, 70,  77,  81,
actividad <- c(rep("G1",24),rep("G2",25),rep("G3",25),rep("G4",25),rep("G5",25))
longevidad <- data.frame(thorax,longevidad,actividad)
# Gráfico
ggplot(longevidad, aes(x = thorax, y = longevidad, color = actividad)) +
  geom_point()
```



Se puede ver como la longevidad aumenta caundo aumenta la longitud del thorax pero ese crecimiento no parece distinto según actividad, dado que las nuebes de puntos están bastante mezcladas. En este caso no prece adecuado un modelo lineal para cada grupo de actividad.

1.2 Modelo ANCOVA

Consideramos el modelo ANCOVA más sencillo donde consideramos dos variables predictoras: una numérica y otra un factor. Consideramos una muestra de tamaño n donde tenemos:

- Una variable respuesta, Y, de tipo numérico con observaciones $y_1, ..., y_n$.
- Una variable predictora, X, de tipo numérico con observaciones $x_1, ..., x_n$.
- Una variable predictora, F, de tipo categórico con I grupos o niveles distintos de tamaños muestrales $n_1, n_2, ..., n_I$, de forma que $n = n_1 + n_2 + ... + n_I$, de forma que el vector de observaciones de la respuesta y de la predictora numérica se pueden escribir como:

$$(Y_1, Y_2, ..., Y_I) = y_{11}, ..., y_{1n_1}, y_{21}, ..., y_{2n_2}, ..., y_{I1}, ..., y_{In_I}$$

$$(X_1, X_2, ..., X_I) = x_{11}, ..., x_{1n_1}, x_{21}, ..., x_{2n_2}, ..., x_{I1}, ..., x_{In_I}$$

  donde el primer subíndice indica el nivel del factor y el segundo la posición dentro del conjunto de datos de dicho nivel del factor.
- Conjunto μ_i de medias de todas las observaciones de la respuesta asociadas con el nivel i del factor, es decir:

$$\mu_i = \frac{\sum_{j=1}^{n_j} y_{ij}}{n_i};\ \text{i = 1, 2,..., I}$$

- Media global de la respuesta, μ, que se puede obtener como:

$$\mu = \frac{\sum_{j=1}^{I} \mu_j}{I}$$

- Incrementos, α_i, de cada una de las medias de cada grupo con respecto a la media global, es decir:

$$\alpha_i = \mu - \mu_i;\ \text{i = 1, 2,..., I}$$

- Pendiente común, β, que representa la posible relación entre las variables de tipo numérico.

- Pendientes diferentes entre las predictoras numéricas asociadas a cada nivel del factor, γ_i con $i = 1, ..., I$.

En esta situación el modelo que describe la posible relación entre respuesta y predictoras se puede escribir como:

$$y_{ij} = \alpha_0 + \alpha_i + \beta x_{ij} + \gamma_i x_{ij} + \epsilon_{ij}; \quad i = 1, ..., I \quad \text{con} \quad \alpha_I = 0 \tag{1.1}$$

Que en forma matricial se puede escribir fácilmente (de forma análoga al ANOVA de un vía) sin más que considerar $1_(n_i) = 1, ..., 1$ un vector de n_i unos, $0_(n_i) = 0, ..., 0$ un vector de n_i ceros, para cada uno de los niveles i del factor la matriz de diseño viene dada por:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_I \end{pmatrix} = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0_{n_1} & ... & 0_{n_1} & X_1 & X_1 & ... & 0_{n_1} \\ 1_{n_2} & 0_{n_2} & 1_{n_2} & ... & 0_{n_2} & X_2 & 0_{n_2} & ... & 0_{n_2} \\ ... & ... & ... & ... & ... & ... & ... & ... & 0_{n_3} \\ 1_{n_I} & 0_{n_I} & 0_{n_I} & ... & 1_{n_I} & X_I & 0_{n_I} & ... & X_I \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_{I_1} \\ \beta \\ \gamma_1 \\ ... \\ \gamma_I \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{pmatrix} = X\beta + \epsilon$$

donde las primeras I columnas representan el efecto del factor (ANOVA de una vía), la siguiente columna representa el efecto común entre predictoras numéricas (Regresión lineal simple), y las últimas I columnas representan el efecto distinto de la predictora numérica para cada nivel del factor. Estas columnas se obtienen fácilmente multiplicando la columna de X por cada de las columnas desde la 2 a la I.

Las posibles modelos anidados que se pueden obtener a partir de la ecuación (1.1), así como sus interpretaciones se presentan a continuación. Además se ofrece la interpretación de dichos modelos en términos de los coeficientes que resulten significativos. La estructura secuencial de contrastes pasa por estudiar si debemos considerar una pendiente distinta para la predictora numérica en función de los niveles del factor, que es equivalente a plantear el contraste:

$$\begin{cases} H_0: & \gamma_1 = \gamma_2 = ... = \gamma_I = 0 \\ H_a: & \text{Al menos hay una pendiente distinto de cero} \end{cases} \tag{1.2}$$

- Si rechazamos el contraste (1.2) tendremos un modelo de regresión entre la respuesta y la predictora numérica para cada nivel del factor, es decir, I rectas de regresión distintas con ecuaciones:

$$\begin{array}{ll} y_{1j} & = (\alpha_0 + \alpha_1) + (\beta + \gamma_1)x_{1j} + \epsilon_{1j} \\ y_{2j} & = (\alpha_0 + \alpha_2) + (\beta + \gamma_2)x_{2j} + \epsilon_{2j} \\ ... & = ... \\ y_{Ij} & = (\alpha_0 + \alpha_I) + (\beta + \gamma_I)x_{Ij} + \epsilon_{Ij} \end{array} \tag{1.3}$$

con interceptaciones $\alpha_0 + \alpha_i$ y pendientes $\beta + \gamma_i$ para cada uno de los I niveles del factor. Tenemos I modelos de regresión distintos (uno por cada nivel del factor), es decir, tenemos I rectas perpendiculares o que se cortan.

- Si no rechazamos el contraste (1.2) tendremos un único modelo de regresión entre la respuesta y la predictora pero con diferentes interceptaciones (la pendiente es la misma), es decir, I rectas paralelas cuyas ecuaciones vienen dadas por:

$$\begin{array}{ll} y_{1j} & = (\alpha_0 + \alpha_1) + \beta x_{1j} + \epsilon_{1j} \\ y_{2j} & = (\alpha_0 + \alpha_2) + \beta x_{2j} + \epsilon_{2j} \\ ... & = ... \\ y_{Ij} & = (\alpha_0 + \alpha_I) + \beta x_{Ij} + \epsilon_{Ij} \end{array} \tag{1.4}$$

En la situación donde el contraste (1.2) es no significativo podemos definir diferentes modelos anidados en función de los incrementos del factor ('s) y la pendiente β. Los contrates son:

$$\begin{cases} H_0: & \alpha_1 = \alpha_2 = ... = \alpha_I = 0 \\ H_a: & \text{Al menos hay un incremento distinto de cero} \end{cases} \tag{1.5}$$

y

$$\begin{cases} H_0: & \beta = 0 \\ H_a: & \beta \neq 0 \end{cases} \tag{1.6}$$

- Si rechazamos la hipótesis nula de (1.5) pero no rechazamos la hipótesis nula de (1.6) diríamos que no hay efecto de la predictora numérica pero que si podemos establecer diferencias entre las medias de la respuesta dadas por los diferentes niveles del factor (Modelo ANOVA con ecuaciones (1.7).
- Si no rechazamos la hipótesis nula de (1.5) pero si rechazamos la hipótesis nula de (1.6) diríamos que no hay efecto del factor pero si que podemos establecer un modelo de regresión entre la respuesta y la predictora numérica (Modelo de Regresión Lineal Simple con ecuación (1.8)).
- Si no rechazamos la hipótesis nula de (1.5) y no rechazamos la hipótesis nula de (1.6) estaríamos ante un modelo nulo donde el comportamiento de la respuesta no viene explicado por las predictoras consideradas.

Las ecuaciones de los modelos resultantes son:

- Modelo ANOVA:

$$\begin{array}{ll} y_{1j} & = (\alpha_0 + \alpha_1) + \epsilon_{1j} \\ y_{2j} & = (\alpha_0 + \alpha_2) + \epsilon_{2j} \\ ... & = ... \\ y_{Ij} & = (\alpha_0 + \alpha_I) + \epsilon_{Ij} \end{array} \tag{1.7}$$

- Modelo de Regresión Lineal Simple:

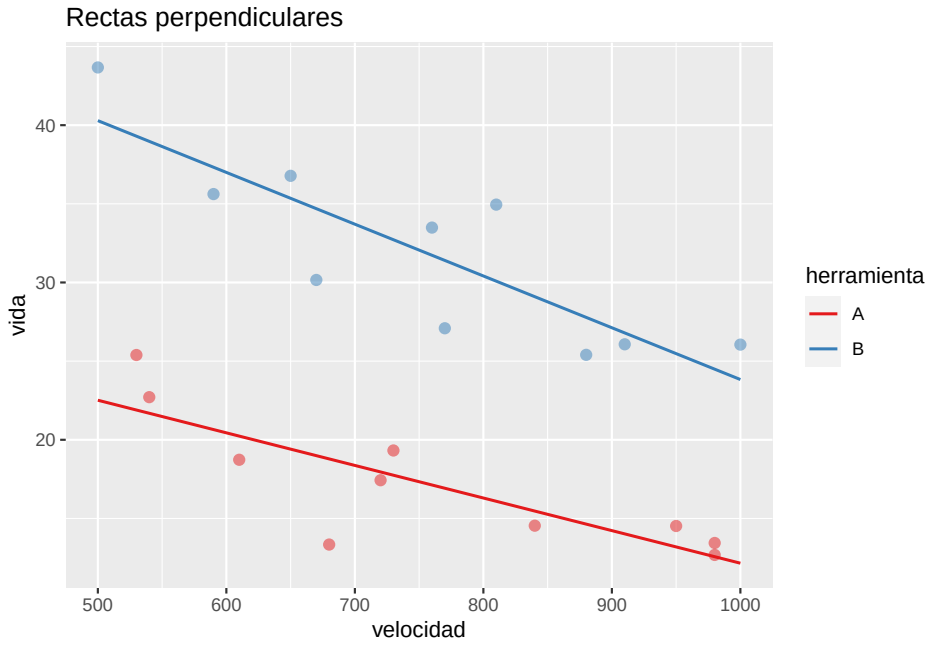
$$y_{ij} \quad = \alpha_0 + \beta x_{ij} + \epsilon_{ij} \tag{1.8}$$

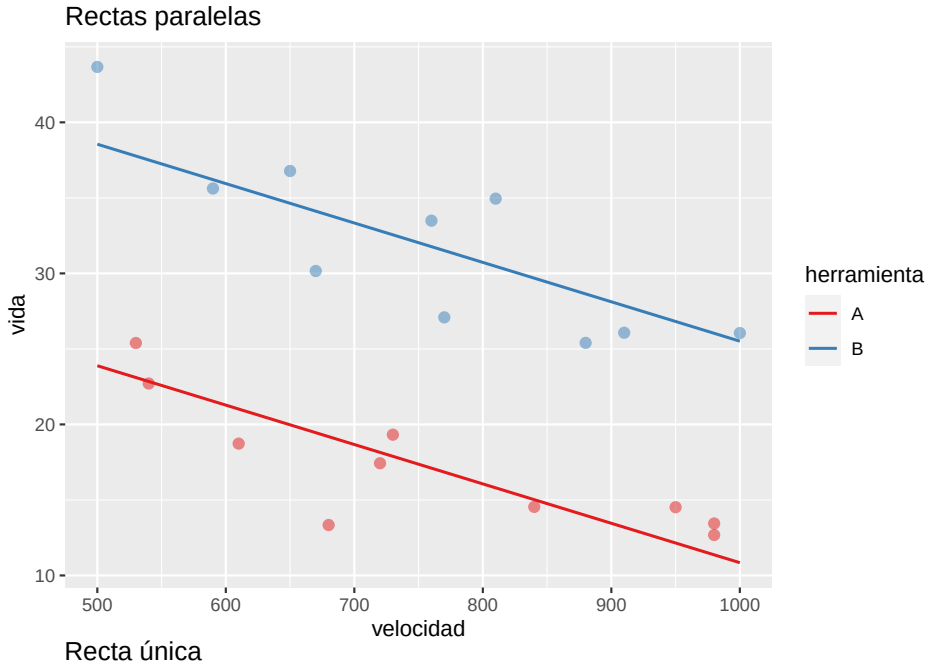
Atendiendo a las ecuaciones obtenidas las posible modelizaciones resultantes del modelo saturado son:

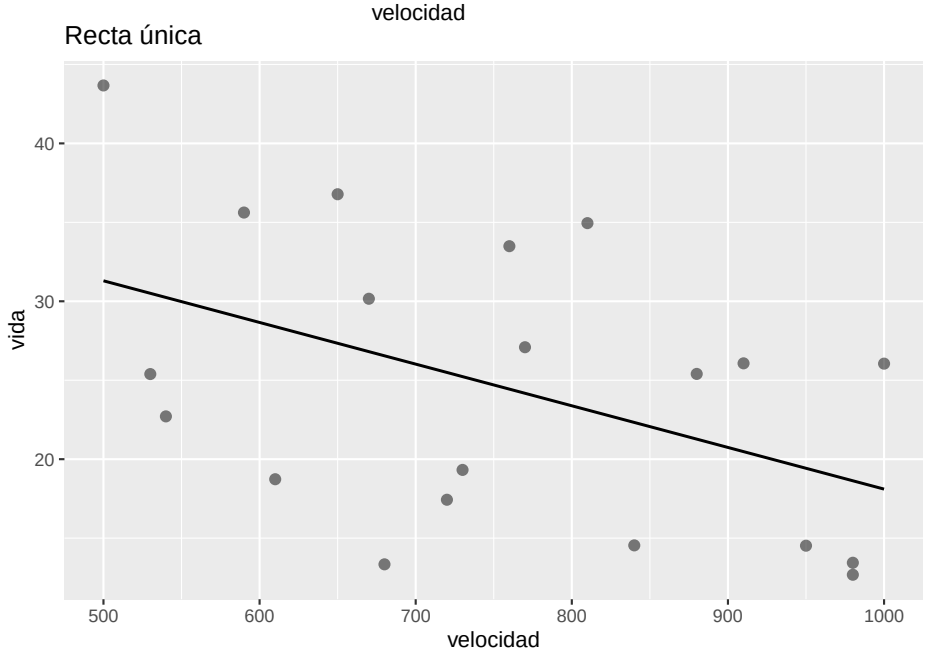
- I rectas que se cortan (modelo con interacción factor - numérica (1.3)).
- I rectas paralelas (modelo sin interacción pero con factor y numérica (1.4)).
- Una recta de regresión (modelo con la predictora numérica únicamente (1.8)).
- Diferencias entre las medias (modelo con el factor únicamente (1.7))

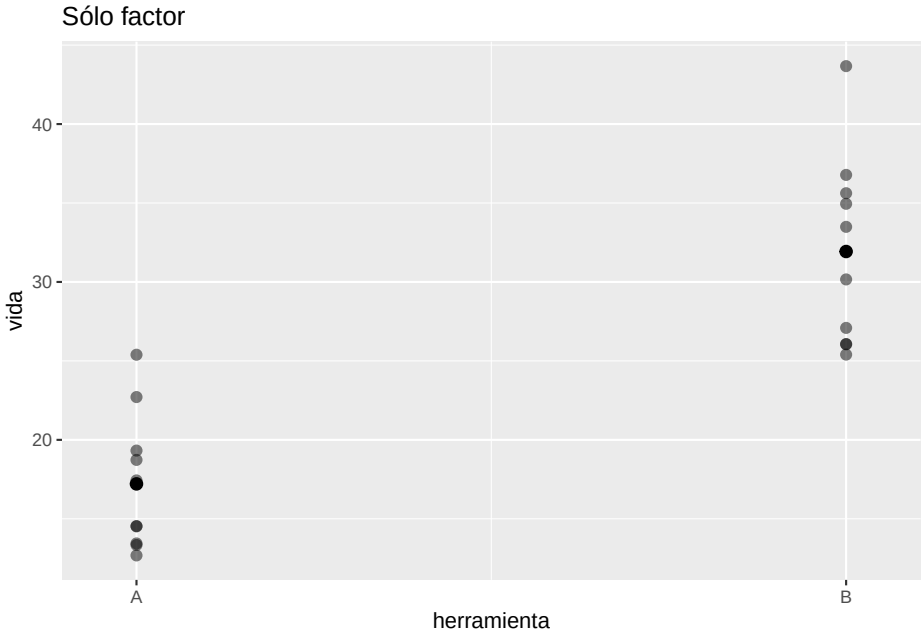
Vemos gráficamente los posibles modelos que podemos construir sobre los dos ejemplos presentados anteriormente:

- Datos de tiempo de vida: ¿Cómo interpretamos los gráficos obtenidos?

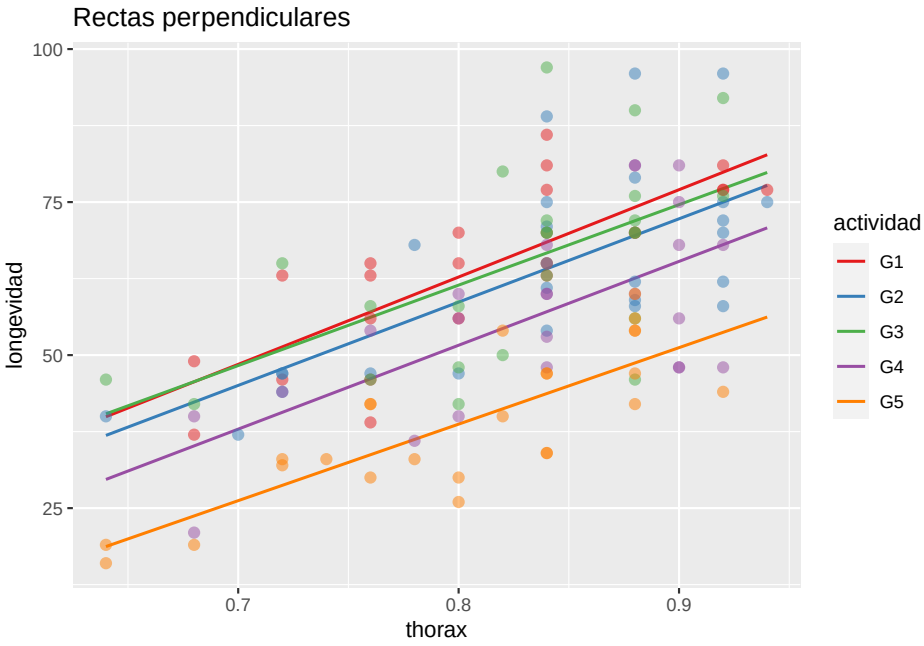


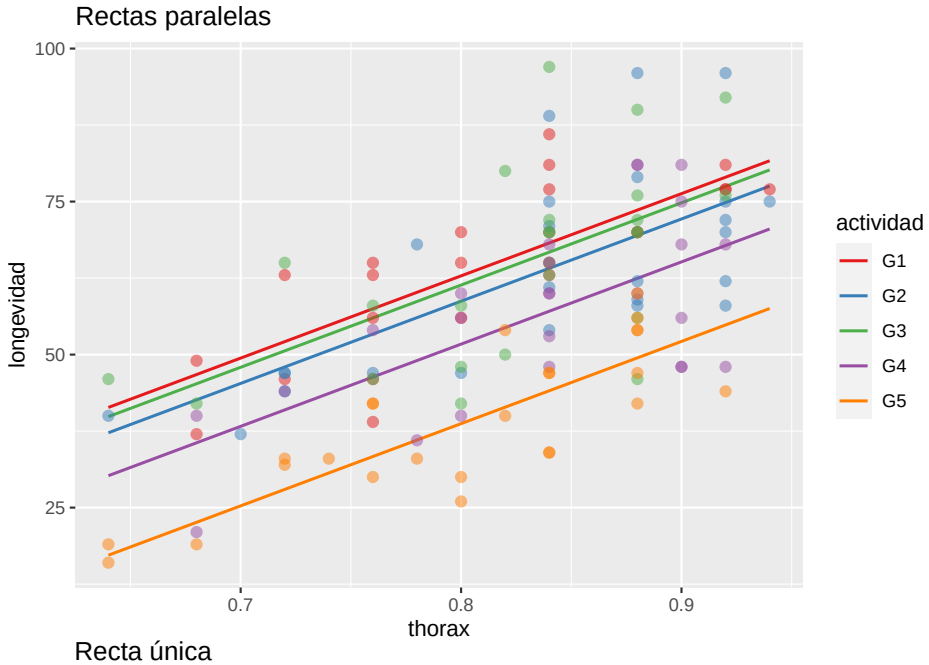

Rectas paralelas
vida
velocidad
herramienta
A
B
40
30
20
10
500
600
700
800
900
1000

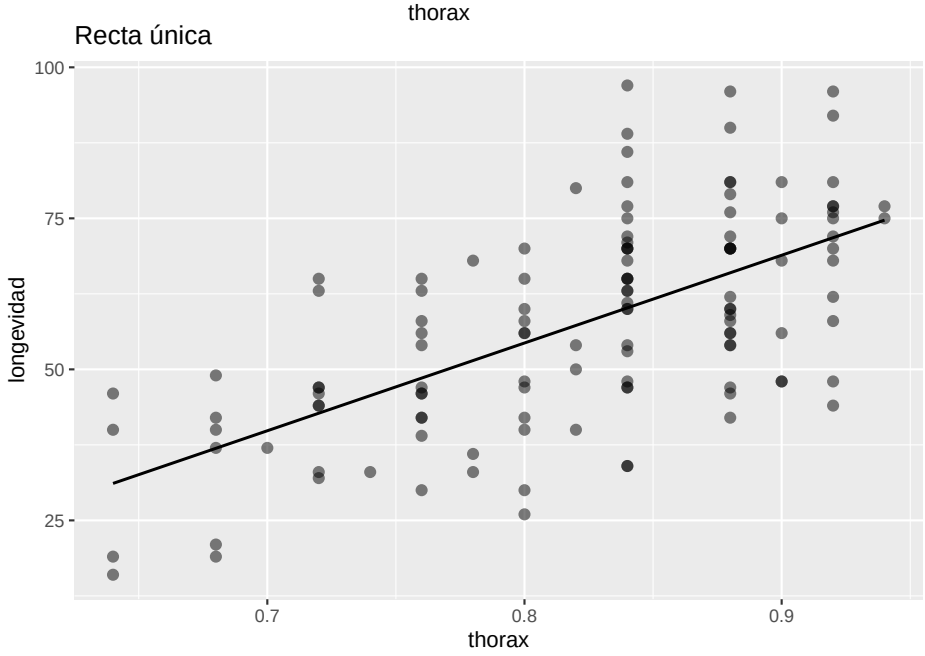

Recta única
vida
velocidad
40
30
20
500
600
700
800
900
1000

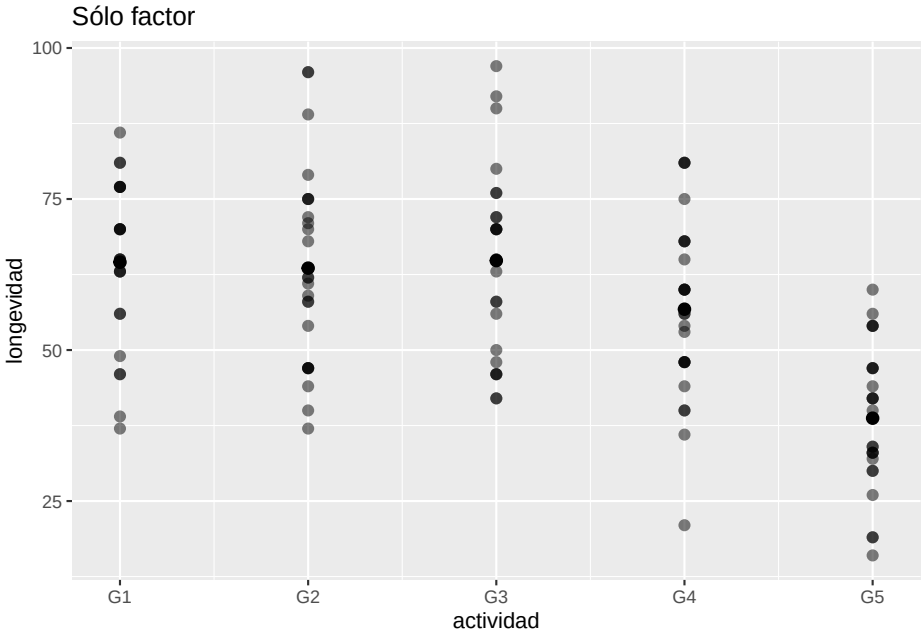


- Datos de longevidad: ¿Cómo interpretamos los gráficos obtenidos?




Rectas paralelas
longevidad
thorax
actividad
G1
G2
G3
G4
G5
100
75
50
25
0.7
0.8
0.9


Recta única
longevidad
thorax
100
75
50
25
0.7
0.8
0.9



Las hipótesis de este modelo es que los errores se distribuyen de forma independiente mediante una distribución Normal de media cero y varianza constante σ^2 para cada uno de los grupos que determina la variable predictora. Estas hipótesis se adaptarán en función del tipo de modelo que finalmente alcanzemos en el proceso de contraste. Todos ellos se pueden resolver mediante un test F, y más adelante veremos como construir la secuencia de modelos para decidir sobre el modelo final.

1.3 Especificación del modelo en R

El modelo ANCOVA planteado en el punto anterior para un factor F y una variable predictora numérica X, se puede escribir en R en su formato reducido como:

$$Y \sim F + X + F : X$$

donde:

- F representa el efecto del factor, es decir, comparamos si las medias de la respuesta para cada grupo pueden considerarse iguales.
- X representa el efecto de regresión asociado con la variable numérica, es decir, la respuesta y X están relacionadas mediante una única pendiente que deberemos estimar.

- $F : X$ representa el efecto de interacción entre predictoras, es decir, que la respuesta se relaciona con la predictora numérica mediante tantas curvas (generalmente líneas) como niveles tenga el factor F.

A continuación se presentan los modelos reducidos para diferentes situaciones experimentales en el número y tipo de varaibles predictoras:

- Modelo para dos factores (F_1 y F_2) y una numérica (X)

$$Y \sim F_1 + F_2 + F_1 : F_2 + X + F_1 : X + F_2 : X + F_1 : F_2 : X$$

o en forma más simplificada $Y \sim F_1 * F_2 * X$

- Modelo para un factor (F) y dos numéricas (X_1 y X_2)

$$Y \sim F + X_1 + X_2 + F : X_1 + F : X_2$$

o en forma más simplificada $Y \sim F * (X_1 + X_2)$

- Modelo para dos factores (F_1 y F_2) y dos numéricas (X_1 y X_2)

$$Y \sim F_1 * F_2 * (X_1 + X_2)$$

Como se puede ver la complejidad del modelo aumenta sustancialmente con la consideración de más variables predictoras. La forma de expresar el modleo saturado debe contemplar tanto os efectos principales asociados a cada predictora, como los efectos de interacción entre factores y entre factores y numéricas.

1.4 Estimación y Selección del modelo

Dado que hemos expresado el modelo ANCOVA como un modelo de tipo lineal con una ecuación similar a los modelos de regresión múltiple, la estimación de los parámetros del modelo se puede realizar utilizando las ecuaciones normales. En el proceso de selección del mejor modelo actuaremos como en los modelos ANOVA, es decir partiremos del modelo saturado y veremos que efectos puede ser considerados como irrelevantes, y por tanto deben desaparecer del modelo. Esta selección nos permitirá elegir el modelo final resultante. En este caso más sencillo podemos escribir todos los modelos posibles y elegir el mejor de ellos, bien mediante la comparación con el test F o con el AIC, pero veremos como utilizar los procedimientos secuenciales automáticos para problemas más complejos.

1.4.1 Ejemplos

Realizamos la selección y estimación del mejor modelo para cada uno de los conjuntos de datos considerados.

1.4.1.1 Datos de tiempo de vida

Construímos el modelo saturado y seleccionamos mediante el test F.

