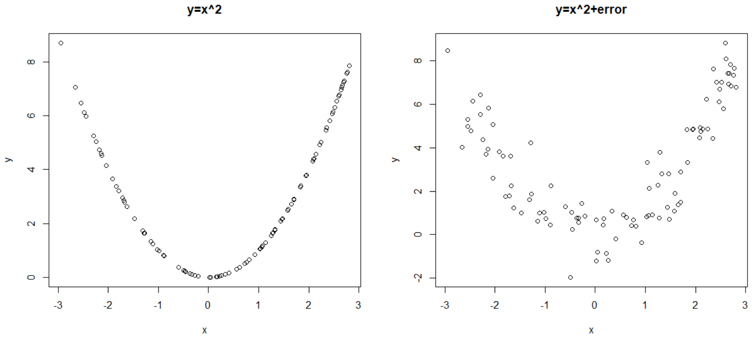
Modelos de regresión lineal

Sesión 5. Curso: 'Análisis estadístico aplicado con R Commander'

Miguel A. Martinez-Beneito
correo: martinez_mig@gva.es

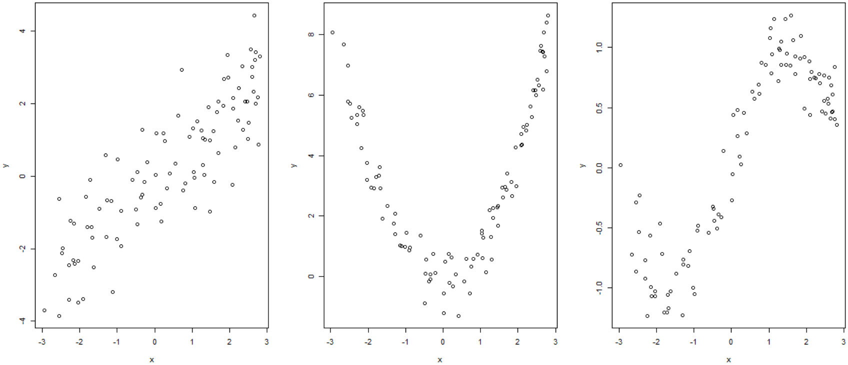
Relación entre variables

Dos variables pueden estar relacionadas de manera <u>determinista</u> o <u>no determinista</u>.



Aunque una relación <u>no sea determinista</u> no quiere decir que ambas variables <u>no estén relacionadas</u> sino que además de la relación entre ellas hay <u>factores distorsionantes</u>.

Dos variables pueden relacionarse de muchas maneras diferentes



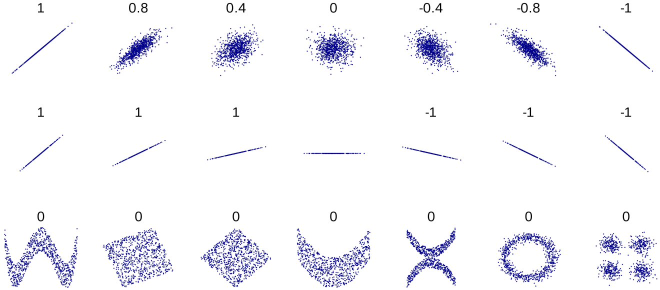
La existencia de relación entre dos variables no implica una relación de causalidad entre ellas sino de asociación entre ambas.

Medición de la linealidad de una relación

Coeficiente de correlación (de Pearson) $\rho_{x,y}$ mide el grado y el sentido de la asociación lineal entre 2 variables x e y.

ρ toma valores en el intervalo [-1,1].

- $\rho \approx 1$ o $\rho \approx -1$ relación lineal determinística, positiva o negativa, respectivamente.
- $\rho \approx 0$ no asociación lineal.



Significación del coeficiente de correlación:

Suele ser habitual plantearse si, para dos variables, ρ es distinto de 0. En ese caso tiene sentido plantear una relación de tipo lineal entre ambas variables.

