

Regresión Lineal Múltiple, verificación de los supuestos del modelo y selección de variables

M. A. Ibáñez y J. C. Martínez Ávila
Dept. Economía Agraria, Estadística y Gestión de Empresas
ETSIAAB*

13 de enero de 2026

Índice

1. Regresión Lineal Múltiple	2
1.1. Introducción	2
1.2. Modelo Probabilístico	5
1.3. Estimación de los coeficientes de regresión parcial	8
1.4. Propiedades de los estimadores	13
2. Verificación de los supuestos del modelo de regresión lineal múltiple	17
2.1. Residuos	17
2.2. Propiedades de los residuos.	17
2.3. Gráficos de componentes más residuos	21
2.4. Modelo estimado e interpretación. Visualización	26
3. Validación cruzada	32
4. Selección de variables	34
4.1. Ajustar un modelo por cada variable	37
4.2. Selección de variables. Modelo más parsimonioso	39
4.3. Inclusión de variables cualitativas en el modelo	43

*Este documento está disponible bajo licencia Creative Commons. Reconocimiento - NoComercial - CompartirIgual.
<http://creativecommons.org/licenses/by-sa/3.0>

1. Regresión Lineal Múltiple

1.1. Introducción

Ampliaremos el modelo de regresión lineal simple, estudiado en otras asignaturas de estadística general, incorporando dos o más variables explicativas que tengan una relación lineal con la variable respuesta.

Estudiaremos como especificar e interpretar un modelo de regresión lineal múltiple (MRLM). Estudiaremos como interpretar los parámetros del modelos de regresión lineal múltiple y como comprobar los supuestos en los que se apoya el modelo de regresión múltiple.

A lo largo del tema utilizaremos un ejemplo de las emisión de metano en purines.

Ejemplo. Emisión de metano en purines

```
> library(readxl)
> Purines <- read_xlsx("Purines_Metano2.xlsx")
> Purines <- as.data.frame(Purines)
```

Tenemos 79 granjas de cerdos comerciales de las que se toma una muestra del purín producido, y en ella se determinan analíticamente la cantidad de metano que emite (mL/gr de sólidos volátiles) y distintas variables sobre la composición del purín con objeto de predecir la cantidad de metano emitido por el purín en función de su composición química.

En este tema vamos a utilizar el archivo `Purines_Metano2.xlsx`. Hay que descargarlo y guardarlo en una carpeta del PC.

Buscaremos una relación entre el metano emitido `CH4`, el contenido de grasa (`EE`) del purín, la fibra neutro detergente `FND` y la lignina ácido detergente `LAD`.

La variable `FND`, fibra neutro detergente, determina el contenido en hemicelulosa, celulosa y lignina del purín, expresado como porcentaje sobre materia seca.

La variable `LAD`, lignina ácido detergente, determina la cantidad de lignina del purín, expresado como porcentaje sobre materia seca.

Disponemos de 79 muestras de purines (explotaciones) seleccionadas al azar en las que se estudian cuatro caracteres: El contenido de grasa (`EE`), de fibra (`FAD`), de lignina (`LAD`) en el purín, y la cantidad de metano (`CH4`) que emite dicha muestra. Cada explotación tiene cuatro valores $(x_{1i}, x_{2i}, x_{3i}, y_i)$ $i = 1, 2, \dots, 79$.

En primer lugar realizamos un resumen numérico de las 4 variables. Para ello usaremos la función `summary`

El resultado es tal y como se muestra a continuación.

	mean	sd	IQR	0%	25%	50%	75%	100%	n
CH4	253.830707	145.416556	227.412489	10.3484879	131.372489	249.054323	358.784979	755.76818	79
EE	10.268617	4.202388	5.642667	2.9300000	6.908885	9.915307	12.551552	24.92174	79
FND	35.230312	12.666784	19.101803	3.7891055	25.873852	35.647972	44.975656	59.62637	79
LAD	7.246044	3.511871	4.563025	0.3905799	4.867909	6.765189	9.430934	18.30531	79

En la figura 1 se muestra un matriz con los gráficos de dispersión dos a dos entre las cuatro variables analizadas.

Para obtener dicho gráfico la forma más rápida es usar la función `plot` sobre el marco de datos que tenemos, la función `pairs` o el paquete `car` y su función `scatterplotMatrix(CH4+FND+LAD+EE, data=Purin` con las opciones `diagonal=list(method="boxplot")`

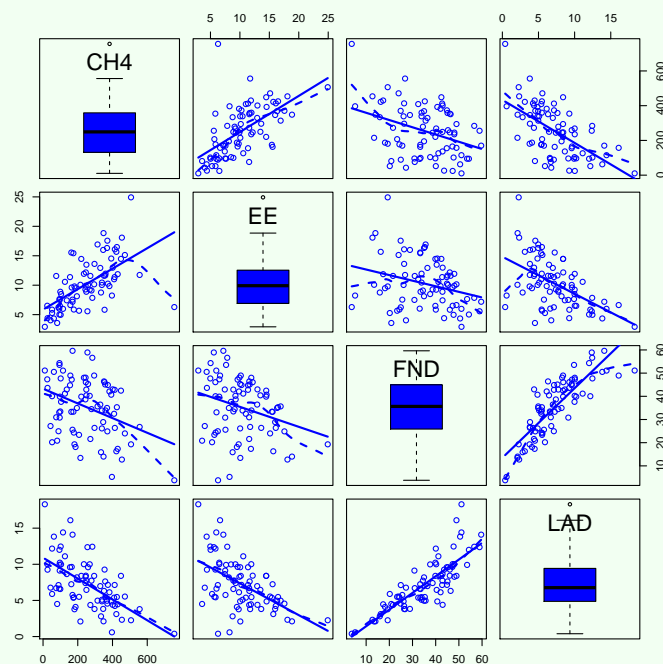


Figura 1: Matriz de gráficos de dispersión de las variables CH4, EE, FND y LAD de las 79 explotaciones analizadas

En la diagonal principal de esta matriz de gráficos se ha representado los gráficos de cajas y bigotes de cada una de las variables (tal y como hemos seleccionado). Fuera de la diagonal principal, se muestran los gráficos de dispersión por pares de variables. Cada fila y cada columna representa a una de las 4 variables.

Así, por ejemplo, el gráfico de la primera fila (CH4) y segunda columna (EE), muestra el gráfico de dispersión de las variable CH4 frente a la variable EE, que realizamos en el tema anterior. Tal y como le hemos indicado a R, para cada gráfico de dispersión, muestra la recta de regresión estimada (línea azul continua) y la tendencia obtenida con una regresión no paramétrica (línea azul discontinua).

De esta forma, la primera fila muestra los gráficos de dispersión de la variable CH4 frente a cada una de las tres variables explicativas. Observamos una relación positiva de las emisiones de metanos con el contenido en grasa, pero negativa con el contenido en fibra y lignina.

En la segunda fila de la matriz de dispersión se representan los gráfico de dispersión de la variable EE frente a CH4 (primera columna), el de la variable EE frente a FND (tercera columna) y el de la variable EE frente a LAD. El resto de filas y columnas de la matriz de gráficos se interpreta de la misma forma.

Es de destacar la relación positiva entre las variable FND y LAD. Esto es debido a que la variable

fibra neutro detergente está determinando, entre otros tipo de fibras, también el contenido de lignina en el purín.

Esto lo podemos comprobar calculando la correlación no sólo entre la variable dependiente y las explicativas sino entre las propias variables explicativas.

```
> round(cor(Purines), 2)
```

	FND	LAD	EE	CH4
FND	1.00	0.83	-0.29	-0.36
LAD	0.83	1.00	-0.53	-0.60
EE	-0.29	-0.53	1.00	0.60
CH4	-0.36	-0.60	0.60	1.00

1.2. Modelo Probabilístico

En el MRLM, asumimos que existen p variables explicativas X_1, X_2, \dots, X_p que están relacionadas con la variable respuesta Y .

La matriz de datos es una matriz con n filas y $p + 1$ columnas $(x_{1i}, x_{2i}, \dots, x_{pi}, y_i)$, $i = 1, 2, \dots, n$.

La notación x_{ji} hace referencia a la observación $i = 1, 2, \dots, n$ de la variable explicativa $j = 1, 2, \dots, p$.

Se asume que existe la siguiente **relación lineal** entre la variable Y y las variables explicativas X_j , $j = 1, 2, \dots, p$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i \quad i = 1, 2, \dots, n \quad (1)$$

donde e_i son realizaciones independientes de una normal de media 0 y varianza σ^2 que no depende de los valores de x_j , $j = 1, 2, \dots, p$.

A los valores e_i , $i = 1, 2, \dots, n$ se les denomina términos del error o perturbación.

La función de regresión asumida implica que:

La media de la variable respuesta Y para diferentes valores de las variables explicativas es:

$$\begin{aligned} \mu_{Y.X} &= E[Y|X_1 = x_{1i}, X_2 = x_{2i}, \dots, X_p = x_{pi}] \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} \quad i = 1, 2, \dots, n \end{aligned} \quad (2)$$

y la varianza de la variable respuesta Y para diferentes valores de las variables explicativas es:

$$var[Y|X_1 = x_{1i}, X_2 = x_{2i}, \dots, X_p = x_{pi}] = \sigma^2 \quad (3)$$

Esto es equivalente a afirmar que las observaciones y_i son realizaciones independientes de una normal cuya media queda determinada por la función de regresión en (2) y su varianza por la función de varianza en (3).

$$Y_i \in N(\mu_{Y.X}; \sigma) \quad i = 1, 2, \dots, n$$

β_0 es el término independiente o **intercept** de la función de regresión y representa el valor medio de la variable explicada Y cuando todas las variables explicativas tienen el valor 0.

A los parámetros β_j , $j = 1, \dots, p$ se les denomina **coeficiente de regresión parcial** y representa cuanto se incrementa o reduce la variable respuesta Y por un incremento en una unidad de la variable explicativa X_j , **manteniendo constante el resto de variables explicativas** (por eso el nombre parcial).

Es importante destacar que en la definición del coeficientes de regresión parcial, va implícita la dependencia de este del resto de variables explicativas en el modelo. De forma que, el coeficiente

de regresión β_2 en un modelo que contiene a las variables explicativas X_1 , X_2 y X_3 es distinto del coeficiente β_2 en un modelo que contiene a las variables explicativas X_1 , X_2 , X_3 y X_4 .

En general, el coeficiente de regresión de una regresión lineal simple entre la variable Y y la variable X_j , no coincidirá con el coeficiente de regresión parcial de la variable X_j cuando otras variables explicativas están incluidas en el modelo.

