

Práctica 6: Análisis de la varianza. Regresión lineal y correlación.

1. Análisis de la varianza

1.1. Introducción

En la primera parte de esta práctica vamos a utilizar `RCommander` de nuevo para realizar un análisis de la varianza. En esta parte de la práctica vamos a utilizar en banco de datos `Tire` del paquete `PASWR`.

EJERCICIO:

Carga el paquete `PASWR` con la instrucción

```
> library(PASWR)
```

o mediante el menú **HERRAMIENTAS->CARGAR PAQUETE(S)** de `RCommander`. Esto hará que los bancos de datos de este paquete estén disponibles para que se puedan cargar desde `RCommander`.

EJERCICIO:

Carga el banco de datos `Tire` del paquete `PASWR` con `RCommander`. Para ello tendrás que seleccionar **DATOS->CONJUNTO DE DATOS EN PAQUETES->LEER CONJUNTO DE DATOS DESDE PAQUETE ADJUNTO**.

Esto hará que el banco de datos `Tire` sea el banco de datos activo en `RCommander`.

Este banco de datos contiene dos variables:

- `stopDist`

Distancia en pies que necesita un coche 'típico' para detenerse completamente cuando va a una velocidad de 60 millas por hora.

- `tire`

Tipo de neumático, que puede ser A, B, C o D.

1.2. Análisis exploratorio de datos

Antes de proceder con el análisis de la varianza es conveniente realizar un análisis exploratorio de datos para ver qué es lo que tenemos entre manos. Es importante ver qué valores tienen las variables y de qué tipo son.

EJERCICIO:

Obtén estadísticos resumen de las dos variables de los datos. Puedes hacerlo utilizando las distintas opciones en `ESTADÍSTICOS->RESÚMENES`. ¿Observas algo extraño? ¿Qué ocurre con los estadísticos resumen de la variable `Tire`?

EJERCICIO:

A continuación haz un histograma de la variable `stopDist`. ¿Por qué crees que la variable `Tire` no aparece en la lista? ¿Puedes hacer algún otro gráfico de `Tire`?

1.2.1. Gráfica de las medias

Para comparar gráficamente las medias de los distintos grupos tenemos la opción `Gráficas->Gráficas de las medias`, que nos representa las medias de cada uno de los grupos junto con un intervalo. Las distintas opciones aparecen en la Figura 1.

En primer lugar, tenemos que definir las variables que nos definen los grupos (variable `Factor`) y la que contiene los datos medidos en cada grupo (variable `Explicada`). Además de la media de cada grupo, podemos añadir un 'intervalo' basado en:

- Errores típicos
- Desviaciones típicas
- Intervalos de confianza
- Sin barras de errores (sólo dibuja las medias)

EJERCICIO:

Realiza 3 gráficas de las medias comparando los distintos intervalos que podemos representar. Si quieres que una nueva gráfica se dibuje en una nueva ventana, ejecuta el comando `windows()` en la `Ventana de instrucciones` de `RCommander` para que aparezca una nueva ventana gráfica en blanco.



Figura 1: RCommander permite realizar un gráfico para comparar las medias de distintos grupos.

¿Qué diferencias observas entre los distintos intervalos? ¿Cuáles son más anchos y cuáles más estrechos? ¿Por qué crees que es?

¿Crees que hay diferencias entre grupos? ¿Se te ocurre alguna forma sencilla e intuitiva de determinar qué grupos tienen medias similares?

1.2.2. Diagrama de cajas por grupos

De manera similar podemos hacer un diagrama de cajas por grupos, de manera que tendremos 4 diagramas de cajas (uno para cada tipo de neumático). La Figura 2 muestra la ventana para realizar el diagrama de cajas y la ventana que aparece cuando pulsamos en **Gráfica por grupos** para seleccionar la variable que nos define los grupos.

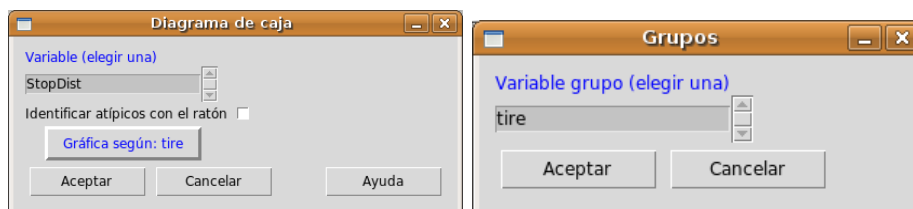


Figura 2: RCommander permite realizar diagramas de cajas por grupos.