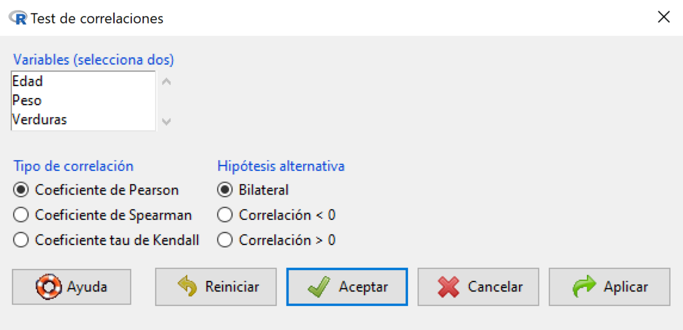
Contraste:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Podemos implementar este contraste de forma sencilla en `R Commander`. Para ello hemos de recurrir al menú `Estadísticos> Resumenes> Test de correlación`

Ejemplo

Para el banco de datos guardado en `Pesos.Rdata` deseamos valorar la significatividad de la correlación entre las variables `Peso` (peso en kg. de un conjunto de niños) y `Verduras` (ingesta diaria (en kg.) de verduras).
Abrimos el menú `Test de correlación`



Seleccionamos las variables `Peso y Verduras` y aceptamos las opciones por defecto

Ejemplo

```
        Pearson's product-moment correlation

data:  Peso and Verduras
t = 2.4442, df = 98, p-value = 0.0163
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04542524 0.41651238
sample estimates:
      cor
0.2397046
```

- Pivote
- P-valor
- Intervalo de confianza (no contiene el valor 0)
- Estimador puntual

Existe una importante relación lineal (positiva) entre ambas variables. Este resultado invita a establecer una relación lineal entre ellas.

El cuadrado del coeficiente de correlación se conoce como *coeficiente de determinación* y tiene una interesante interpretación: ρ^2 es la proporción de varianza de una de las variables que puede ser explicada por la otra (mediante una relación lineal).

Ejemplo

El coeficiente de determinación de las variables estudiadas es $0.240^2 = 0.057$, por tanto la ingesta de verduras explica un 5.8 % del peso de los niños.

Regresión lineal simple, el concepto

Un modelo de regresión lineal simple asume que disponemos de dos variables:

- y (variable dependiente, variable respuesta).
- x (variable independiente, variable explicativa, covariable).

El modelo de regresión lineal simple asume una relación lineal no determinista entre ambas variables -> existe un factor aleatorio adicional que altera la variable respuesta.

El modelo de regresión lineal asume que el valor esperado de la variable respuesta depende de forma lineal de la covariable: $E(y_i) = \beta_0 + \beta_1 x_i$.

El termino de la derecha de la expresión anterior se dice *termino o predictor lineal del modelo*.

Adicionalmente, el modelo de regresión lineal asume que la variable respuesta se distribuye de forma Normal, de media la que acabamos de describir y varianza constante.

En resumen, el modelo de regresión lineal se suele presentar de alguna de las siguientes maneras:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \; i = 1, \cdots, n$$

o

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), \; i = 1, \cdots, n$$

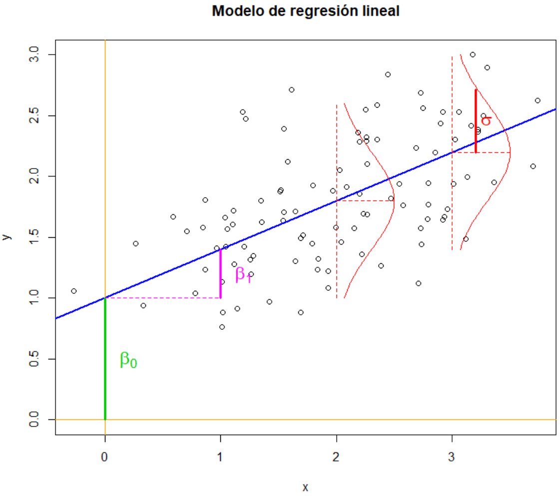
o

$$y_i \sim N(\mu_i, \sigma^2), \; \mu_i = \beta_0 + \beta_1 x_i, \; i = 1, \cdots, n$$

Modelo de regresión lineal, interpretación

β_0 se conoce como el *intercept* y es el valor de la recta de regresión para $x = 0$.