Si expandimos el modelo en (1) para las n observaciones obtenemos las siguientes n ecuaciones:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{p1} + e_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_p x_{p2} + e_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_p x_{pn} + e_n \end{aligned}$$

que podemos expresar de forma más compacta utilizando la siguiente notación matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

donde:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}_{n \times (p+1)}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n \times 1}$$

\mathbf{y} : es el vector columnas de orden $n \times 1$ con las n observaciones de la variable respuesta Y .

\mathbf{X} : es la matriz de orden $n \times (p+1)$ que contiene una columna de unos seguida de p vectores columnas con el valor de las p variables explicativas en las n observaciones.

$\boldsymbol{\beta}$: es el vector columna de orden $(p+1) \times 1$ con los $p+1$ coeficientes de regresión.

\mathbf{e} : es el vector columna aleatorio de orden $n \times 1$ con las n perturbaciones o errores.

Ejemplo. Emisiones de metano en purines

A partir del conjunto de datos **Purines** con los 79 explotaciones vamos a establecer el modelo de regresión lineal múltiple entre la variable respuesta **CH4** y las variables explicativas **EE**, **FND** y **LAD**.

El vector **y** tendrá 79 filas y 1 columna. La matriz **X** tendrá 79 filas y cuatro columnas: la primera será una columna de unos, la segunda serán los valores de las 79 explotaciones para la variable **EE**, la tercera los 79 valores para la variable **FND** y la cuarta con los 79 valores de la variable **LAD**.

A continuación se muestra las 5 primeras filas: del vector **y**, de la matriz **X** y el vector **β** con los 4 coeficientes de regresión que se corresponden con las 5 primeras observaciones del conjunto de datos **Purines**:

	FND	LAD	EE	CH4
1	17.341	4.4647	11.8732	429.91
2	44.773	9.5933	11.3771	155.30
3	26.373	4.2359	16.0940	413.37
4	51.101	14.4115	5.9854	115.06
5	59.626	14.1016	7.1883	169.95

$$\mathbf{y} = \begin{bmatrix} 429.90588 \\ 155.29569 \\ 413.36533 \\ 115.05719 \\ 169.94931 \\ . \\ . \\ . \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 11.87317 & 17.34117 & 4.46474 \\ 1 & 11.37714 & 44.77334 & 9.59333 \\ 1 & 16.09403 & 26.37251 & 4.23589 \\ 1 & 5.98538 & 51.1013 & 14.41146 \\ 1 & 7.18825 & 59.62637 & 14.10164 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Las 5 primera fila representa las 5 primeras ecuaciones del MRLM:

$$429.90588 = \beta_0 + 11.87317\beta_1 + 17.34117\beta_2 + 4.46474\beta_3 + e_1$$

$$155.29569 = \beta_0 + 11.37714\beta_1 + 44.77334\beta_2 + 9.59333\beta_3 + e_2$$

$$413.36533 = \beta_0 + 16.09403\beta_1 + 26.37251\beta_2 + 4.23589\beta_3 + e_3$$

$$115.05719 = \beta_0 + 5.98538\beta_1 + 51.1013\beta_2 + 14.41146\beta_3 + e_4$$

$$169.94931 = \beta_0 + 7.18825\beta_1 + 59.62637\beta_2 + 14.10164\beta_3 + e_5$$

1.3. Estimación de los coeficientes de regresión parcial

Al igual que en el modelo de regresión lineal simple, utilizaremos mínimos cuadrados ordinarios (MCO) para estimar el vector de parámetros β .

El estimador mínimo cuadrado de β es aquel vector que minimice la suma de cuadrados de los residuos:

$$e_i = y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}$$

$$\begin{aligned}\Phi &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi})^2 = (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta) = \\ &= \mathbf{e}'\mathbf{e} = \sum_{i=1}^n e_i^2\end{aligned}$$

donde $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$

\mathbf{e}' indica la traspuesta del vector \mathbf{e} (el vector columna $n \times 1$ se convierte en el vector fila $1 \times n$).

Los valores del vector β que minimizan la expresión anterior se obtienen resolviendo un sistema de $p + 1$ ecuaciones lineales al que se denomina **ecuaciones normales**.

La expresión matricial del sistema de ecuaciones normales es:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y} \quad (4)$$

donde \mathbf{b} es el estimador MCO de β y:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \cdots & \sum_{i=1}^n x_{pi} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{pi} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 & \cdots & \sum_{i=1}^n x_{2i}x_{pi} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{pi} & \sum_{i=1}^n x_{1i}x_{pi} & \sum_{i=1}^n x_{2i}x_{pi} & \cdots & \sum_{i=1}^n x_{pi}^2 \end{bmatrix}_{(p+1) \times (p+1)}$$

La matriz $\mathbf{X}'\mathbf{X}$ es simétrica, los elementos de la diagonal principal son las sumas de los cuadrados de los elementos de cada columna de la matriz \mathbf{X} y fuera de la diagonal son las sumas de los productos cruzados entre los elementos de cada columna de \mathbf{X} .

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \dots \\ \dots \\ \dots \\ \sum_{i=1}^n x_{pi}y_i \end{bmatrix}_{(p+1) \times 1}$$

El primer elemento del vector $\mathbf{X}'\mathbf{y}$ es la suma de las observaciones y_i consecuencia de multiplicar el vector de unos de la primera columna de \mathbf{X} por el vector \mathbf{y} . El resto de elementos del vector $\mathbf{X}'\mathbf{y}$ es la suma de los productos cruzados entre las variables explicativas x_j y la variable y_i como consecuencia de multiplicar cada columna de la matriz \mathbf{X} por el vector \mathbf{y} .

La solución al sistema de ecuaciones normales en (4) se obtiene de la siguiente forma:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (5)$$

donde $(\mathbf{X}'\mathbf{X})^{-1}$ es la inversa de la matriz $\mathbf{X}'\mathbf{X}$.

Para que la matriz tenga inversa, su rango ha de ser $p + 1$, es decir las columnas de la matriz \mathbf{X} deben ser linealmente independientes, o lo que es lo mismo, no debe ser posible expresar ninguna columna de \mathbf{X} como combinación lineal del resto de columnas de \mathbf{X} .

Ejemplo. Emisiones de metano en purines

Continuando con el ejemplo para establecer el modelo de regresión lineal múltiple entre la variable CH₄ y las variables EE, FND y LAD. Para obtener la matriz \mathbf{X} es necesario utilizar la función `model.matrix` y recordar que el producto de matrices en R es `%*%`

El vector $\mathbf{X}'\mathbf{y}$ y la matriz $\mathbf{X}'\mathbf{X}$ tendrán la siguiente forma:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 79 & 811.22073 & 2783.19468 & 572.43746 \\ 811.22073 & 9707.59957 & 2.73916 \times 10^4 & 5271.82373 \\ 2783.19468 & 2.73916 \times 10^4 & 1.10568 \times 10^5 & 2.30557 \times 10^4 \\ 572.43746 & 5271.82373 & 2.30557 \times 10^4 & 5109.89925 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2.00526 \times 10^4 \\ 2.34573 \times 10^5 \\ 6.54222 \times 10^5 \\ 1.215 \times 10^5 \end{bmatrix}$$

La inversa de la matriz $\mathbf{X}'\mathbf{X}$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.29312 & -0.01416 & -0.00118 & -0.01293 \\ -0.01416 & 0.00112 & -1.84797 \times 10^{-4} & 0.00126 \\ -0.00118 & -1.84797 \times 10^{-4} & 2.90759 \times 10^{-4} & -9.89536 \times 10^{-4} \\ -0.01293 & 0.00126 & -9.89536 \times 10^{-4} & 0.00481 \end{bmatrix}$$

y al multiplicar esta matriz por el vector $\mathbf{X}'\mathbf{y}$ obtenemos el vector de soluciones \mathbf{b}

$$\begin{bmatrix} 0.29312 & -0.01416 & -0.00118 & -0.01293 \\ -0.01416 & 0.00112 & -1.84797 \times 10^{-4} & 0.00126 \\ -0.00118 & -1.84797 \times 10^{-4} & 2.90759 \times 10^{-4} & -9.89536 \times 10^{-4} \\ -0.01293 & 0.00126 & -9.89536 \times 10^{-4} & 0.00481 \end{bmatrix} \begin{bmatrix} 2.00526 \times 10^4 \\ 2.34573 \times 10^5 \\ 6.54222 \times 10^5 \\ 1.215 \times 10^5 \end{bmatrix} =$$

$$= \begin{bmatrix} 217.14257 \\ 11.77139 \\ 3.06856 \\ -26.5378 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 217.14257 \\ 11.77139 \\ 3.06856 \\ -26.5378 \end{bmatrix}$$

Con el resultado obtenido, la ecuación de regresión estimada es:

$$\widehat{CH_4} = 217.14 + 11.77EE + 3.07FND - 26.54LAD$$

El coeficiente de 11.77139, estima un incremento de 11.77139 unidades en la cantidad media de emisiones de CH₄ por cada incremento en una unidad de EE en la composición del purín, manteniendo constante el contenido de FND y LAD en la composición del purín.

Es decir, para purines con un contenido de, por ejemplo, 30 % de FND y 5 % de LAD, un incremento en 1 unidad en el contenido en grasa supone un incremento de 11.77139 unidades en la cantidad media de metano emitido.

Este incremento es el mismo si el purín tiene un contenido de 35 % de FND y 6 % de LAD .

Es como si se hubiese estimado un modelo de regresión lineal simple entre CH₄ y EE para una población purines con el mismo contenido de FND y LAD y se considera que la recta de regresión es válida para cualquiera de los contenidos de FND y LAD de los purines.

El valor 217.14257 del término independiente se interpreta como las emisiones medias de metano cuando el contenido de grasa, fibra neutro detergente y lignina en los purines es 0.

El resumen estadístico que hemos realizado de las variables explicativas, nos muestra en qué rango de valores se ha estimado el MRLM y es solo en esa región para la que el modelo es válido, siendo arriesgado extrapolar los resultados fuera de dicha región, por ejemplo cuando las tres variables explicativas se hacen 0.

A diferencia del MRLS, en el que podíamos dibujar el gráfico de dispersión y la recta de regresión estimada, en el modelo de RLM es más complicado visualizar el modelo de regresión estimado ya que ahora no nos movemos en el plano sino que la representación sería en el espacio de $p + 1$ dimensiones.

A modo de ejemplo, en la figura 2 se muestra cómo sería el plano de regresión estimado si hubiésemos estimado un modelo de regresión lineal con solo las variables EE y FND como variables explicativas. En este caso estamos en el espacio de tres dimensiones.