```
##
##
##                                 Elimination Summary
## ------------------------------------------------------------------------------------
##         Variable                                 Adj.
## Step           Removed           R-Square    R-Square     C(p)       AIC        RMSE
## ------------------------------------------------------------------------------------
##    1    velocidad:herramienta      0.8969      0.8847    3.9652    106.6591    3.0919
## ------------------------------------------------------------------------------------
```

El proceso de selección identifica el efecto de interacción entre `velocidad` y `herramienta` como no significativo, de forma que el modelo final viene dado por:

$$vida \sim velocidad + herramienta$$

Ajustamos el modelo y estudiamos los parámetros obtenidos:

vida

Predictors

Estimates

CI

(Intercept)

36.93

29.56 – 44.30

velocidad

-0.03

-0.04 – -0.02

herramienta [B]

14.67

11.75 – 17.58

Observations

20

En este caso tenemos un modelo con dos rectas paralelas (una por cada tipo de herramienta) cuyas ecuaciones de estimación vienen dadas por:

$$\begin{cases} \text{Herramienta A}: & \widehat{Vida}_A = 36.93 + 0 + 14.67 * velocidad & = 36.93 + 14.67 * velocidad \\ \text{Herramienta B}: & \widehat{Vida}_B = 36.93 - 0.03 + 14.67 * velocidad & = 36.90 + 14.67 * velocidad \end{cases}$$

Tenemos una interceptación mayor para la herramienta A indicando que la recta asociada con dicha herramienta está por encima de la de la herramienta B. Además, la pendiente negativa asociada con la velocidad indica que conforme aumenta esta disminuye el tiempo de vida. Ese efecto es mayor si usamos la herramienta de tipo B dado que la recta ajustada está por debajo de la de la herramienta A.

1.4.1.2 Datos de longevidad

Construímos el modelo saturado y seleccionamos mediante el test F.