β_1 se conoce como la *pendiene de la recta* (o efecto de x). Se interpreta como la modificación de $E(y)$ por cada unidad de aumento en x.

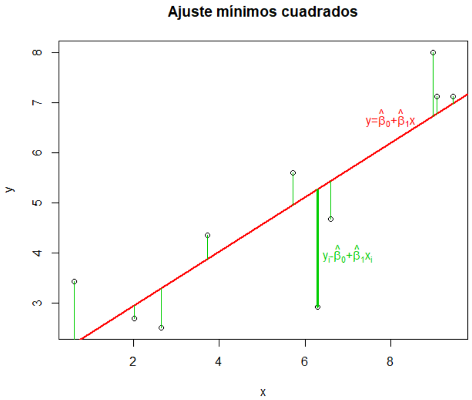


σ es la desviación típica de los datos alrededor de la recta de regresión.

Estimación en el modelo de regresión lineal

Una vez planteado el modelo de regresión lineal, de la forma que hemos introducido, habremos de estimar los valores de β_0, β_1 y σ.

Los parámetros del modelo de regresión se suelen estimar mediante el método de mínimos cuadrados, que estima aquellos parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$.



La recta de regresión que estimemos será óptima en el sentido que será la que pase más cerca de los datos observados.

Además de estimar los parámetros del modelo de regresión lineal, en general estaremos interesados en determinar si éstos son significativamente distintos de cero.

Así, suele ser habitual plantearse contrastes del tipo:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

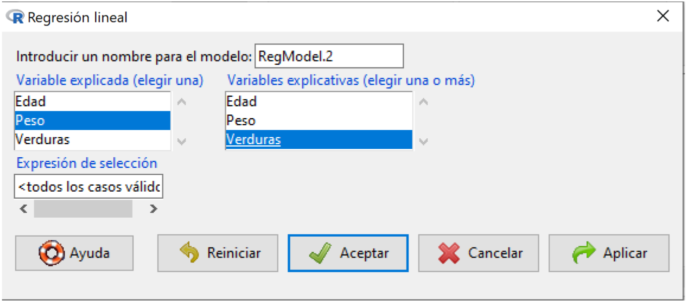
para todo valor de i.

En general la estimación de los coeficientes de regresión se suele acompañar de los P-valores correspondientes a estos contrastes.

Regresión lineal en `R Commander`

El menú `Estadísticos> Ajuste de Modelos> Regresión lineal` permite ajustar modelos de regresión lineal en `R Commander`.

Ese menú abre una ventana como la siguiente donde habremos de determinar la variable respuesta del modelo (varianle explicada) y la(s) variable(s) explicativa(s).



La casilla `nombre` nos permitirá especificar un nombre para el modelo ajustado, que nos será útil para otras opciones de los modelos.

Ejemplo

En el ejemplo del peso y consumo de verduras, obtenemos:

```
> RegModel.2 <- lm(Peso ~ Verduras, data = Pesos)

> summary(RegModel.2)

Call:
lm(formula = Peso ~ Verduras, data = Pesos)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5892 -2.4007 -0.2971  2.3678  7.0790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3984     0.9033  19.261   <2e-16 ***
Verduras      9.0962     3.7215   2.444   0.0163 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.153 on 98 degrees of freedom
Multiple R-squared:  0.05746, Adjusted R-squared:  0.04784
F-statistic: 5.974 on 1 and 98 DF,  p-value: 0.0163
```

Estimadores:

- β_0
- β_1
- σ
- ρ^2

Significación de parámetros del modelo

Con el menú `Modelos> Intervalos de confianza`, donde sólo habremos de completar el nivel de confianza, calcularemos intervalos de confianza para los parámetros del modelo.

```
> Confint(RegModel.2, level = 0.95)
             Estimate    2.5 %   97.5 %
(Intercept) 17.398442 15.60591 19.19097
Verduras     9.096218  1.71096 16.48148
```

Validación del modelo de regresión

El modelo de regresión asume una serie de supuestos que habremos de comprobar para validar el modelo hemos ajustado.