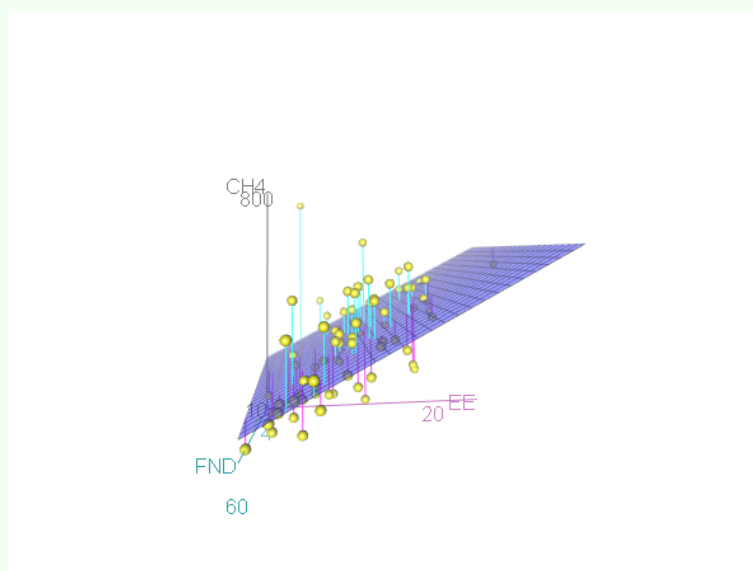


Figura 2: Plano de regresión estimado en el MRLM que relaciona la variable CH₄ con las variables EE y FND

Ese gráfico se ha generado,

```
> library(rgl)
> library(nlme)
> library(mgcv)
> scatter3d(CH4 ~ EE + FND, data = Purines, surface = TRUE, residuals = TRUE,
+          bg = "white", axis.scales = TRUE, grid = TRUE, ellipsoid = FALSE)
```

Interpretación de los coeficientes de regresión

En la tabla 1 se muestran las estimas de los coeficientes de regresión parcial del modelo de regresión lineal múltiple que relaciona la variable CH₄ con las variables explicativas EE, FND y LAD y los tres modelos de regresión lineal simple en los que solo se utiliza como variable explicativa una de ellas.

Podemos observar, que los coeficientes que multiplican a la variable LAD en el modelo de regresión lineal simple con la variable LAD y en el modelo de regresión múltiple son similares.

Sin embargo, la interpretación de dichos coeficientes es distinta. En el MRLS con la variable LAD, el valor de -24.74 indica una reducción en valor medio de la emisiones de metano por un incremento de una unidad en la cantidad de lignina en el purín, ignorando el contenido de grasa y fibra en el mismo.

El valor de -26.54 en el MRLM, indica la reducción en valor medio de la emisiones de metano por un incremento de una unidad en la cantidad de lignina en el purín manteniendo constante el contenido de grasa y fibra en el mismo.

Las diferencias son más evidentes en los coeficientes que multiplican a la variable FND.

Cuando lo estimamos en el MRLM teniendo en cuenta el contenido en grasa y lignina del purín, el valor 3.07 indica un incremento de 3.07 en valor medio de la emisiones de metano por un incremento de una unidad en la cantidad de fibra en el purín manteniendo constante el contenido de grasa y lignina en el mismo.

Por el contrario, la estima del coeficiente de regresión es de -4.17 cuando ignoramos el contenido de grasa y lignina del purín. En este caso, hay una reducción de 4.17 unidades en valor medio de la emisiones de metano por un incremento de una unidad en la cantidad de fibra en el purín.

Esta estima, no mantiene constante el contenido en grasa y lignina del purín y por tanto, no tiene en cuenta que existe una relación positiva entre el la cantidad de fibra y el contenido de lignina en el purín, tal y como veíamos en el gráfico de dispersión de las variables FND y LAD. Al incrementar el contenido de fibra del purín, también se incrementa el contenido de lignina y por tanto el efecto estimado de la fibra sobre las emisiones no es solo de la fibra sino que también lo es de la lignina.

	EE	FND	LAD
RLM con EE, FND y LAD	11.77	3.07	-26.54
RLS con EE	20.81		
RLS con FND		-4.17	
RLS con LAD			-24.74

Cuadro 1: Estimaciones de los coeficientes de regresión de cada una de las variables explicativas en el MRLM con las tres variables explicativas y en los MRLS con cada una de las variables explicativas

1.4. Propiedades de los estimadores

Como ocurre en la regresión lineal simple, los estimadores MCO b_j contenidos en \mathbf{b} , si los supuestos asumidos en el modelo 1 son correctos, también tienen en el muestreo una distribución normal.

Se verifica que:

$$b_j \in N(\beta_j; \sigma\sqrt{c_{jj}}) \quad j = 0, 1, \dots, p$$

siendo c_{jj} el elemento j -ésimo de la diagonal de la matriz \mathbf{C}

donde $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$.

La diagonal principal de la matriz \mathbf{C} nos proporciona las varianzas de los estimadores y fuera de la diagonal las covarianzas entre los estimadores.

$$\text{var}(b_j) = \sigma^2 c_{jj} \quad j = 0, 1, \dots, p$$

$$\text{cov}(b_j, b_k) = \sigma^2 c_{jk} \quad j = 0, 1, \dots, p \quad k = 0, 1, \dots, p$$

siendo c_{jj} el elemento j -ésimo de la diagonal de la matriz \mathbf{C} y c_{jk} el elemento de la fila j y la columna k de la matriz \mathbf{C} .

A partir de los estimadores \mathbf{b} podemos obtener la estima del vector de medias condicional:

$$\hat{\mu}_{Y.X} = \hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

El vector de residuos del modelo estimado es:

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Cada elemento $\hat{e}_i = y_i - \hat{y}_i$ del vector de residuos $\hat{\mathbf{e}}$ es la diferencia entre la observación y la media estimada con el modelo de regresión.

$$y_i - \hat{y}_i = y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_px_{pi}) = y_i - \mathbf{X}_i\mathbf{b}$$

siendo \mathbf{X}_i la fila i -ésima de la matriz \mathbf{X} .

La suma de los residuos elevados al cuadrado puede expresarse:

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \\ &= (\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}) = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \end{aligned}$$

Bajo los supuestos de normalidad, independencia y homogeneidad de varianzas, se verifica que el estimador de σ^2 es el cuadrado medio residual (MSE):

$$MSE = \frac{SSE}{n - (p + 1)} = \frac{(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{n - (p + 1)} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n - p - 1}$$

es un estimador insesgado que tiene una distribución proporcional a una chi cuadrado con $n - p - 1$ grados de libertad.

La cantidad $p + 1$ es por el número de parámetros que se estiman en el modelo de regresión lineal múltiple.

$$\frac{(n - p - 1)MSE}{\sigma^2} \in \chi^2_{n-p-1}$$

Se verifica además, que los estimadores \mathbf{b} son independientes del MSE.

Ejemplo. Emisión de metano en purines

En R estimamos un modelo de regresión lineal múltiple de la misma forma que cuando estimamos el MRLS, con la función `lm`. CH4 es la variable dependiente y las variables explicativas EE, FND y LAD. El resultado lo deberemos guardar en un objeto, que al llamarlo con la función `summary` genera la siguiente salida.

```
Call:
lm(formula = CH4 ~ EE + FND + LAD, data = Purines)

Residuals:
    Min       1Q   Median       3Q      Max
-194.51  -68.41   -3.94   53.05  463.38

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  217.143     57.319   3.788 0.000304 ***
EE           11.771      3.546   3.319 0.001396 **
FND           3.069      1.805   1.700 0.093317 .
LAD          -26.538      7.340  -3.616 0.000540 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 105.9 on 75 degrees of freedom
Multiple R-squared:  0.4903, Adjusted R-squared:  0.4699
F-statistic: 24.05 on 3 and 75 DF, p-value: 5.226e-11
```

En el encabezado **Coefficients**: Las columnas **Estimate** y **Std. Error** proporciona la estima y el error estándar de cada uno de los coeficientes de regresión del modelo (uno por cada fila). La fila **(Intercept)** se refiere al coeficiente β_0 , la fila **EE** al coeficiente de regresión β_1 que multiplica a la variable **EE**, la fila **FND** al coeficiente de regresión β_2 que multiplica a la variable **FND** y la fila **LAD** al coeficiente de regresión β_3 que multiplica a la variable **LAD**. Observamos que coinciden con las estimas que habíamos obtenido operando con las matrices.

Las columnas **t value** y **Pr(>|t|)** nos proporcionan el valor del estadístico T y el p_{val} para

testar las hipótesis de nulidad de cada uno de los parámetros. Podemos comprobar que el valor del estadístico T es el cociente entre la estima y el error estándar.

En este ejemplo, para el parámetro β_2 , $t \text{ value}=1.69977$ y $\Pr(>|t|)=0.09332$ nos indica que no podemos afirmar que β_2 sea distinta de cero. Sin embargo, para el coeficiente de regresión β_1 , $t \text{ value}=3.31917$ y $\Pr(>|t|)=0.0014$ nos permite afirmar que β_1 es distinto de cero.

Lo mismo ocurre con el test de hipótesis para el parámetro β_3 , $t \text{ value}=-3.61563$ y $\Pr(>|t|)=5.40162 \times 10^{-4}$ nos permite afirmar que β_3 es distinto de cero.

Podemos obtener los intervalos de confianza de los parámetros de la recta de regresión utilizando la función `confint` sobre el objeto que contiene el modelo. El nivel de confianza por defecto es 0.95.

	Estimate	2.5 %	97.5 %
(Intercept)	217.142566	102.9576717	331.327459
EE	11.771395	4.7064297	18.836360
FND	3.068558	-0.5277308	6.664847
LAD	-26.537800	-41.1593303	-11.916269

La columna **Estimate** proporciona las estimas de los dos coeficientes de regresión. Las columnas **2.5 %** y **97.5 %** nos muestran los límites inferior y superior, respectivamente, del intervalo de confianza para cada uno de los coeficientes de regresión.

Observamos que el intervalo de confianza del coeficiente de regresión que multiplica a la variable **FND** contiene al valor 0 y por eso no rechazamos la hipótesis nula de que $\beta_2 = 0$.

Sin embargo, el intervalo de confianza de los coeficiente de regresión que multiplican a las variable **EE** y **LAD** no contienen al valor 0 y podemos afirmar, con un nivel de confianza del 95 %, que su valor es cualquiera de los valores comprendidos en el intervalo.

En el caso de la variable **EE** el rango de valores son positivos, indicando una asociación positiva entre las variables **CH4** y **EE** y pudiendo afirmar que, por cada unidad que incrementamos el contenido de grasa en los purines, el metano emitido se incrementa entre 4.71 y 18.84 unidades.

Para la variable **LAD** el rango de valores son negativos, indicando una asociación negativa entre las variables **CH4** y **LAD** y pudiendo afirmar que, por cada unidad que incrementamos el contenido de lignina en los purines, el metano emitido se reduce entre -41.16 y -11.92 unidades.

Ejemplo. Emisiones de metano en purines

Cuando hemos estimado el modelo de regresión **ModReg1**, con la función `summary` tenemos la información sobre el MSE, el coeficiente de determinación y el valor del estadístico F para el test de hipótesis de que los coeficientes de regresión son todos 0.