```
##
##
##                                 Elimination Summary
## ------------------------------------------------------------------------------------
##          Variable                            Adj.
## Step        Removed          R-Square    R-Square      C(p)        AIC         RMSE
## ------------------------------------------------------------------------------------
##    1    thorax:actividad       0.6527       0.638    -3.7881    943.8165    10.5394
## ------------------------------------------------------------------------------------
```

El proceso de selección identifica el efecto de interacción entre `thorax` y `actividad` como no significativo, de forma que el modelo final viene dado por:

$$longevidad \sim thorax + actividad$$

Ajustamos el modelo y estudiamos los parámetros obtenidos:

longevidad

Predictors

Estimates

CI

(Intercept)

-44.61

-65.53 – -23.69

thorax

134.34

109.13 – 159.55

actividad [G2]

-4.14

-10.13 – 1.86

actividad [G3]

-1.50

-7.48 – 4.47

actividad [G4]

-11.15

-17.15 – -5.16

actividad [G5]

-24.14

-30.12 – -18.17

Observations

124

En este caso tenemos un modelo con dos rectas paralelas (una por cada tipo de actividad) cuyas ecuaciones de estimación vienen dadas por:

$$\begin{cases} \text{G1}: & \widehat{logevidad}_{G1} = -44.61 + 134.34 * thorax \\ \text{G2}: & \widehat{logevidad}_{G2} = -48.75 + 134.34 * thorax \\ \text{G3}: & \widehat{logevidad}_{G3} = -46.11 + 134.34 * thorax \\ \text{G4}: & \widehat{logevidad}_{G4} = -55.76 + 134.34 * thorax \\ \text{G5}: & \widehat{logevidad}_{G5} = -68.75 + 134.34 * thorax \end{cases}$$

Se observa un efecto de aumneto de la longevidad conforme aumenta la longitud del thorax. El grupo con mayor longevidad es el G1, ya que tiene la interceptación más grande, mientras que el que tiene menor longevidad es G5. El orden vendría dado por $G1 > G3 > G2 > G4 > G5$.

1.5 Diagnóstico del modelo

En este caso el diagnóstico es similar al de los modelos de regresión pero teniendo en cuenta que las hipótesis se deben verificar para los residuos asociados a cada nivel del factor (si este está presente en el modelo). Las hipótesis son linealidad, normalidad y varianza constante. Para verificar las hipótesis utilizamos los procedimientos gráficos y tests tratados en las unidades anteriores.

De nuevo utilizaremos la distancia der Cook para establecer posibles observaciones influyentes.

1.5.1 Ejemplos

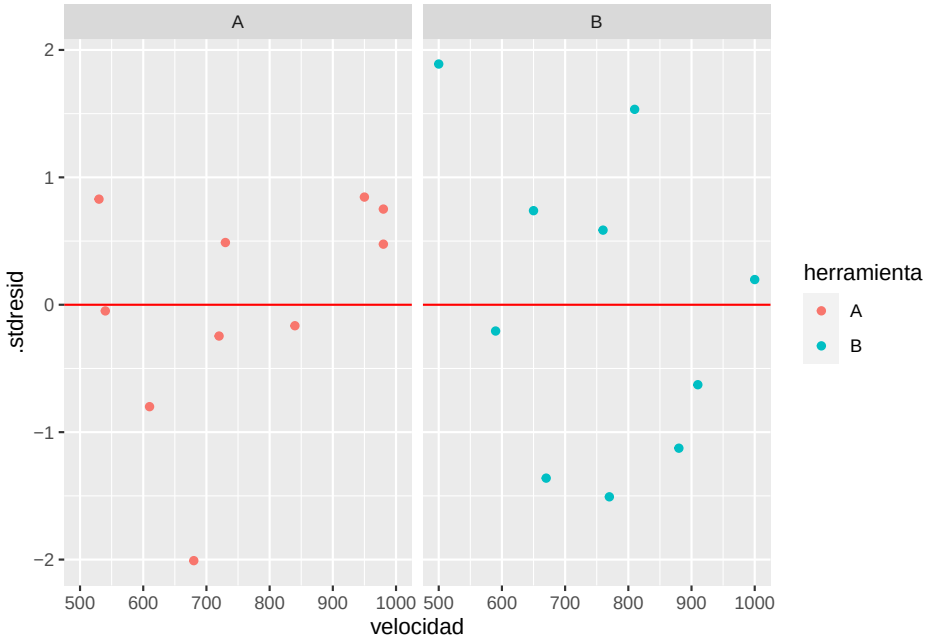
A continuación realizamos el diagnóstico de los ejemplos que venimos trabajando en esta unidad. Para obtener el diagnóstico partimos del modelo obtenido en la sección anterior y realizamos el gráfico de residuos versus ajustados y los tests de diagnóstico.

Como en los dos ejemplos el factor considerado se encuentra en el modelo final, utilizaremos el test de Levene para verificar la igualdad de varianzas entre los niveles del factor.

1.5.1.1 Datos de tiempo de vida

Obtenemos los valores de diagnóstico y realizamos los correspondientes tests de hipótesis y análisis de influencia.