Modelo de regresión lineal:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \; i = 1, \cdots, n$$

Hipótesis del modelo:

- Normalidad: La variable respuesta se debe distribuir de forma Normal alrededor de la recta.
- Homoscedasticidad: La varianza de la variable respuesta alrededor de su media debería ser constante para todas las observaciones.
- Independencia: La variable dependiente, dada la recta de regresión, se distribuye de manera independiente.

La principal herramienta para la validación de las hipótesis del modelo será el análisis de sus residuos.

Se definen los *residuos* del modelo ajustado como la diferencia entre el valor observado de la variable respuesta y el valor ajustado: $R_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

`R Commander` permite el calculo de residuos mediante el uso de comandos de `R` usando la función `resid(modelo)` (donde `modelo` es el nombre del modelo correspondiente).

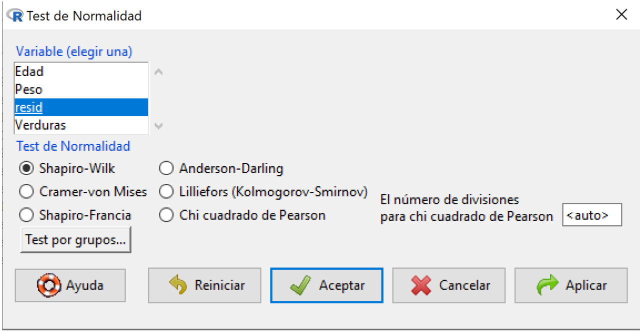
Tras ejecutar un modelo en `R Commander` resulta conveniente guardar sus residuos como una variable adicional del banco de datos.

Ejemplo

En el ejemplo anterior añadimos los residuos como variable del banco de datos de la siguiente manera: `Pesos$resid<-resid(RegModel.2)`
Para que `R Commander` reconozca la nueva variable como una más del banco de datos lo editaremos (botón de la barra de herramientas) y aceptaremos los cambios.

La Normalidad de la variable respuesta, dado el modelo de regresión, se puede valorar mediante un contraste de normalidad sobre los residuos.

Contrastaremos dicha normalidad con el menú `Estadísticos> Resúmenes> Tests de normalidad` de `R Commander`.



De entre las opciones disponibles los tests de Shapiro-Wilks y Kolmogorov-Smirnov son los más comunes.

La hipótesis nula para ambos casos es la Normalidad de la variable analizada.

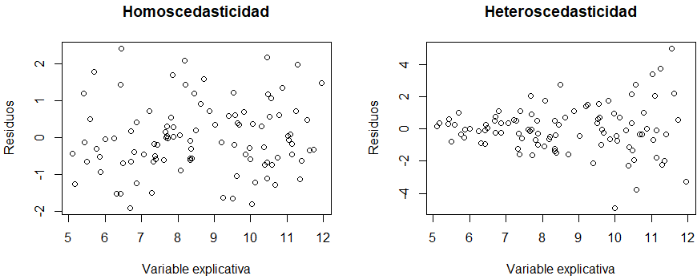
Ejemplo

Valoración de la Normalidad de los residuos en nuestro ejemplo:

```
> normalityTest(~resid, test = "shapiro.test", data = Pesos)

	Shapiro-Wilk normality test

data:  resid
W = 0.98556, p-value = 0.3483


> normalityTest(~resid, test = "lillie.test", data = Pesos)

	Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid
D = 0.070149, p-value = 0.2629
```

Ninguno de los contrastes rechaza la Normalidad de los residuos.

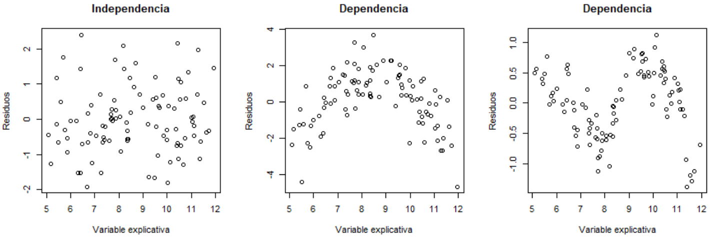
La representación gráfica de los residuos es la herramienta más útil para valorar su homoscedasticidad. Deberíamos comprobar que ningún rango de valores de la variable explicativa tiene menor/mayor varianza que el resto.