Volvemos a obtener la misma información, tal y como se muestra a continuación:

```

Call:
lm(formula = CH4 ~ EE + FND + LAD, data = Purines)

Residuals:
    Min       1Q   Median       3Q      Max
-194.51  -68.41   -3.94   53.05  463.38

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  217.143     57.319   3.788 0.000304 ***
EE           11.771      3.546   3.319 0.001396 **
FND           3.069      1.805   1.700 0.093317 .
LAD          -26.538      7.340  -3.616 0.000540 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 105.9 on 75 degrees of freedom
Multiple R-squared:  0.4903, Adjusted R-squared:  0.4699
F-statistic: 24.05 on 3 and 75 DF, p-value: 5.226e-11

```

El valor mostrado como **Residual estándar error**:105.87084 es la raíz cuadrada del MSE que tiene 75 grados de libertad. Si lo elevamos al cuadrado y lo multiplicamos por sus grados de libertad obtenemos la suma de cuadrados residual (SSE).

$$SSE = MSE \times gl = 105.87084^2 \times 75 = 8.40648 \times 10^5$$

El valor mostrado como **Multiple R-squared**: 0.49033 se refiere al valor del coeficiente de determinación, R^2 . Es decir, el modelo de regresión estimado explica el 49.03 % de la variabilidad de la variable CH4. El resto, es variabilidad no explicada, residual (variación de las observaciones alrededor del modelo estimado).

El valor mostrado como **Adjusted R-squared**:: 0.46994 se refiere al valor del coeficiente de determinación ajustado.

El valor mostrado como **F-statistic**: 24.05106 se refiere al valor del estadístico F de la tabla ANOVA. Este estadístico es el que utilizamos para realizar el test de hipótesis:

$$\begin{aligned}
 H_0 &: \beta_1 = \beta_2 = \beta_3 = 0 \\
 H_1 &: \text{al menos un } \beta_i \neq 0, i = 1, 2, 3
 \end{aligned}$$

Si la hipótesis nula es cierta, el valor obtenido debe ser la realización de una F de Fisher-Snedecor con 3 y 75 grados de libertad.

El **p-value**: 5.22643×10^{-11} que es menor de 0.05 nos permite rechazar la hipótesis nula y concluir que al menos uno de los coeficientes de regresión es distinto de 0.

2. Verificación de los supuestos del modelo de regresión lineal múltiple

Como ya vimos en el tema del modelo de regresión lineal simple, para que toda la inferencia realizada en el modelo de regresión lineal múltiple sea correcta, han de ser correctos los supuestos asumidos.

1. Relación lineal entre $E[Y|X_1 = x_{1i}, X_2 = x_{2i}, \dots, X_p = x_{pi}]$ y cada una de las variables explicativas X_j $j = 1, 2, \dots, p$
2. Homogeneidad de varianzas
3. Independencia de las observaciones
4. Normalidad de las observaciones

En este apartado, veremos algunas de las herramientas gráficas que podemos utilizar para detectar fallos en los supuestos en los que hemos basado la estimación del modelo de regresión lineal múltiple.

Aunque en el tema de regresión lineal simple ya desarrollamos algunas de estas herramientas, veremos como aplicarlas en el caso de la regresión lineal múltiple y estudiaremos algunas particularidades que surgen cuando tenemos más de una variables explicativa.

2.1. Residuos

Al igual que en el MRLS vamos a utilizar los residuos para detectar anomalías en los supuestos asumidos en la estimación y análisis del modelo de regresión lineal múltiple.

Recordamos que el vector de residuos del modelo estimado es:

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Cada elemento $\hat{e}_i = y_i - \hat{y}_i$ del vector de residuos $\hat{\mathbf{e}}$ es la diferencia entre la observación y la media estimada con el modelo de regresión.

Los residuos, $\hat{e}_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$, son las estimas de los errores e_i .

Si el modelo es adecuado para el conjunto de datos analizado, entonces, los residuos deben reflejar las propiedades de los e_i . Esta es la idea básica por la que se utilizan los residuos en el diagnóstico del MRLM.

2.2. Propiedades de los residuos.

- Los residuos tiene media 0 y su varianza no es constante

La varianza de los residuos es:

$$\text{var}(\hat{e}_i) = \sigma^2(1 - h_{ii}) \quad i = 1, 2, \dots, n$$

siendo h_{ii} es el elemento i -esimo de la diagonal de la matriz \mathbf{H} .

- Los elementos fuera de la diagonal de la matriz \mathbf{H} son distintos de cero por lo que los residuos no son independientes.

$$\text{cov}(\hat{e}_i, \hat{e}_k) = -\sigma^2 h_{ik}$$

Sin embargo, cuando el tamaño de la muestra es relativamente grande en comparación con el número de parámetros estimados, la dependencia entre los residuos es relativamente despreciable.

Por este motivo se suelen utilizar los residuos estandarizados:

$$r_i = \frac{\hat{e}_i}{\sqrt{CME(1 - h_{ii})}} \quad i = 1, 2, \dots, n$$

que aunque no son independientes tienen media 0 y varianza 1.

Al igual que hacíamos el modelo de regresión lineal simple, podemos utilizar diferentes tipos de gráficos en lo que intervienen los residuos o los residuos estandarizados para detectar alguna anomalía en los supuestos del modelo de regresión lineal múltiple.

En concreto utilizaremos: los gráficos de los residuos frente a los valores predichos (\hat{e}_i vs \hat{y}_i), los gráficos de dispersión-centralidad y los gráficos cuantil-cuantil para la normalidad.

La forma de construir, interpretar los gráficos y los posibles remedios que podemos utilizar en caso de detectar alguna anomalía ya han sido estudiados en el tema de regresión lineal simple.

Ejemplo. Emisiones de metano en purines

Para obtener estos gráficos en R tecleamos `plot(ModReg1)`. Automáticamente se muestran los 4 gráficos mostrados en la figura 3.

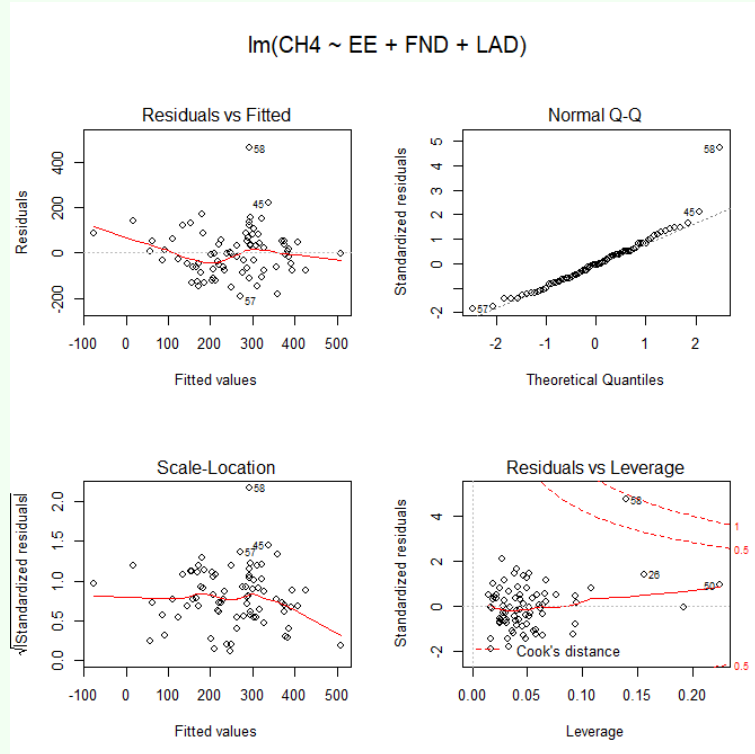


Figura 3: Gráfico de los residuos para verificar los supuesto del modelo de regresión estimado

ModReg1

En la esquina superior izquierda se representa es gráfico de los residuos frente a los valores predichos, en la esquina superior derecha el gráfico de normalidad y en la esquina inferior izquierda el gráfico dispersión-centralidad. El gráfico representado en la esquina inferior derecha lo veremos más adelante.

En el caso de la regresión lineal múltiple, el gráfico de de los residuos frente a los valores predichos nos permite detectar si hay heterogeneidad de varianzas pero en general, no suelen ser adecuados para detectar si el supuesto de linealidad asumido es correcto.

En los gráficos de los residuos frente a los valores predichos y de dispersión-centralidad no observamos ninguna anomalía grave sobre el supuesto de homogeneidad de varianzas.

Transformación de la variable respuesta

Aunque en el gráfico de normalidad los residuos se ajustan a la recta de referencia, los residuos positivos tienden a separarse mostrando valores más alto que lo esperado en una normal, en especial los dos residuos marcados como posibles valorea atípicos (45 y 58)

Vamos a utilizar la transformación de Box-Cox para determinar si es necesario transformar la variable respuesta.

Para ello, vamos R y llamamos a `powerTransform`, en concreto

```
> summary(powerTransform(ModReg1, family = "bcPower"))
```

y el resultado se muestra a continuación.

```
bcPower Transformation to Normality
  Est Power Rounded Pwr Wald Lwr Bnd Wald Up Bnd
Y1    0.4927          0.5    0.2563    0.7291

Likelihood ratio test that transformation parameter is equal to 0
(log transformation)
              LRT df      pval
LR test, lambda = (0) 20.27428  1 0.0000067096

Likelihood ratio test that no transformation is needed
              LRT df      pval
LR test, lambda = (1) 14.79478  1 0.00011987
```

El valor de λ estimado es de 0.4927 que es redondeado a 0.5, lo que indica que para conseguir que la variable CH4 se asemeje a una distribución normal debemos utilizar la transformación raíz cuadrada.

También nos muestra el intervalo de confianza del parámetro λ . Con un nivel de confianza del 95 % el parámetro λ puede ser cualquier valor comprendido entre 0.26 y 0.73.

Vemos que cuando realizamos el test de hipótesis de que no es necesario transformar la variable respuesta ($\lambda = 1$), rechazamos la hipótesis nula ya que el valor de LRT=14.79478 y el

$p_{val} = 0.00011987$.

Una vez que hemos determinado cual es la transformación adecuada, debemos estimar el modelo utilizando como variable respuesta la raíz cuadrada de la variable CH4.

Los pasos a seguir son :

1. Crear la variable `tCH4` utilizando la raíz cuadrada de la variable `CH4`.
2. Estimar el modelo de regresión que relaciona las variables `tCH4` con las tres variables explicativas `EE`, `FND` y `LAD`.
3. Verificar los supuestos del modelo de regresión utilizando los gráficos de los residuos.