```
# Valores de diagnóstico
diagnostico <- fortify(fit.vida)
# Gráfico
ggplot(diagnostico,aes(x = velocidad, y = .stdresid, colour = herramienta)) +
   geom_point() +
   geom_hline(yintercept = 0, col = "red") +
   facet_wrap(. ~ herramienta)
```



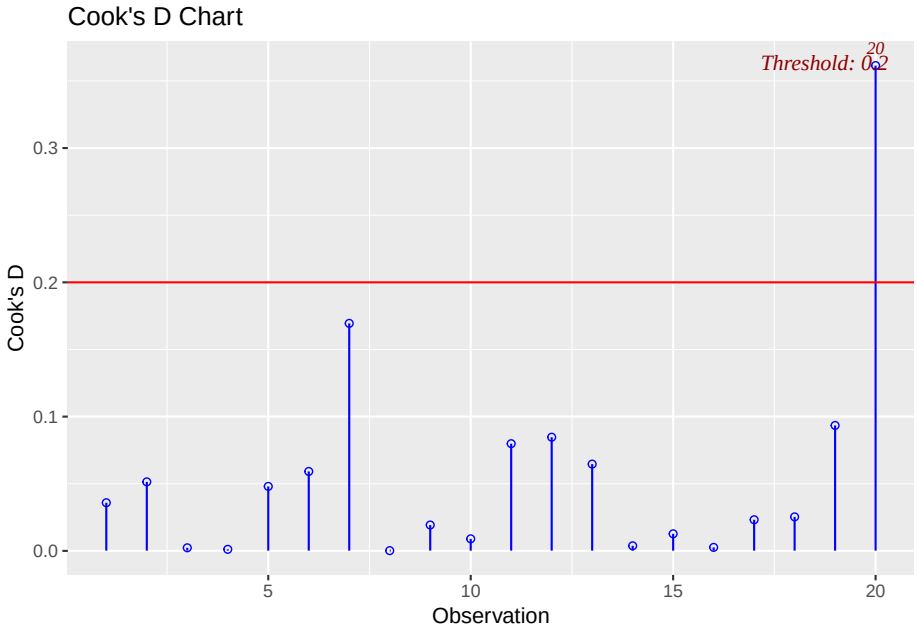