Además de la inspección visual de los residuos podremos valorar también su homoscedasticidad mediante el test de Levene, que valora la igualdad de varianzas para grupos definidos según cierta variable categórica.

Para implementar el test de Levene habremos de categorizar previamente la variable explicativa.

La representación gráfica de los residuos también nos permite valorar la independencia de la variable respuesta dada la recta de regresión.



Si los residuos son independientes no deberían mostrar ningún patrón.

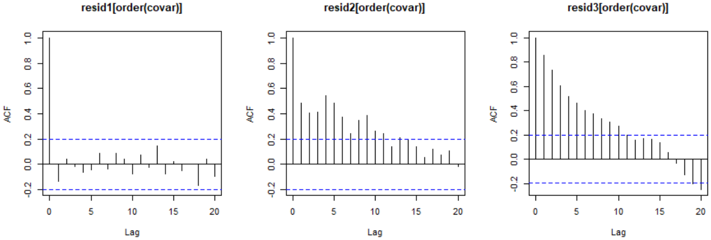
La *función de autocorrelación* (ACF) nos permite cuantificar la dependencia de los residuos.

`R Commander` no dispone de ningún menu que permita el cálculo de la ACF.

Habremos de calcular la ACF desde línea de comando con la expresión: `acf(resid[order(covar)]` donde `resid` es la variable que contiene los residuos y `covar` la covariable que analicemos (sustituir por los valores oportunos).

La ACF mide la correlación de cada observación con la anterior (lag 1) o la anterior de la anterior (lag 2) . . .

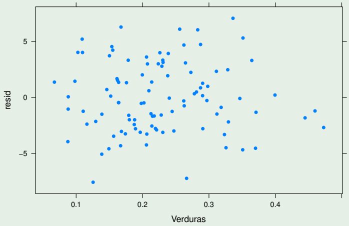
Si los datos son independientes, todas las autocorrelaciones (con lag>1) deberían ser próximas a 0.



Ejemplo

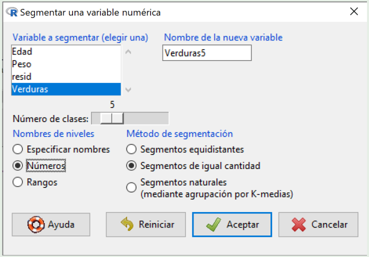
Vamos a valorar las hipótesis de homocedasticidad e independencia del modelo lineal que hemos ajustado. Comenzamos representando los residuos del modelo frenta a la variable explicativa, menú: `Gráficas> Gráfica XY`.

Ejemplo



No parece observarse evidencia de heteroscedasticidad ni dependencia del residuo. Aun así vamos a contrastar dichas hipótesis.

Para contrastar la heteroscedasticidad dividimos la variable `Verduras` siguen sus quintiles, menú `Datos> Modificar ...> Segmentar ...:`



Ejemplo

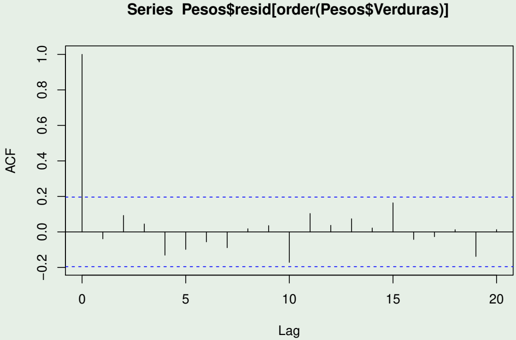
Un vez segmentada la variable podremos comparar las varianzas de los residuos mediante el test de Levene o Barlett, menú: `Estadísticos> Varianzas> Test de Levene` o `Test de Barlet`:

```
> bartlett.test(resid ~ Verduras5, data = Pesos)

        Bartlett test of homogeneity of variances

data:  resid by Verduras5
Bartlett's K-squared = 2.0896, df = 4, p-value = 0.7193


> leveneTest(resid ~ Verduras5, data = Pesos, center = "median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  4  0.4041 0.8053
      95
```

Por último valoraremos la dependencia de los residuos en base a su función de autocorrelación, con la sentencia:

`acf(Pesos$resid[order(Pesos$Verduras)])`.