Para construir la variable con la raíz cuadrada seguimos los mismos pasos que en el tema de RLS. Tecleamos `Purines$tCH4<-sqrt(Purines$CH4)`

En el conjunto de datos `Purines` hay una nueva columna `tCH4` con la raíz cuadrada de los valores en la columna `CH4` y estimamos el modelo con la variable respuesta transformada

```
> Purines$tCH4 <- sqrt(Purines$CH4)
> ModReg2t <- lm(tCH4 ~ EE + FND + LAD, data = Purines)
```

```
Call:
lm(formula = tCH4 ~ EE + FND + LAD, data = Purines)

Residuals:
    Min       1Q   Median       3Q      Max
-7.5767 -1.8473 -0.1736  2.1178 11.7506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.85599    1.83758   6.996 9.44e-10 ***
EE           0.43473    0.11370   3.824 0.00027 ***
FND          0.14205    0.05788   2.454 0.01643 *
LAD         -0.98774    0.23530  -4.198 7.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.394 on 75 degrees of freedom
Multiple R-squared:  0.5401, Adjusted R-squared:  0.5217
F-statistic: 29.36 on 3 and 75 DF, p-value: 1.157e-12
```

Construimos los gráficos de los residuos con `plot(ModReg2t)`. Automáticamente se muestran los 4 gráficos mostrados en la figura 4.

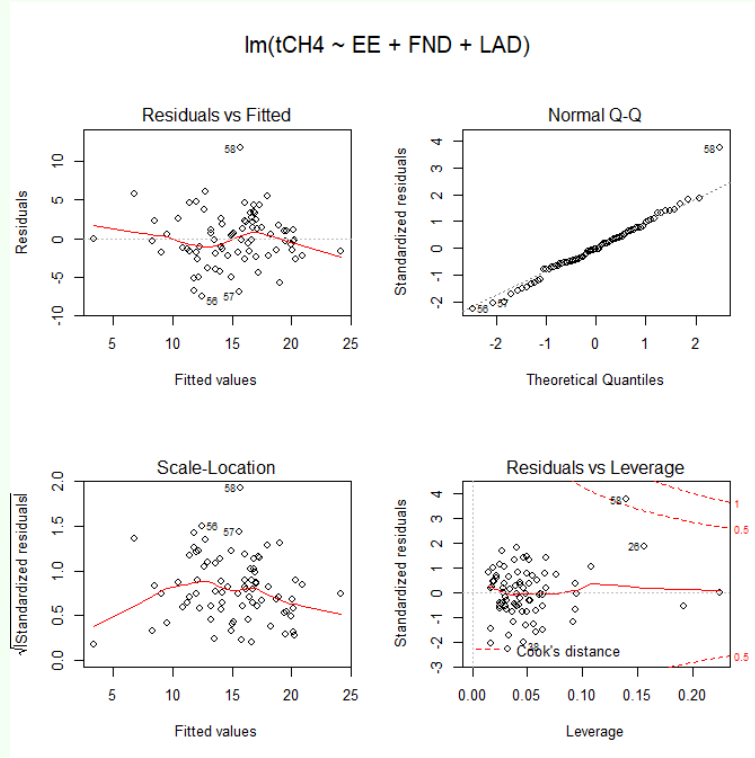


Figura 4: Gráfico de los residuos para verificar los supuestos del modelo de regresión estimado ModRegt2

Observamos que los residuos extremos en el gráfico Q-Q Norm se ajustan mejor a la línea de referencia.

2.3. Gráficos de componentes más residuos

Los gráficos de componentes más residuos son adecuados para verificar el supuesto de linealidad entre la variable respuesta y cada una de las variables explicativas.

Para cada variable explicativa se realiza uno de estos gráficos.

Es decir, para la variable X_j y para cada observación, se calcula el siguiente residuo:

$$\hat{e}_i^{(j)} = \hat{e}_i + b_j \times x_{ji} \quad i = 1, 2, \dots, n$$

El superíndice (j) indica que es el residuo aumentado para la variable explicativa $j = 1, 2, \dots, p$. Cada variable explicativa se asocia con n distintos residuos aumentados.

Con estos residuos $e_i^{(j)}$ y los valores de la variable X_{ji} , se realiza un gráfico de dispersión.

Tal y como están contruidos, la pendiente de la recta de regresión estimada para esta nube de puntos coincide con el coeficiente de regresión b_j del modelo completo y los residuos que se distribuyen alrededor de la recta de regresión coinciden con los residuos del modelo completo.

En estos gráficos, cualquier relación no lineal de la variable explicativa representada se mostrará de forma clara. Además, veremos si la relación no lineal es monótona, en cuyo caso, una transformación de la variable podría conseguir la relación lineal.

Ejemplo. Emisiones de metano en purines

Vamos a construir en R los gráficos de componentes más residuos en el modelo estimado `ModReg2t`.

La función a usar será

```
> crPlots(ModReg2t, smooth = list(span = 0.5))
```

Obtenemos los tres gráficos mostrados en la figura 5

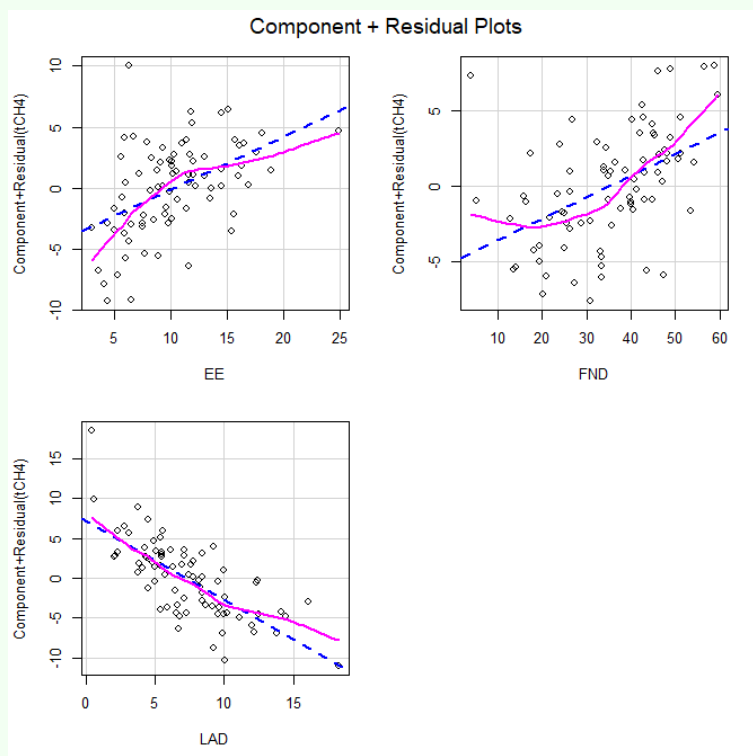


Figura 5: Gráfico de los residuos más componentes para verificar el supuesto de relación lineal de las variables explicativas en el modelo estimado `ModReg2t`

Como siempre, las líneas rojas dibujadas son las tendencias estimadas con una regresión no paramétrica.

El gráfico residuos más componentes de la variable explicativa `EE` es muy similar al gráfico que habíamos obtenido en el modelo de regresión lineal simple que relacionaba la variable `tCH4` y la variable `EE`. Observamos una tendencia no lineal más o menos monótona. Hay un incremento desde valores pequeño de la variable `EE` hasta valores intermedio y a partir de estos valores es incremento se reduce o incluso se hace cero.

En el caso de la variable `FND` la no linealidad es más marcada pero no es monótona. Hay una pendiente negativa para valores bajos de `FND`, con un mínimo en valores intermedios y luego cambia a una pendiente positiva. En este caso una transformación de la variable `FND` no soluciona el problema.

El gráfico de la variable `LAD` muestra que la relación lineal parece adecuada.

Transformación de la variable EE

Igual que hicimos cuando estimamos el modelo de regresión lineal que relacionaba a las variables CH4 y EE vamos a utilizar la transformación logarítmica para la variable EE. Veámos que esta transformación hacía más simétrica la distribución de esta variable y mejoraba la relación lineal entre la variable tCH4 y log(EE).

Los pasos a seguir son:

1. Crear la variable transformada tEE utilizando la función logaritmo.
2. Estimar el modelo de regresión lineal múltiple que relaciona la variable respuesta tCH4 con las variables explicativas tEE, FND y LAD.
3. Verificar los supuestos del modelo estimado mediante los gráficos de los residuos y los gráficos de residuos más componentes.

El modelo estimado lo llamamos ModReg3t y en la figura 6 se muestra los gráficos de residuos más componentes para las tres variables explicativas utilizadas en el modelo de regresión lineal.

```
> Purines$tEE <- log(Purines$EE)
> ModReg3t <- lm(tCH4 ~ tEE + FND + LAD, data = Purines)
```

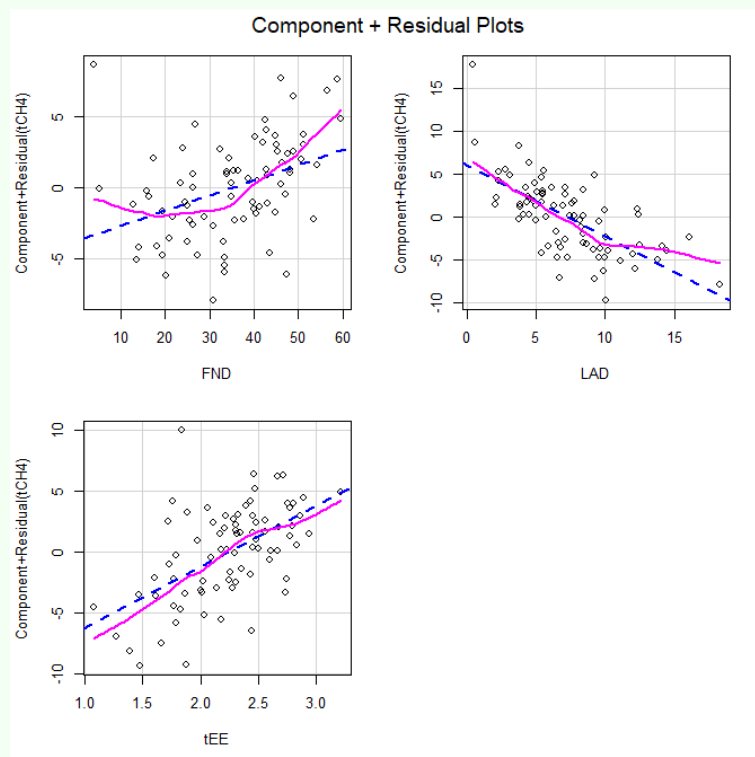


Figura 6: Gráfico de los residuos más componentes para verificar el supuesto de relación lineal de las variables explicativas en el modelo estimado ModReg3t

Podemos comprobar como ha cambiado el gráfico para la variable tEE

Regresión polinómica de orden 2 para la variable FND

En el caso de la variable FND donde la relación no lineal no es monótona utilizaremos una **regresión polinómica de orden dos** en la variable FND. En este caso, en el modelo de regresión incorporamos la variable explicativa FND y su valor elevado al cuadrado FND^2 , es decir el modelo a estimar será:

$$CH_4^{1/2} = \beta_0 + \beta_1 \log(EE) + \beta_2 FND + \beta_3 FND^2 + \beta_4 LAD$$

Para estimar este modelo, al que vamos a llamar **ModReg3t** los pasos a seguir son los siguientes:

1. Crear la variable FND2 con los valores de la variable FND elevados al cuadrado.
2. Estimar el modelo de regresión múltiple que relaciona la variable **tCH4** y las variables **tEE**, **FND**, **FND2** y **LAD**.
3. Realizar los gráficos de residuos y componentes para verificar los supuestos del modelo.