```
# Tests de hipótesis
ols_test_normality(fit.vida)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9715         0.7858
## Kolmogorov-Smirnov        0.1232         0.8859
## Cramer-von Mises          1.4412          2e-04
## Anderson-Darling          0.2652         0.6555
## -----------------------------------------------
```

```
leveneTest(.stdresid ~ herramienta, data = diagnostico)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.3888  0.254
##       18
```

```
# Análisis de influencia
ols_plot_cooksd_chart(fit.vida)
```

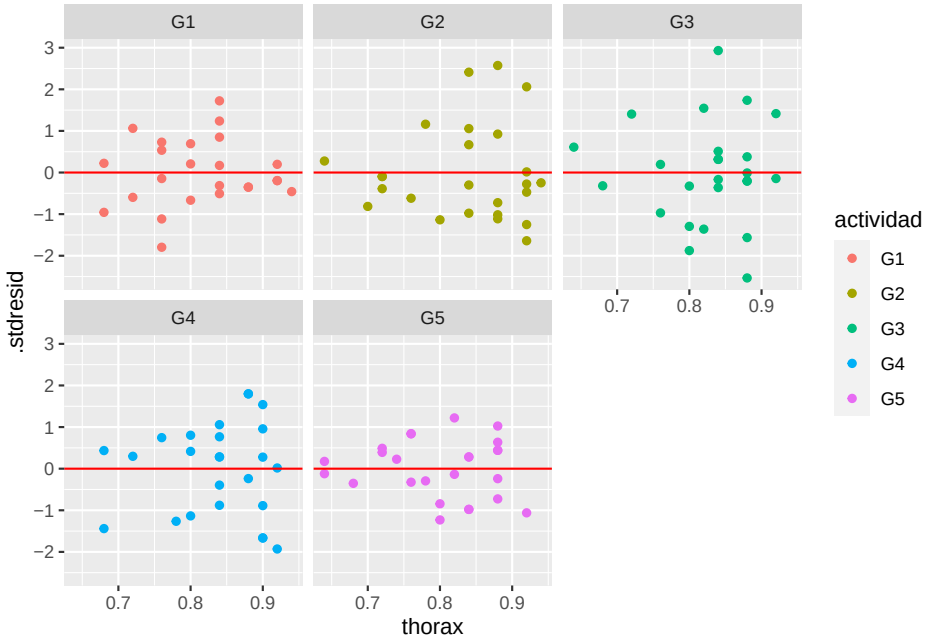


Los residuos tienen un comportamiento aleatorio, se verifican las hipótesis del modelo y no se detectan observaciones influyentes. El modelo obtneido parace adecuado para estudiar el tiempo de vida en función de la velocidad y la herramienta utilizada.

1.5.1.2 Datos de longevidad

Obtenemos los valores de diagnóstico y realizamos los correspondientes tests de hipótesis y análisis de influencia.