Ejemplo



Todas las <u>autocorrelaciones</u> caen dentro del intervalo de confianza <u>alrededor de cero</u> por lo que <u>no</u> encontramos <u>evidencia en contra</u> de la <u>independencia</u> de los datos.

Regresión lineal múltiple

En ocasiones querremos:

- Establecer una relación lineal entre la variable respuesta y más de una variable.
- Hacer regresiones lineales controlando el efecto confusor de una tercera variable.

En ese caso la *regresión lineal múltiple* será la técnica de regresión adecuada, a diferencia de la regresión lineal (simple) que hemos visto hasta ahora.

El modelo de regresión lineal múltiple se expresa como:

$$y_i \sim N(\mu_i, \sigma^2),\ \ i = 1, \cdots, n$$

$$\mu_i = \beta_0 + \beta_1 (x_1)_i + \cdots + \beta_p (x_p)_i$$

o

$$y_i = \beta_0 + \beta_1 (x_1)_i + \cdots + \beta_p (x_p)_i + \epsilon_i,\ \ \epsilon_i \sim N(0, \sigma^2)$$

El modelo de regresión lineal múltiple asume una relación lineal entre la variable respuesta y y una serie de variables explicativas $x_1, \cdots, x_p$.

Los parámetros del modelo $\beta_0, \cdots, \beta_p$ se interpretan como los efectos de cada variable, controlando el efecto del resto.

El modelo de regresión lineal puede utilizarse tanto para incluir nuevas variables en el modelo como para incluir funciones no lineales de la variable explicativa original. Así, podremos reproducir relaciones más complejas que una simple relación lineal. Por ejemplo:

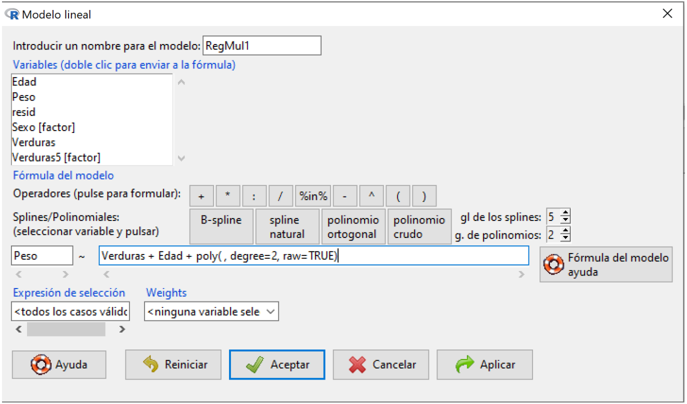
$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

reproduce una relación polínomica (de orden p) entre y y x.

Todas las herramientas (intervalos de confianza, contrastes, análisis de residuos . . .) introducidas para la regresión lineal simple siguen siendo igualmente válidos para la regresión lineal múltiple.

Regresión lineal múltiple en `R Commander`

`R Commander` hace posible el ajuste de modelos de regresión lineal múltiple mediante el menú: `Estadísticos> Ajuste de modelos> Modelo lineal`



En la casilla precedida por $\sim$ introduciremos el predictor lineal de $E(y_i)$. Un simbolo $+$ separará los distintos términos del modelo.

El botón `polinomio crudo` introducirá un término polinómico del grado que queramos para la variable correspondiente. Tanto el grado (`degree`) como la variable que reproduce el polinomio (poner justo antes de la primera coma) habrán de ser especificadas.

Podemos valorar si la variabilidad residual de este modelo es significativamente menor que la del modelo que no tiene ninguna variable explicativa. Podemos hacer esto con un test ANOVA que compara la varianza de ambos modelos. Este test ANOVA se considera un contraste global de ajuste del modelo.

Ejemplo

En el ejemplo que hemos estado estudiando querríamos añadir la variable `Edad` ya que prevemos que pudiera tener un importante efecto confusor.

Simplemente habremos de llenar el formulario tal y como aparece en la imagen anterior quitando el término polinómico del predictor lineal.

