CASO PRÁCTICO: ANÁLISIS MULTIVARIANTE.

Vamos a valorar el efecto del trabajo en la mina de una serie de variables que describimos a continuación. El fichero de trabajo es una hoja de Excel que como siempre vamos a importar desde R Commander. Para ello nos vamos a Datos → Importar datos → desde un archivo Excel. Nos aparece una ventana en la que nos pide que demos un nombre al fichero, en nuestro caso le llamaremos "mineros", y a continuación se nos abre el explorador de Windows, buscamos nuestro fichero y damos a aceptar. Veremos el nombre de nuestro fichero en la ventana de "Conjunto de datos".

El fichero contiene las siguientes variables:

ID: número de identificación del registro. No la vamos a utilizar en el estudio, ya que no tiene interés.

EDAD: contiene la edad de los integrantes en el estudio expresada en años

ALTURA: contiene la altura de los participante expresada en cm

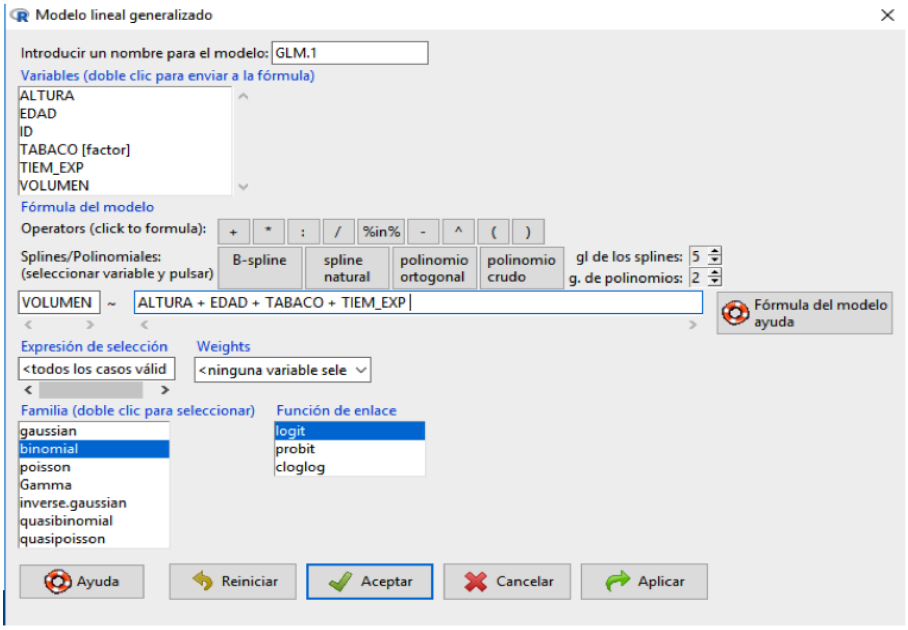
TIEM_EXP: Tiempo trabajado en la mina expresado en años

TABACO: variable cualitativa con tres categorías: No fumadores (que no han fumado nunca), Exfumadores (Que han fumado alguna vez pero que en el momento del estudio no fumaban), y Fumadores en el momento del estudio

VOLUMEN: Volumen espiratorio pulmonar, medido en mL. Es nuestra variable resultado o variable dependiente.

Nuestra hipótesis de trabajo es que pensamos que el volumen espiratorio de los mineros depende de su edad, altura, tiempo de trabajo en la mina y de sus hábitos de tabaco.

Construimos nuestro modelo. Para ello nos vamos a R Commander y en Estadísticos → Ajuste de modelos → Modelo lineal introducimos las variables. La primera como siempre nuestra variable resultado, y después el resto de variables.



```
Residuals:
    Min      1Q  Median      3Q     Max
-1250.74 -367.89   86.43  397.28 1082.10

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3899.471   1867.375  -2.088  0.04008 *
ALTURA             54.662      9.930   5.505  4.68e-07 ***
EDAD              -26.936     10.230  -2.633  0.01023 *
TABACO[T.Fumador] -393.261    148.462  -2.649  0.00979 **
TABACO[T.No fuma] -408.062    195.212  -2.090  0.03989 *
TIEM_EXP          -23.708      9.655  -2.456  0.01632 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 551.5 on 77 degrees of freedom
Multiple R-squared:  0.6868, Adjusted R-squared:  0.6665
F-statistic: 33.78 on 5 and 77 DF,  p-value: < 2.2e-16
```