Para crear la variable FND2 y estimar el modelo con las variables transformadas, tecleamos

```
> Purines$FND2 <- Purines$FND^2
> ModReg4t <- lm(tCH4 ~ tEE + FND + FND2 + LAD, data = Purines)
```

Call:

```
lm(formula = tCH4 ~ FND + FND2 + LAD + tEE, data = Purines)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.447	-1.584	-0.076	1.770	7.209

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.142159	2.954153	3.772	0.000324 ***
FND	-0.400674	0.134587	-2.977	0.003931 **
FND2	0.007814	0.001903	4.105	0.000103 ***
LAD	-0.836447	0.220216	-3.798	0.000296 ***
tEE	5.914999	1.053733	5.613	0.000000328 ***

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

Residual standard error: 3.005 on 74 degrees of freedom

Multiple R-squared: 0.6444, Adjusted R-squared: 0.6252

F-statistic: 33.52 on 4 and 74 DF, p-value: 6.052e-16

En la figura 7 se muestra los gráficos de residuos más componentes para las cuatro variables explicativas utilizadas en el modelo de regresión lineal.

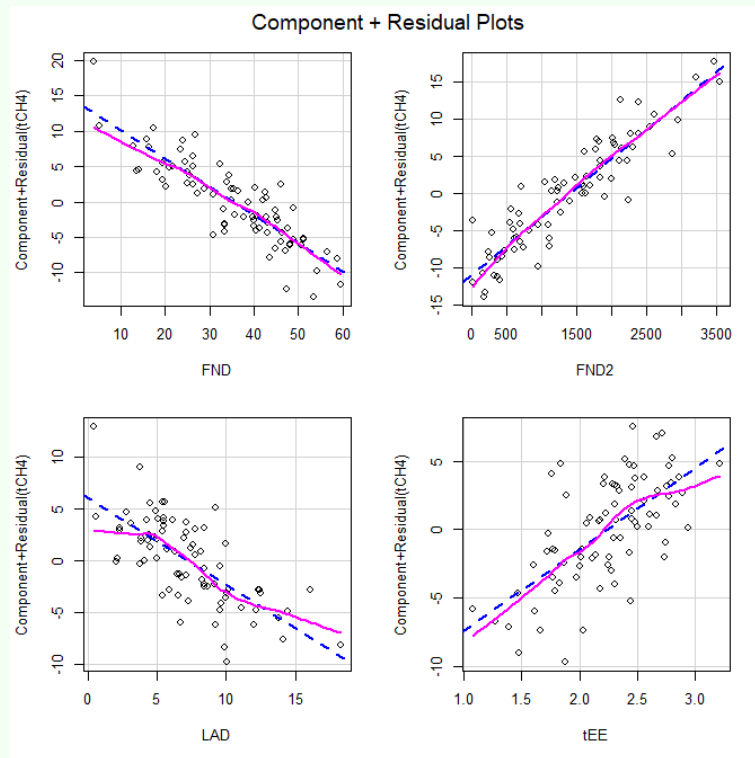


Figura 7: Gráfico de los residuos más componentes para verificar el supuesto de relación lineal de las variables explicativas en el modelo estimado **ModReg4t**

Observamos como, para las variables **FND** y **FND2** las tendencias estimadas coinciden con las rectas de regresión con pendientes -0.40 y 0.0078 que son, respectivamente, las estimas de los coeficientes de regresión de las dos variables.

2.4. Modelo estimado e interpretación. Visualización

El modelo estimado es lineal en las variables transformadas y por lo tanto los coeficientes de regresión estimado se interpretan en la escala transformada. En esta escala, el efecto del coeficiente de regresión parcial, β_j , se mantiene constante cualquier que sea el valor del resto de variables explicativas, siempre y cuando se mantengan constante.

Sin embargo, puede resultar interesante visualizar como es la relación parcial de la variable respuesta y la variable explicativa en las escalas originales de las dos variables. En la escala original de las variables, la relación lineal desaparece y por tanto, la relación parcial entre la variable explicativa X_j y la variable respuesta Y cambia cuando el resto de variables explicativas cambian de valor.

Vamos a utilizar el modelo finalmente estimado en el ejemplo de emisiones de purines, para ver como podemos visualizar el efecto parcial de las variables relacionada con la composición de los purines en la cantidad de metano emitido.

Ejemplo. Emisiones de metano en purines

Modelo estimado e interpretación

El modelo final estimado tiene la siguiente forma:

$$\widehat{CH4}^{1/2} = 11.14 + 5.915 \log(E E)$$

$(2.954) \quad (1.054)$

$$-0.4007 FND + 0.00781 FND^2$$

$(0.1346) \quad (0.0019)$

$$-0.8364 LAD$$

(0.2202)

$$R_{adj}^2 = 0.62515$$

Debajo de cada coeficiente de regresión estimado se muestra entre paréntesis el error estándar de dicha estima.

Esta es la forma habitual de mostrar un modelo de regresión lineal múltiple estimado.

Vamos a analizar como se relacionan parcialmente cada una de las variables explicativas con las emisiones de metano.

Variable explicativa EE

En el caso de la variable EE, el coeficiente de regresión estimado, 5.915, nos indica cuanto cambia el valor medio de la variable respuesta, $\widehat{CH4}$ por incrementar una unidad el valor de la variable $\log EE$, manteniendo constantes el resto de variable del modelo.

Vamos a fijar el valor de las variables FND y LAD. Se puede utilizar cualquier valor dentro del rango de valores de las variables en la muestra, pero habitualmente se les asigna su valor medio.

Para estos valores asignados, la relación entre la variable $\widehat{CH4}$ y $\log EE$ quedaría de la siguiente forma:

$$\widehat{tCH4} = 11.14 + 5.915tEE - 0.4007\overline{FND} + 0.00781\overline{FND}^2 - 0.8364\overline{LAD}$$

donde $\overline{FND} = 35.2$ y $\overline{LAD} = 7.25$ son las media muestrales de las variables FND y LAD, respectivamente.

El modelo queda reducido a un modelo de regresión lineal simple entre las dos variables.

$$\widehat{tCH4} = k_0 + 5.915tEE \quad (6)$$

siendo

$$k_0 = 11.14 - 0.4007\overline{FND} + 0.00781\overline{FND}^2 - 0.8364\overline{LAD} = 0.653$$

un valor constante.

Con este modelo operamos de la misma forma que hacíamos en el tema del MRLS. Damos distintos valores a la variable tEE, para cada valor obtenemos las medias estimada de $\widehat{tCH4}$ e invertimos las transformaciones para las dos variables.

A modo de ejemplo, en la tabla 2, se muestran 5 valores de la variable tEE y los valores estimados, con el modelo 6, de la variable $\widehat{tCH4}$.

Para estos valores invertimos la transformación realizada en cada variable. Para ello, el número e se eleva a los valores de tEE y obtenemos los valores de EE. Los valores de $\widehat{tCH4}$ se elevan al cuadrado para obtener $\widehat{CH4}$. A partir de estos valores haríamos la representación gráfica de la relación entre las variables $\widehat{CH4}$ y EE.

Si cambiamos los valores de las variables FND y LAD, cambiará el valor de k_0 , desplazando la recta de regresión verticalmente. El efecto de la variable tEE sobre la variable tCH4 no cambia, pero si lo hará cuando las representamos en la escala original.

Variable explicativa FND

Para el caso de la variable explicativa FND operamos de la misma forma, pero teniendo en cuenta que en el modelo tenemos es un polinomio de orden 2 de esta variable.

Si fijamos los valores de las variables tEE y LAD en su valor medio, el modelo de regresión estimado quedaría:

$$\widehat{tCH4} = 11.14 + 5.915\overline{tEE} - 0.4007\overline{FND} + 0.00781\overline{FND}^2 - 0.8364\overline{LAD}$$

donde $\overline{tEE} = 2.24$ y $\overline{LAD} = 7.25$ son las media muestrales de las variables tEE y LAD, respectivamente.

El modelo queda reducido a un modelo de regresión con un polinomio de orden 2 para la variable FND.

$$\widehat{tCH4} = k_0 - 0.4007FND + 0.00781FND^2 \quad (7)$$

siendo

$$k_0 = 11.14 + 5.915\overline{tEE} - 0.8364\overline{LAD} = 18.3$$

Para distintos valores de la variable FND obtenemos los valores estimados de la variable $\widehat{tCH4}$ pudiendo representar el polinomio de orden 2 que relaciona a las dos variables, o bien, elevando al cuadrado los valores de $\widehat{tCH4}$, la función que relaciona $\widehat{CH4}$ con FND.

Variable explicativa LAD

Para el caso de la variable explicativa LAD operamos de la misma forma.

Si fijamos los valores de las variables tEE y FND en su valor medio, el modelo de regresión estimado quedaría:

$$\widehat{tCH4} = 11.14 + 5.915\overline{tEE} - 0.4007\overline{FND} + 0.00781\overline{FND}^2 - 0.8364LAD$$

donde $\overline{tEE} = 2.24$ y $\overline{FND} = 35.2$ son las media muestrales de las variables tEE y FND, respectivamente.

El modelo queda reducido a un modelo de regresión lineal simple entre las dos variables.

$$\widehat{tCH4} = k_0 - 0.8364LAD \quad (8)$$

siendo

$$k_0 = 11.14 + 5.915\overline{tEE} - 0.4007\overline{FND} + 0.00781\overline{FND}^2 = 20$$

Para distintos valores de la variable LAD obtenemos los valores estimados de la variable $\widehat{tCH4}$ pudiendo representar la recta que relaciona a las dos variables, o bien, elevando al cuadrado los valores de $\widehat{tCH4}$, representamos la función que relaciona $\widehat{CH4}$ con LAD.

EE	tEE	tCH4	CH4
2.90	1.06	26.30	692.00
7.90	2.07	32.20	1040.00
12.90	2.56	35.10	1230.00
17.90	2.88	37.10	1370.00
22.90	3.13	38.50	1480.00

Cuadro 2: Valores medios estimados de la variable tCH4 para distinto valores de la variable tEE y sus valores en la escala original en el modelo de regresión estimado ModReg4t cuando las variables FND y LAD están fijadas en su media

Ejemplo.

Para hacer esto en R debemos volver a estimar el modelo

teclemos.