Ejemplo

```
> summary(RegMull)

Call:
lm(formula = Peso ~ Verduras + Edad, data = Pesos)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6714 -1.3960 -0.1028  1.2335  4.6685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.6173     1.0019   8.601  1.4e-13 ***
Verduras     -8.5212     2.9648  -2.874  0.00498 **
Edad          1.4827     0.1348  11.001  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.114 on 97 degrees of freedom
Multiple R-squared:  0.5807, Adjusted R-squared:  0.572
F-statistic: 67.16 on 2 and 97 DF,  p-value: < 2.2e-16
```

- Todas las variables tienen un efecto significativo.
- La ingesta de verduras !!reduce ahora el peso corporal de los niños!!
- El error standard residual se reduce y el coeficiente de determinación aumenta claramente respecto al modelo de regresión lineal simple por la variabilidad explicada por la edad.
- El test ANOVA certifica que las variables de este modelo explican una proporción significativa de la variabilidad de Peso.

El chequeo de las hipótesis del modelo es también pertinente ahora, al igual que para la regresión lineal simple.

Sería interesante estudiar los residuos en función tanto de las covariables (¿hay tendencia/dependencia en los residuos?) como del valor del predictor lineal que hemos ajustado (¿son los residuos Normales? ¿hay homoscedasticidad? ¿hay tendencia en los residuos?)

Podemos obtener el predictor lineal del modelo mediante la sentencia `predict(modelo)` donde `modelo` es el nombre del modelo ajustado.

Regresión con covariables categóricas.

De la misma forma que hemos utilizado covariables numéricas, podremos incluir también covariables categóricas en modelos de regresión.

Un modelo con sólo covariables categóricas es equivalente aun modelo ANOVA en el que se valoran diferencias en las media de la variable explicativa para cada grupo definido por la(s) covariable(s). Sin embargo, `R` trata y presenta de forma distinta los resultados en función de si hacemos un ANOVA o una regresión lineal.

La inclusión de variables categóricas en modelos de regresión requiere una recodificación de la variable en variables binarias, tantas como categorías tengamos:

$$\begin{pmatrix} 1 \\ 3 \\ 1 \\ \dots \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ \dots & \dots & \dots \end{pmatrix}$$

Tras esta recodificación procederíamos al ajuste del modelo como si se tratara de un modelo de regresión múltiple normal y corriente.

Sin embargo el modelo correspondiente estaría sobreparametrizado. Por ejemplo, en un modelo con una sóla variable categórica tendríamos que estimar el efecto de sus p categorías y tendríamos $p+1$ valores a estimar: el intercept y las p variables en las que hemos descompuesto la variable original.

En consecuencia, se suele quitar la primera columna (efecto de la primera categoría) de la matriz binaria anterior. Este es el procedimiento que emplea `R Commander`.

Tras esta recodificación, interpretamos:

- β_0 como el efecto de la primera categoría de la variable.
- β_1 el efecto de la segunda categoría de la variable en relación a la primera.
- ...
- β_{p-1} el efecto de la p-ésima categoría en relación a la primera.

Ejemplo

Sobre el último modelo que hemos ajustado vamos a añadir el efecto del sexo. Añadiremos en el término lineal que ya tuviéramos un nuevo término de la forma `+ Sexo`.

En nuestro caso `Sexo` es un factor con categorías: `{Niña, Niño}`.

```
> summary(LinearModel.5)

Call:
lm(formula = Peso ~ Verduras + Edad + Sexo, data = Pesos)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5988 -1.3811 -0.0966  1.2947  4.6932 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    8.6777     1.0240   8.475 2.78e-13 ***
Verduras      -8.3795     3.0109  -2.783  0.00649 ** 
Edad           1.4803     0.1356  10.916  < 2e-16 ***
Sexo[T.Niño]  -0.1384     0.4299  -0.322  0.74816    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.124 on 96 degrees of freedom
Multiple R-squared:  0.5811, Adjusted R-squared:  0.568 
F-statistic: 44.39 on 3 and 96 DF,  p-value: < 2.2e-16
```

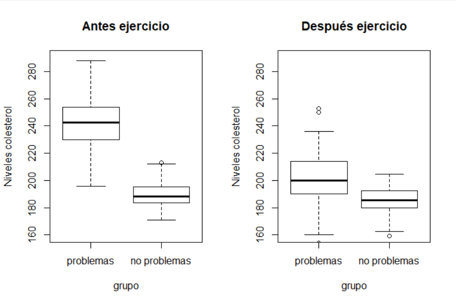
- El peso medio estimado para las niñas es de 8.677 kilos.
- Los niños pesan en promedio -0.1384 kilos menos, aunque la diferencia entre niños y niños no es significativa.
- La inclusión de la variable sexo apenas mejora las medidas de ajuste del modelo.