A la ordenada en el origen no le vamos a prestar atención ya que no tiene mucho valor su dato ya que es difícil de interpretar o la interpretación es absurda. Tenemos dos variables numéricas, en el caso de la altura, por cada cm de altura se incrementa el volumen espiratorio en 54,7 mL, en cambio por cada año de edad el volumen espiratorio desciende en 26,9 mL. En el caso del tiempo trabajado en la mina, por cada año trabajado el volumen espiratorio disminuye en 23,7 mL

En el caso del tabaco nos encontramos ante una variable *dummy*. La referencia que toma R es un poco rara, serían los exfumadores, es decir que el volumen espiratorio disminuye en 408,1 mL de ser no fumador a ser exfumador, e igualmente desciende en 393,1 mL de un fumador respecto de un exfumador.

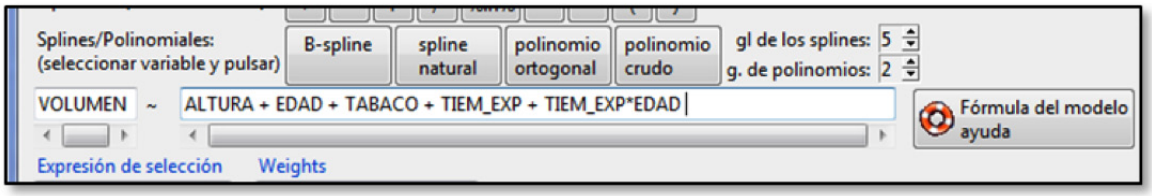
Vamos a cambiar la referencia para que la interpretación de la variable TABACO sea más razonable. Para ello Datos → Modificar variables del conjunto activo → Reordenar niveles de factor, le decimos que sobrescribimos la variable y ponemos como referencia a los no fumadores.

```
TABACO[T.Exfumador]  408.062   195.212  2.090  0.0399 *
TABACO[T.Fumador]     14.801   168.943  0.088  0.9304*
```

Observad como cambian los coeficientes. Ahora el volumen espiratorio disminuye 408,1 en un exfumador respecto de un no fumador y curiosamente en un fumador no hay diferencias significativas respecto al no fumador.

Podemos pensar que a más edad el efecto del tiempo trabajado en la mina sea mayor, es decir puede haber un efecto de interacción. Vamos pues a estudiarlo, para ello construimos un nuevo modelo con el término de interacción que quedaría:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_2 x_2 \times \beta_3 x_3$$



El resultado es el siguiente:

```
Coefficients:
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)      -4400.5786  1843.1907  -2.387  0.0194 *
ALTURA              53.7878     9.9757   5.392 7.58e-07 ***
EDAD               -20.6794    12.1254  -1.705  0.0922 .
TABACO[T.Exfumador] 403.0306   195.3756   2.063  0.0425 *
TABACO[T.Fumador]    23.5881   169.2702   0.139  0.8895
TIEM_EXP              7.0995    33.4361   0.212  0.8324
```

Vemos en último lugar el término de interacción entre la edad y el tiempo trabajado en la mina. Podéis observar que la *p* es muy alta por tanto parece que no existe el término de interacción. Podemos comparar los modelos como ya hemos visto, el resultado sería:

```
> anova(LinearModel.2, LinearModel.4)
Analysis of Variance Table

Model 1: VOLUMEN ~ ALTURA + EDAD + TABACO + TIEM_EXP
Model 2: VOLUMEN ~ ALTURA + EDAD + TABACO + TIEM_EXP + TIEM_EXP * EDAD
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     77 23420369
2     76 23138366  1    282002 0.9263 0.3389
```

Vemos que no hay diferencias significativas entre los dos modelos por tanto la variable de interacción no aporta más información y por lo que la podemos eliminar del estudio.