```
# Valores de diagnóstico
diagnostico <- fortify(fit.longevidad)
# Gráfico
ggplot(diagnostico,aes(x = thorax, y = .stdresid, colour = actividad)) +
   geom_point() +
   geom_hline(yintercept = 0, col = "red")+
   facet_wrap(. ~ actividad)
```



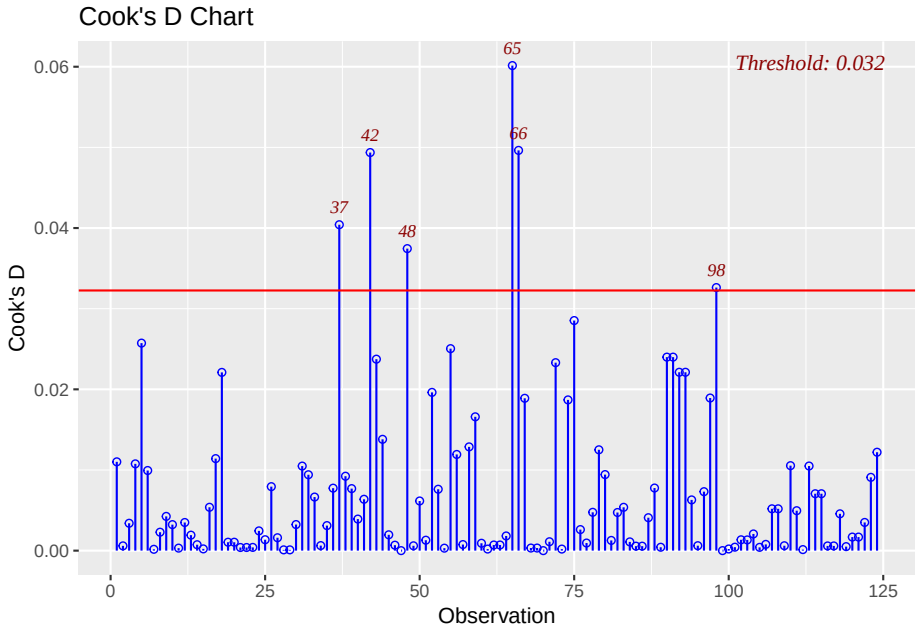

```
# Tests de hipótesis
ols_test_normality(fit.longevidad)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9916         0.6607
## Kolmogorov-Smirnov        0.0538         0.8654
## Cramer-von Mises         10.2413         0.0000
## Anderson-Darling          0.3224         0.5241
## -----------------------------------------------
```

```
leveneTest(.stdresid ~ actividad, data = diagnostico)
```

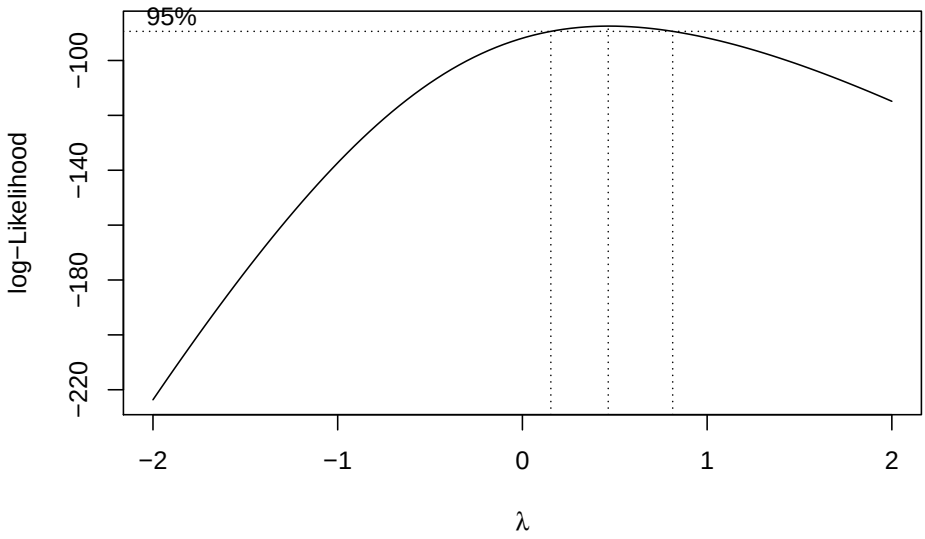
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  1.2925 0.2769
##       119
```

```
# Análisis de influencia
ols_plot_cooksd_chart(fit.longevidad)
```



Aunque se verifican las hipótesis del modelo y no se detectan observaciones influyentes, si que es cierto que los gráficos de residuos muestram cierto comportamiento de embudo con varaibilidades más pequeñas en valores más pequeños de thorax, y mayor dispersión al aumentar la longitud del thorax. Sería recomendable probar Box-Cox para tratar de obtner una transfromación de la respuesta que nos permita obteenr gráficos sin esos efectos indeseables.

```
MASS::boxcox(fit.longevidad)
```



La transformación raíz cuadrada parece adecuada en esta situación. Obtenemos la nueva variable y ajustamos de nuevo el modelo.

```
# Trnasformación
longevidad <- longevidad %>% mutate(rlongevidad = sqrt(longevidad))
# Modelo saturado
fit.longevidad <- lm(rlongevidad ~ thorax * actividad, data = longevidad)
# Selección del modelo
ols_step_backward_p(fit.longevidad, prem = 0.05)
```

```
##
##
##                                Elimination Summary
## ------------------------------------------------------------------------------------
##         Variable                            Adj.
## Step        Removed         R-Square    R-Square     C(p)        AIC        RMSE
## ------------------------------------------------------------------------------------
##    1    thorax:actividad      0.6868      0.6736    -3.0694    267.2943    0.6888
## ------------------------------------------------------------------------------------
```

De nuevo el modelo seleccionado prescinde del efecto de interacción. Ajustamos y estudiamos el nuevo modelo.

rlongevidad

Predictors

Estimates

CI

(Intercept)

0.34

-1.03 – 1.71

thorax

9.41

7.77 – 11.06

actividad [G2]

-0.30

-0.69 – 0.09

actividad [G3]

-0.12

-0.51 – 0.27

actividad [G4]

-0.76

-1.15 – -0.37

actividad [G5]

-1.73

-2.12 – -1.34

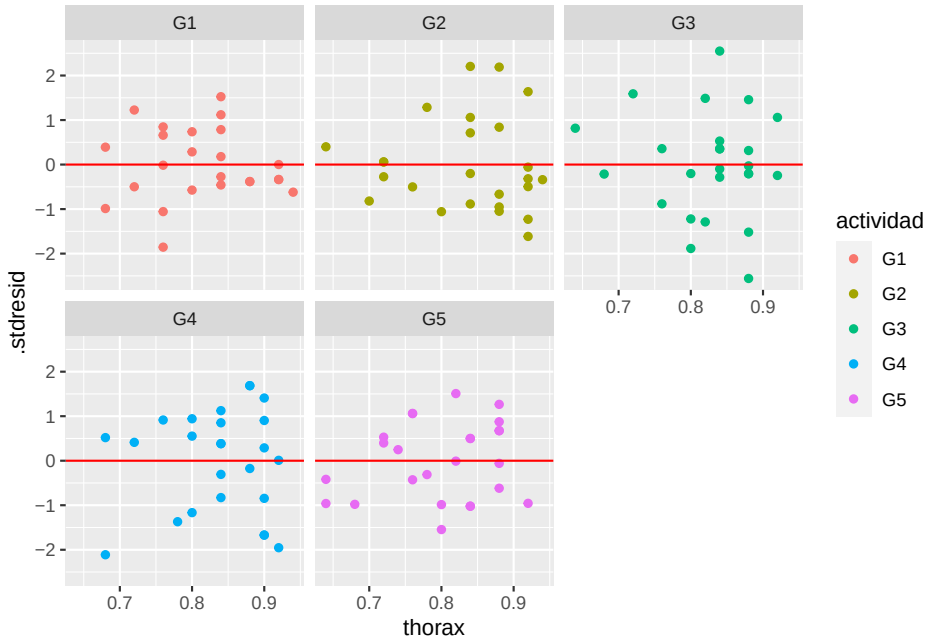
Observations

124

¿Cuáles son las ecuaciones de estimación en este caso?