```
> ModReg5t <- lm(sqrt(CH4) ~ log(EE) + poly(FND, degree = 2, raw = TRUE) +  
+ LAD, data = Purines)
```

Call:

```
lm(formula = sqrt(CH4) ~ log(EE) + poly(FND, degree = 2, raw = TRUE) +  
LAD, data = Purines)
```

Residuals:

```
      Min       1Q   Median       3Q      Max  
-7.447 -1.584 -0.076  1.770  7.209
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	11.142159	2.954153	3.772	0.000324	***
log(EE)	5.914999	1.053733	5.613	0.000000328	***
poly(FND, degree = 2, raw = TRUE)1	-0.400674	0.134587	-2.977	0.003931	**
poly(FND, degree = 2, raw = TRUE)2	0.007814	0.001903	4.105	0.000103	***
LAD	-0.836447	0.220216	-3.798	0.000296	***

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

Residual standard error: 3.005 on 74 degrees of freedom

Multiple R-squared: 0.6444, Adjusted R-squared: 0.6252

F-statistic: 33.52 on 4 and 74 DF, p-value: 6.052e-16

La salida es exactamente la misma que la obtenida con el modelo estimado ModReg4t. La única diferencia es que no ha sido necesario crear previamente las variables transformadas tCH4, tEE y FND2 ya que se han indicado las transformaciones en la especificación del modelo.

Para representar visualmente el efecto de cada una de la variables explicativas utilizadas en el modelo estimado, teclemos

```
> plot(predictorEffects(ModReg5t))
```

y obtenemos los gráficos mostrados en la figura 8.

La línea azul son los valores medios estimados con el modelo de regresión y las bandas azules los intervalos de confianza. Las barras verticales que se muestran en los ejes horizontales de los tres gráficos marcan los valores de las variables explicativas utilizados para estimar el modelo de regresión lineal múltiple.

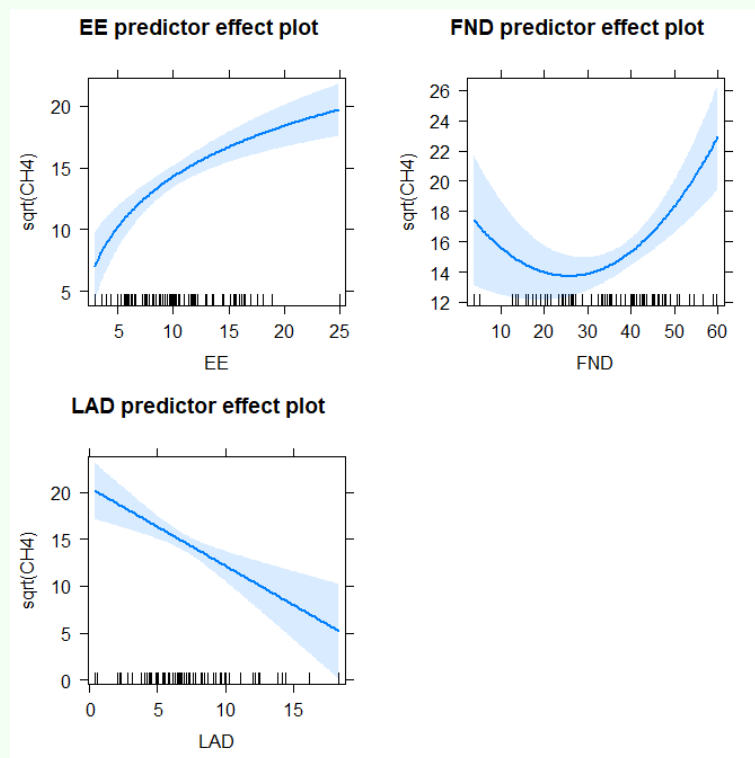


Figura 8: Gráficos de las relaciones parciales de las variables explicativas EE, FND y LAD y la variable respuesta $\sqrt{\text{CH4}}$ en el modelo estimado ModReg5t

Sin embargo, en el gráfico 8 la variable respuesta sigue en la escala transformada, en este ejemplo la raíz cuadrada.

Si queremos hacer los mismos gráficos pero con la variable respuesta en la escala original, tenemos que teclear en R la misma instrucción que ha utilizado R para hacer el gráfico en 8 añadiendo lo siguiente:

```
plot(predictorEffects(ModReg5t,
  transformation=list(link=sqrt,inverse=function(x){x**2})),
  axes=list(y=list(lab="CH4",type="response")))
```

Se ha añadido en la instrucción el argumento `transformation=list(inverse=function(x)x**2)` para indicarle que la inversa de la transformación son los valores de $\sqrt{\text{CH4}}$ elevados al cuadrado y el argumento `axes=list(y=list(lab="CH4",type="response"))` para indicarle que el título del eje vertical es CH4 y que represente la variable en la escala original.

Al ejecutar esta instrucción, obtenemos los gráficos mostrados en la figura 9

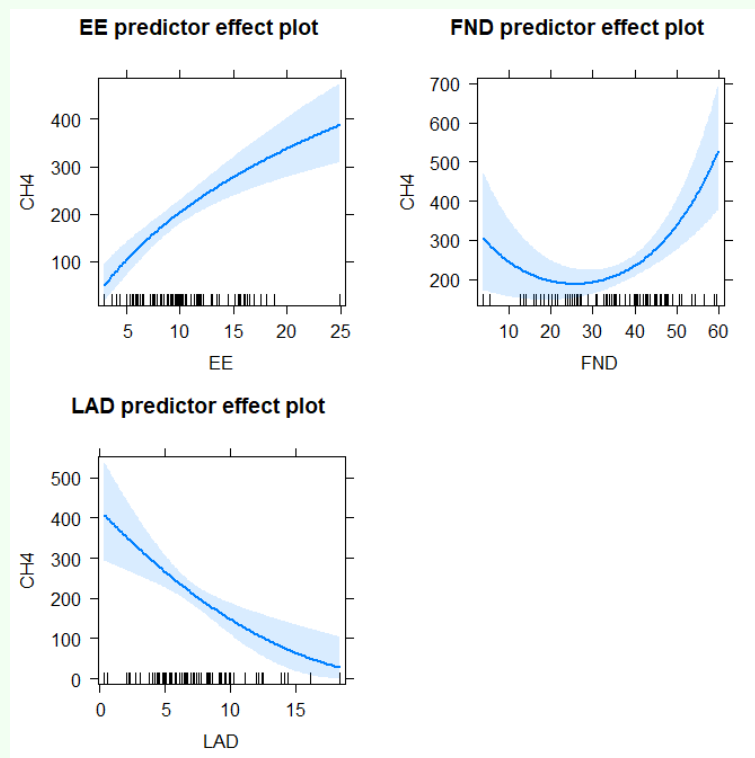


Figura 9: Gráficos de las relaciones parciales de las variables explicativas EE, FND y LAD y la variable respuesta CH₄ en el modelo estimado ModReg5t

3. Validación cruzada

El uso de métodos de regresión lineal puede tener dos objetivos. Por un lado nos puede interesar explicar un fenómeno que observamos mediante nuestra variable dependiente y , con lo cual revisando que se cumplen las premisas del modelo lineal, revisando que posibles valores atípicos y pudiendo explicar la plausibilidad biológica de las variables elegidas, habríamos terminado.

Pero por otro lado también podríamos estar interesados en predecir nuestra variable dependiente, y , a través del modelo lineal que estimemos. En este caso la selección del mejor modelo y del error cometido debe hacerse mediante validación cruzada. La validación cruzada tiene los siguientes pasos,

1. Dividir el total de observaciones, N en k subconjuntos de manera aleatoria
2. Elegir uno de los k subconjuntos y con él, ajustar el modelo.
3. Con el resto de subconjuntos estimar el error cometido comparando y_i vs \hat{y}_i
4. Elegir otro subconjunto k y repetir los pasos 2) y 3) hasta que todos los subconjuntos k hallan sido utilizados para ajustar el modelo
5. Repetir todo el proceso n iteraciones
6. Calcular la media del error cometido en cada una de las iteraciones

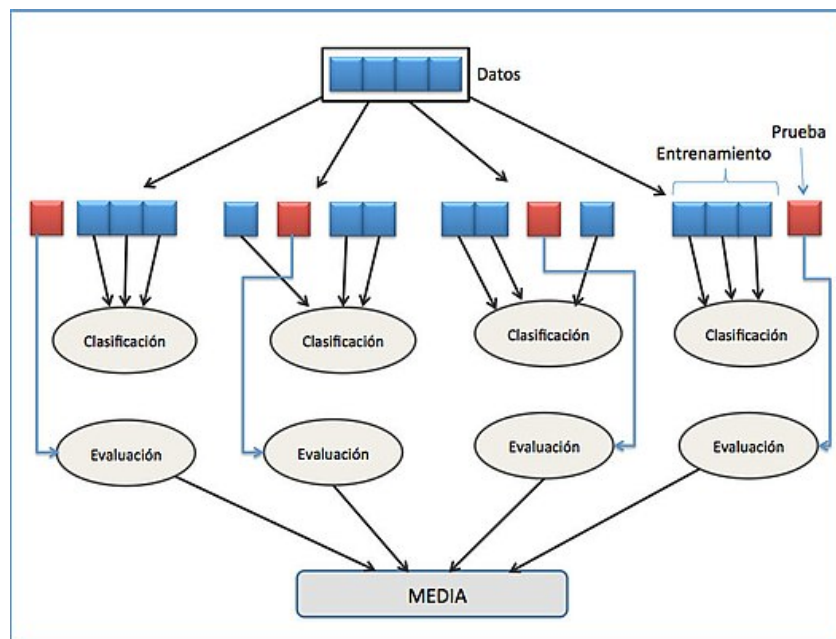


Figura 10: Esquema de validación cruzada con 4 subconjuntos, sets. Fuente: Wikipedia

El primer paso es fundamental. Los k subconjuntos deben ser una *m.a.s.* del conjunto inicial de datos.