EJERCICIO:

Realiza un diagrama de cajas por grupos para los datos con los que estamos trabajando. ¿Qué similitudes (y diferencias) observas con respecto a los gráficas de las medias producidos anteriormente?

1.3. Análisis de la varianza y tabla ANOVA

Para realizar el análisis de la varianza que hemos visto en clase podemos seleccionar ESTADÍSTICOS->MEDIAS->ANOVA DE UN FACTOR. La ventana que aparece es la que vemos en la Figura 3 Como puede verse, tenemos que seleccionar la variable que define los grupos y la variable que contiene las mediciones de cada grupo (variable explicada).

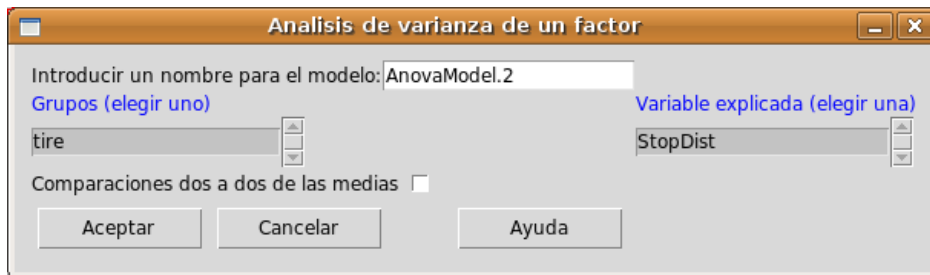


Figura 3: Ventana para hacer la ANOVA.

Además, tenemos la opción de seleccionar **Comparaciones dos a dos de las medias**, que nos dará un gráfico en el que se nos comparan los grupos de dos en dos.

EJERCICIO:

Realiza un análisis ANOVA de los datos y las comparaciones de grupos dos a dos. ¿Cuáles son tus conclusiones?

EJERCICIO:

El banco de datos FCD contiene datos de un experimento en el que se sometieron a una serie de gatos a 4 tipos de dietas. Las variables son **Weight** (diferencia de peso) y **Diet** (tipo de dieta). Haz una análisis similar al que has hecho con los datos de enumerativos y determina si existen diferencias entre las distintas dietas.

EJERCICIO:

El banco de datos `TireWear` de un experimento en el que se midió la pérdida de goma de unos neumáticos después de conducir durante 10000 millas. Las variables que se midieron son la pérdida de goma (`Wear`), el tipo de coche (`Block`) y cuatro tipos de neumáticos (variable `Treat`).

Haz una análisis similar al que has hecho con los datos de neumáticos y determina si existen diferencias entre los distintos tipos de neumáticos.

2. Regresión lineal y correlación

A continuación vamos a ver cómo se puede realizar una regresión lineal con R y RCommander.

EJERCICIO:

Como en la práctica anterior, carga el banco de datos `cars` del paquete `datasets`.

El objetivo de esta parte de la práctica es estudiar la relación entre las dos variables de este banco de datos (`speed` y `dist`). `speed` será la variable independiente (x) y `dist` la variable dependiente (y).

2.1. Análisis exploratorio de datos

El primer paso a la hora de estudiar la relación entre las dos variables es hacer un diagrama de puntos de las dos variables. Ésto se puede hacer en el menú GRÁFICAS->GRÁFICAXY. La ventana que aparece se puede ver en la Figura 4.

EJERCICIO:

Produce una gráfica de puntos con las dos variables del banco de datos `cars` en la que `speed` sea la variable independiente y `dist` la dependiente. ¿Parece que la relación es lineal?

2.2. Ajuste de la recta de regresión

La manera en la que ajustamos una recta a estos datos con RCommander se encuentra en el menú ESTADÍSTICOS->AJUSTE DE MODELOS->REGRESIÓN LINEAL. La ventana que aparece se puede ver en la Figura 5.