El proceso de diagnóstico para el nuevo modelo permite verificar el cumplimineto de las hipótesis y la leve mejora de los gráficos de residuos.

```
# Valores de diagnóstico
diagnostico <- fortify(fit.longevidad)
# Gráfico
ggplot(diagnostico,aes(x = thorax, y = .stdresid, colour = actividad)) +
   geom_point() +
   geom_hline(yintercept = 0, col = "red")+
   facet_wrap(. ~ actividad)
```



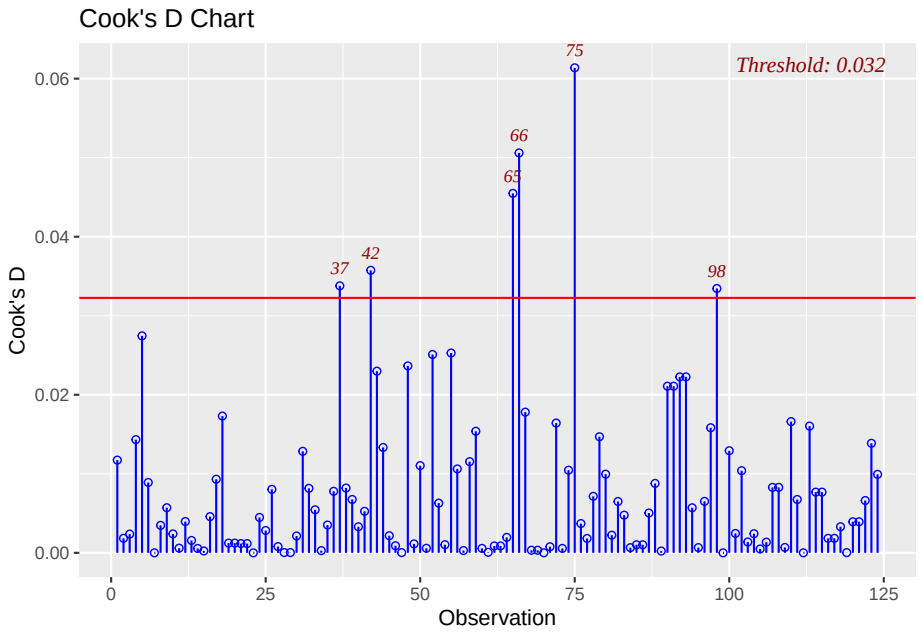
```
# Tests de hipótesis
ols_test_normality(fit.longevidad)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9951         0.9465
## Kolmogorov-Smirnov        0.0474         0.9432
## Cramer-von Mises         10.7431         0.0000
## Anderson-Darling          0.2137         0.8484
## -----------------------------------------------
```

```
leveneTest(.stdresid ~ actividad, data = diagnostico)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.6856 0.6033
##       119
```

```
# Análisis de influencia
ols_plot_cooksd_chart(fit.longevidad)
```



1.6 Predicción

El proceso de predicción en este tipo de modelos es muy simple a partir de las ecuaciones de los modelos obtenidas. De hecho, en las secciones anteriores ya hemos visto gráficamente la predicción para todos estos modelos en los ejemplos que hemos ido trabajando. Básicamente, si queremos obtener una predicción especifica deberemos dar un valor del factor y otro de la predictora numérica para calcular el valor de predicción y su correspondiente intervalo. En este caso nos imitamos a representar las bandas de predicción que podemos obtener para cada modelo.

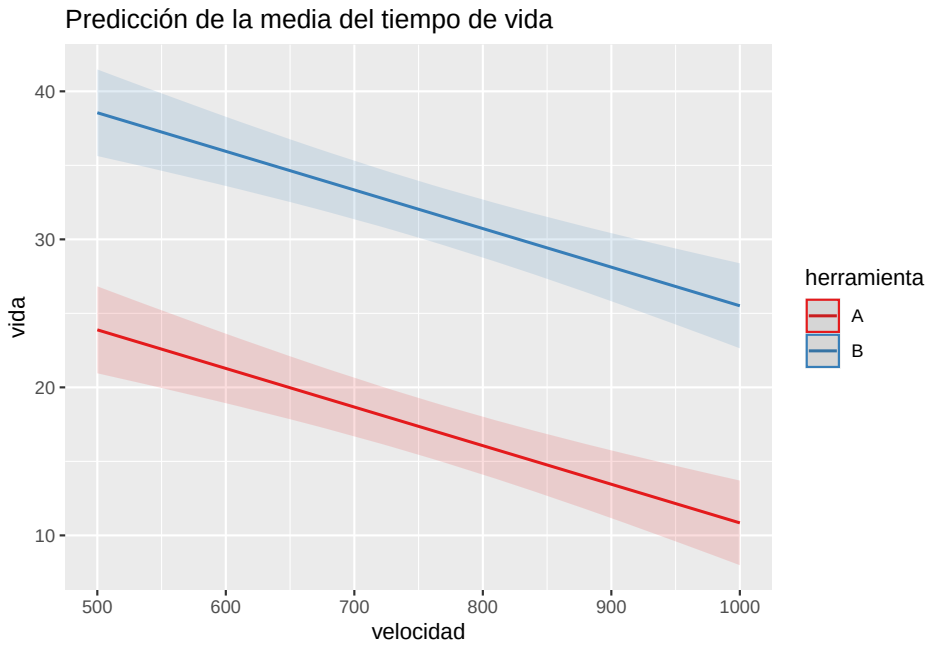
1.6.1 Ejemplos

Construimos la predicción y bandas de confianza para cada uno de los ejemplos a aprtir de los modelos determinados tras la fase de diagnóstico.

1.6.1.1 Datos de tiempo de vida

A continuación se presentan las rectas de predicción para el modelo ajustado a este abnco de datos.

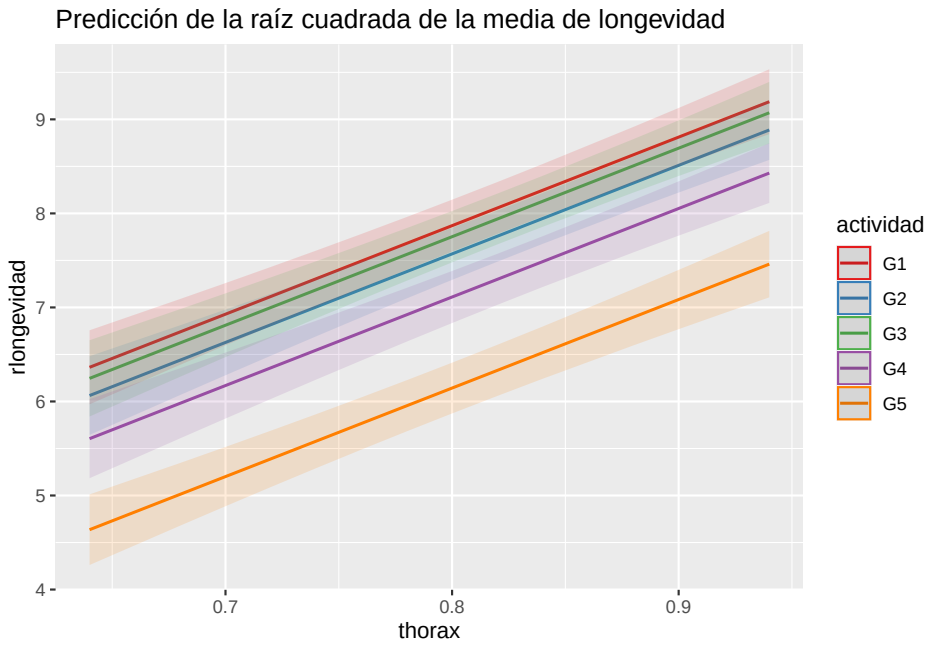
```
plot_model(fit.vida, "pred", terms = c("velocidad","herramienta"),
           title ="Predicción de la media del tiempo de vida")
```



1.6.1.2 Datos de longevidad

A continuación se presentan las rectas de predicción para el modelo ajustado a este abnco de datos.

```
plot_model(fit.longevidad, "pred", terms = c("thorax","actividad"),
           title ="Predicción de la raíz cuadrada de la media de longevidad")
```



1.7 Ejercicios

A continuación se presenta la colección de ejercicios de esta unidad.