Interacción entre variables explicativas

Los modelos de regresión múltiple que hemos planteado asumen que el efecto de cualquier covariable es siempre el mismo, independientemente del valor de las otras. Sin embargo, En ocasiones el efecto de una variable se puede modificar según el valor de las otras.

Ejemplo

En un estudio sobre el control del colesterol queremos evaluar el efecto del ejercicio físico sobre esta variable. En el estudio disponemos de dos grupos: personas con y sin problemas de colesterol.



Ejemplo

Un modelo de regresión con covariables `grupo` y `ejercicio` daría un efecto debilmente significativo para `ejercicio`.
Necesitamos un modelo más complejo que describa el efecto distinto de hacer ejercicio para cada grupo de individuos.

Consideramos la *interacción* entre 2 covariables como el efecto que pudiera tener su producto sobre la variable respuesta.

Modelo de regresión lineal con interacción:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i * z_i$$

Podemos entender, alternativamente, la relación anterior como:

$$E(y_i) = \beta_0 + (\beta_1 + \beta_3 * z_i) x_i + \beta_2 z_i$$

donde el efecto de la variable x depende del valor observado de z.

De la misma forma, podemos entender el modelo con interacción como:

$$E(y_i) = \beta_0 + \beta_1 x_i + (\beta_2 + \beta_3 x_i) z_i$$

donde el efecto de la variable z dependerá de x.

En la expresión anterior x (y z) podrán ser tanto variables numéricas como categóricas.

- Si x es continua: el efecto de z variará gradualmente con x, si β_3 es positivo aumentará con x y si es negativo disminuirá.
- Si x es binaria: tendríamos dos efectos para la variable z, una para la primera categoría de x (β_2) y otro para la segunda ($\beta_2 + \beta_3$).

Para introducir una interacción en el predictor lineal en `R Commander` utilizaremos el operador `:`, así `+ var1:var2` introducirá en el modelo la interacción de `var1` y `var2`.

El operador `*` introduce en el modelo tanto las variables originales más su interacción, es decir: `+ var1 * var2` equivale a `+ var1 + var2 + var1:var2`.

Ejemplo

Sobre el modelo de regresión lineal anterior vamos a considerar también la interacción entre edad y sexo para valorar si influye la edad, de la misma forma la edad para ambos sexos, sobre el peso corporal.
Añadimos al predictor lineal el término: `+ Edad : Sexo`.

```
Call:
lm(formula = Peso ~ Verduras + Edad + Sexo + Edad:Sexo, data = Pesos)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1717 -1.2878 -0.1722  1.4242  4.3078

Coefficients:
                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)         5.6967     1.3455   4.234 0.0000531 ***
Verduras           -8.9632     2.8794  -3.113   0.00245 **
Edad                1.8429     0.1715  10.747   < 2e-16 ***
Sexo[T.Niño]        5.9342     1.9282   3.078   0.00273 **
Edad:Sexo[T.Niño]  -0.7028     0.2180  -3.223   0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.027 on 95 degrees of freedom
Multiple R-squared:  0.6224, Adjusted R-squared:  0.6065
F-statistic: 39.15 on 4 and 95 DF,  p-value: < 2.2e-16
```

- Al incluir la interacción edad:sexo todos los parámetros pasan a ser significativos.
- Los niños de pequeños pesan más que las niñas, a los 8.5 años pesan (5.9-0.7*8.5) parecido y de ahí en adelante pesan más las niñas.
- La inclusión de la interacción ha disminuido el error residual y aumenta el coeficiente de determinación.