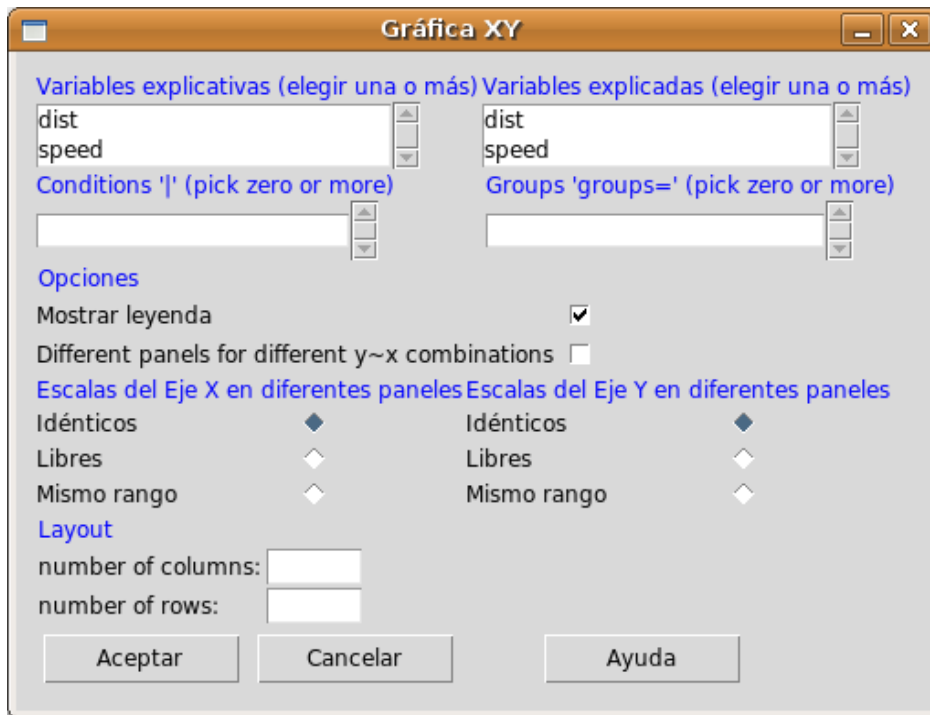


Figura 4: Ventana para dibujar dos variables una frente a otra con RCommander.



Figura 5: Ventana para ajustar un modelo de regresión lineal.

EJERCICIO:

Ajusta un modelo de regresión utilizando las mismas variables que en el ejercicio anterior. ¿Qué elementos reconoces de la salida que produce RCommander?

2.3. Transformación de variables

Cuando la relación entre dos variables no es lineal tenemos que transformar los datos para intentar que las nuevas variables sí que tengan una relación lineal entre ellas. Desde **RCommander** podemos añadir nuevas variables al banco de datos con **DATOS->MODIFICAR DATOS DEL CONJUNTO DE DATOS ACTIVO**. En concreto, con la opción **Calcular una nueva variable** podemos añadir una nueva variable que sea función de las variables que ya existen. La ventana que aparece es la que se puede ver en la Figura 6.

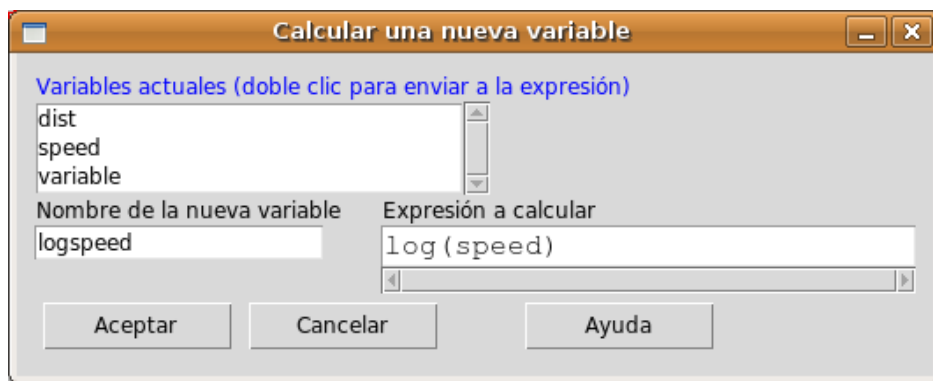


Figura 6: Ventana para transformar variables.

EJERCICIO:

El banco de datos **Animals** (incluido en el paquete **MASS**) contiene datos del peso corporal (**body**) y del cerebro de 28 animales terrestres. Investiga la relación entre las dos variables gráficamente. ¿Es sensato hacer una regresión lineal? En caso de que no lo sea, busca una transformación de los datos que permita hacer un ajuste lineal. Utiliza la variable de peso del cuerpo como independiente y la variable **brain** como dependiente.