Ejercicio 1. Disponemos de los datos de peso de 24 niños recién nacidos (`peso`), su sexo (`sexo`; "H" = Hombres y "M" = Mujeres) y la edad de sus madres (`edad`). Nos gustaría ser capaces de determinar un modelo que explique el peso de los niños recién nacidos en función de su sexo y de la edad de sus madres.

```
# Lectura de datos
edad <- c(40, 38, 40, 35, 36, 37, 41, 40, 37, 38, 40, 38,
          40, 36, 40, 38, 42, 39, 40, 37, 36, 38, 39, 40)
peso <- c(2968, 2795, 3163, 2925, 2625, 2847, 3292, 3473, 2628, 3176,
          3421, 2975, 3317, 2729, 2935, 2754, 3210, 2817, 3126, 2539,
          2412, 2991, 2875, 3231)
sexo <- gl(2,12, labels=c("H","M"))
ejer01 <- data.frame(edad, peso, sexo)
```

Ejercicio 2. Se lleva a cabo una investigación sobre diversas malformaciones del sistema nervioso central registradas en nacidos vivos en Gales del Sur, Reino Unido. El estudio fue diseñado para determinar el efecto de la dureza del agua sobre la incidencia de tales malformaciones. La información registrada son: `NoCNS` = recuento de nacimientos sin problema CNS; `An` = conteo de nacimientos de Anencephalus; `Sp` = conteo de nacimientos de espina bífida; `Otro` = recuento de otros nacimientos del SNC; `Agua` = endurecimiento del agua; `Trabajo` = un factor con niveles Manual no manual en función del tipo de trabajo realizado por los padres Se está interesado en predecir el número total de malformaciones en función de la calidad del agua y el trabajo realizado por los padres.

```
# Lectura de datos
previo <- read_csv("https://goo.gl/bNOSxt", col_types = "cdddddc")
# Calculamos el número total de malformaciones
ejer02 <- previo %>% mutate(CNS=An+Sp+Other)
```

Ejercicio 3. Se ha realizado un estudio para establecer la calidad de los vinos de la variedad Pino Noir en función de un conjunto de características analizadas. Las características analizadas son `claridad`, `aroma`, `cuerpo`, `olor` y `matiz`. Para medir la `calidad` se organiza una cata ciega a un conjunto de expertos y se calcula la puntuación final de cada vino a partir de la información de todos ellos. Además se registra la región (`region`) de procedencia del vino por si puede influir en la calidad del vino.

```
# Lectura de datos
ejer03 <- read_csv("https://goo.gl/OX9wgM", col_types = "ddddddc")
```

Ejercicio 4. Un fabricante de ropa que suministra uniformes militares debe cortar chaquetas, camisas, pantalones (variable `Prenda`) y otros complementos (en muchas tallas diferentes), de rollos de tela. La tela es cara, de modo que el desperdicio (`Desperdicio`) tiene un efecto muy grande en los beneficios. El fabricante tiene que elegir entre tres máquinas (`Maquina`) cortadoras asistidas por computadora: A, B y C. El fabricante decide experimentar haciendo que cada máquina corte varios lotes de chaquetas, varios más de camisas otros más de pantalones y complementos para determinar que máquina es más eficiente en cada caso, es decir, tratamos de conocer el desperdicio que se producirá para cada prenda y máquina.

```
# Lectura de datos
ejer04 <- read_csv("https://bit.ly/2GcVn3R", col_types = "ccd")
ejer04 <- ejer04 %>%
  mutate_if(sapply(ejer04,is.character),as.factor)
```

Ejercicio 5. Una empresa dedicada a la fabricación de aislantes térmicos y acústicos establece un experimento que mide la pérdida de calor (`Calor`) a través de cuatro tipos diferentes de cristal para ventanas (`Cristal`) utilizando cinco graduaciones diferentes de temperatura exterior (`TempExt`). Se prueban tres hojas de cristal en cada graduación de temperatura, y se registra la pérdida

de calor para cada hoja.

```
# Lectura de datos
ejer05 <- read_csv("https://goo.gl/V6hyVW", col_types = "ddc")
ejer05 <- ejer05 %>%
  mutate_if(sapply(ejer05,is.character),as.factor)
```

Ejercicio 6. El grupo de asesores LearnStatistics ha realizado un estudio para comprobar si las empresas destinan parte de los beneficios de sus ventas en la formación de sus empleados para mejorar su competitividad. Para ellos se recoge la información sobre ventas (`Ventas`) en miles de euros, capital invertido en formación (`Capital`) en miles de euros, y el nivel de productividad de la empresa establecido por un asesor externo (`Productividad`).

```
# Lectura de datos
ejer06 <- read_csv("https://bit.ly/2rCATaO", col_types = "dcd")
ejer06 <- ejer06 %>%
  mutate_if(sapply(ejer06,is.character),as.factor)
```

Ejercicio 7. Una empresa recibe cargamentos de material para procesar en sus almacenes (`almacen`). El objetivo básico del estudio es determinar el tiempo de procesado (`tiempo`) de los cargamentos recibidos como función del tamaño del cargamento (`tamanyo`).

```
# Lectura de datos
lectura <- read.table("https://goo.gl/kuMNpD", header = TRUE)
ejer07 <- as_tibble(lectura)
```

Ejercicio 8. La lista Forbes500 la componen las 500 empresas con mayores ventas mundiales. En este caso se desea realizar un estudio de un grupo de empresas de la lista Forbes para intentar relacionar las `Ventas` (en millones de dólares) con los `Activos` de la empresa (en millones de dólares), el `ValorBursatil` (en millones de dólares), el número de `Empleados` (en miles de personas) y el `Sector` económico al que pertenece la empresa (Alta Tecnología, Energía, Finanzas, Manufacturas, Medico, Transportes, Venta al detalle, y Otros). Como en todos los problemas económicos resulta mucho más cómodo trabajar con las variables transformadas con el logaritmo. La variable `Empresa` es un identificador y no debe ser utilizada en el proceso de modelización.

```
# Lectura de datos
lectura <- read.table("https://goo.gl/PHQXaW", header = TRUE)
ejer08 <- as_tibble(lectura)
```

Ejercicio 9. Se conoce como infiltración al proceso por el cual el agua (riego o lluvia) se va introduciendo bajo la superficie de un terreno cultivado. Este proceso es vital para determinar las cantidades de agua de riego necesarias, para mantener el terreno en condiciones óptimas. Un parámetro habitual que sirve para estudiar dicho proceso es la carga hidráulica (`cargahid`). Este depende tanto de la profundidad de la infiltración (`profundidad`) como del procedimiento

de riego usado. Se diseña un experimento para estudiar la carga hidráulica de un terreno bajo diferentes condiciones de riego (`trata`). los resultados obtenidos aparecen en la base de datos correspondiente.

```
# Lectura de datos
lectura <- read.table("https://goo.gl/1JnUVX", header = TRUE)
ejer09 <- as_tibble(lectura)
```