Una forma de comparar y_i vs \hat{y}_i es con RMSEP,

$$RMSEP = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

Ejemplo. Metano en purines

Con los datos de los purines vamos a ver la habilidad predictiva del polinomio de orden 2 y el de orden 3, programando un esquema de validación cruzada en 3 sets.

```
> Purines$tCH4 <- sqrt(Purines$CH4)
> Purines$tEE <- log(Purines$EE)
> ind <- sample(c(1, 2, 3), replace = TRUE, size = 79)
> niter <- 100
> Solmod1 <- matrix(0, ncol = 3, nrow = niter)
> Solmod2 <- matrix(0, ncol = 3, nrow = niter)
> for (i in 1:niter) {
+   ind <- sample(c(1, 2, 3), replace = TRUE, size = 79)
+   for (j in 1:3) {
+     datf <- Purines[ind == j, ]
+     datp <- Purines[ind != j, ]
+     fit1 <- lm(tCH4 ~ LAD + poly(FND, 2) + tEE,
+               data = Purines)
+     fit2 <- lm(tCH4 ~ LAD + poly(FND, 3) + tEE,
+               data = Purines)
+     pre1 <- predict(fit1, datp)
+     pre2 <- predict(fit2, datp)
+     Solmod1[i, j] <- sqrt(sum((datp$tCH4 - pre1)^2)/length(pre1))
+     Solmod2[i, j] <- sqrt(sum((datp$tCH4 - pre2)^2)/length(pre2))
+   }
+ }
> modFND2 <- round(mean(apply(Solmod1, 1, mean)), 3)
> modFND3 <- round(mean(apply(Solmod2, 1, mean)), 3)
> modFND2

[1] 2.902

> modFND3

[1] 2.863
```

4. Selección de variables

En esta sección vamos a afrontar uno de los problemas prácticos más comunes cuando se usa la regresión lineal, ¿qué variables incluyo en el modelo?

A lo largo del tema utilizaremos unos datos forenses.

Ejemplo. Cristales

Para estudiar los modelos de regresión, usaremos los datos del paquete **MASS** *Los datos fgl tienen 214 filas y 10 columnas. Fueron recogidos por B. German con fragmentos de cristal en trabajos forenses*

RI es el índice de refracción, la variable dependiente. Las demás variables indican el porcentaje de ese compuesto químico en la muestra. Sabiendo el RI, podemos clasificar el cristal.

```
> library(car)
> library(MASS)
> data(fgl)
> fgl2 <- fgl[, -10]
> datos <- fgl2
```

En primer lugar realizamos un resumen numérico de las variables

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Min.	-6.85	10.73	0.00	0.29	69.81	0.00	5.43	0.00	0.00
1st Qu.	-1.48	12.91	2.11	1.19	72.28	0.12	8.24	0.00	0.00
Median	-0.32	13.30	3.48	1.36	72.79	0.56	8.60	0.00	0.00
Mean	0.37	13.41	2.68	1.44	72.65	0.50	8.96	0.18	0.06
3rd Qu.	1.16	13.83	3.60	1.63	73.09	0.61	9.17	0.00	0.10
Max.	15.93	17.38	4.49	3.50	75.41	6.21	16.19	3.15	0.51
SD	3.04	0.82	1.44	0.50	0.77	0.65	1.42	0.50	0.10
n	214.00	214.00	214.00	214.00	214.00	214.00	214.00	214.00	214.00

```
> scatterplotMatrix(fgl2, diagonal = list(method = "boxplot"),
+   smooth = FALSE)
```

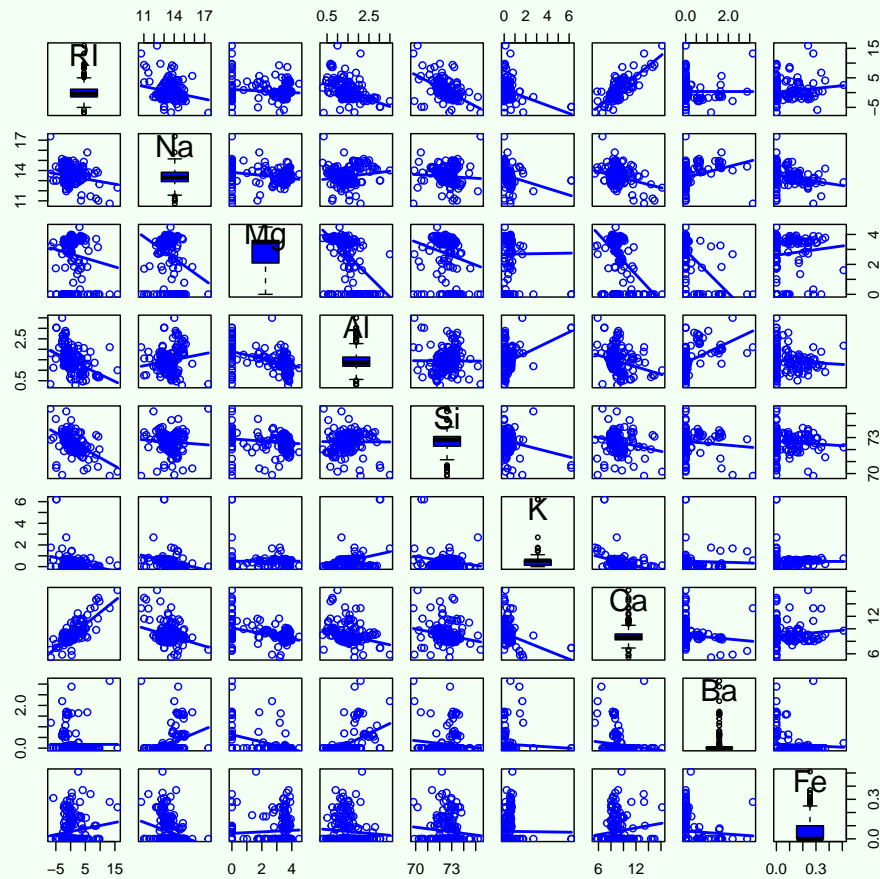
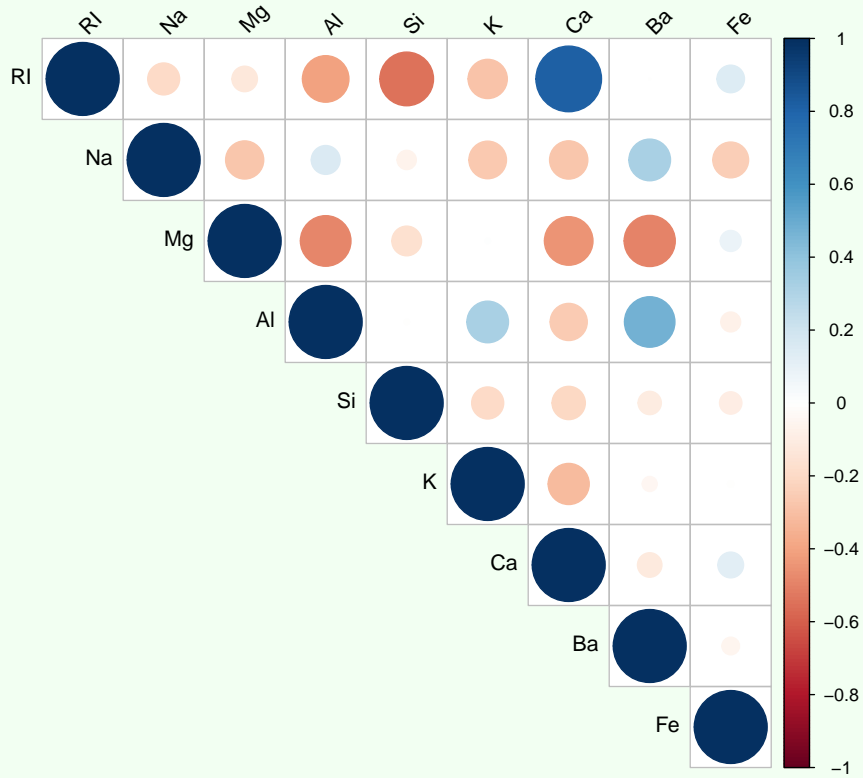


Figura 11: Matriz de gráficos de dispersión de las variables RI, Na, Mg, Al, Si, K, Ca, Ba y Fe de las 214 muestras analizadas

- *Na*, porcentaje de Sodio
- *Mg*, porcentaje de Magnesio
- *Al*, porcentaje de Aluminio
- *Si*, porcentaje de Silicio
- *K*, porcentaje de Potasio
- *Ca*, porcentaje de Calcio
- *Ba*, porcentaje de Bario
- *Fe*, porcentaje de Hierro

Un aspecto a tener en cuenta es que no exista correlación entre variables explicativas, para evitar la colinealidad.

```
> library(corrplot)
> k <- cor(fgl2)
> corrplot(k, type = "upper", tl.col = "black", tl.srt = 45)
```



En el gráfico de correlaciones entre variables. Hay que prestar especial atención a la correlación entre variables explicativas

Basta notar que la suma de estas variables, es 100 o casi 100, eso nos indica que estamos antes un tipo de datos especial, *compositional* data, que no abordaremos en profundidad.

4.1. Ajustar un modelo por cada variable

Si se tienen p variables explicativas, consiste en

- ajustar un modelo para cada una de esas p variables
- recoger un p valor ó bien de la *significación* del coeficiente de regresión o de la prueba F de ese modelo.
- recoger los límites del intervalos de ocnfianza de los coeficientes de regresión
- ajustar por comparaciones los p valores anteriores
- seleccionar las variables cuyos p valores corregidos sean menores de 0.05
- ajustar un modelo con las variables seleccionadas

Ejemplo.

```
> vary <- colnames(datos)[-1]
> sol <- matrix(0, nrow = length(vary), ncol = 4)
> colnames(sol) <- c("lower", "upper", "p.beta", "p.beta.adjust")
> rownames(sol) <- vary
> # Bucle que ajusta modelos variable por variable y recoge
> # IC y p valor
> for (i in 1:length(vary)) {
+   m <- lm(formula(paste("RI~", vary[i], sep = "")), data = datos)
+   sol[i, 1:2] <- confint(m)[2, ]
+   sol[i, 3] <- summary(m)$coefficients[2, 4]
+ }
> sol[, 4] <- p.adjust(sol[, 3])
> round(sol, 4)

      lower    upper p.beta p.beta.adjust
Na -1.2077 -0.2195 0.0049      0.0194
Mg -0.5403  0.0255 0.0743      0.1485
Al -3.2297 -1.7255 0.0000      0.0000
Si -2.5714 -1.6792 0.0000      0.0000
K  -1.9529 -0.7462 0.0000      0.0001
Ca  1.5601  1.8986 0.0000      0.0000
Ba -0.8292  0.8245 0.9955      0.9955
Fe  0.2810  8.6333 0.0366      0.1097

> # Ahora selecciono las variables que tienen un p valor
> # corregido menor de 0.05
> rownames(sol)[sol[, 4] < 0.05]

[1] "Na" "Al" "Si" "K"  "Ca"

> var.f <- rownames(sol)[sol[, 4] < 0.05]
> mf <- lm(formula(paste("RI~", paste(var.f, collapse = "+", sep = ""))),
+   data = datos)
```

Las estimas de ese modelo son:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	121.261	8.908	13.612	0.000
Na	-0.227	0.114	-1.986	0.048
Al	-1.210	0.171	-7.074	0.000
Si	-1.753	0.111	-15.797	0.000
K	-0.629	0.155	-4.074	0.000
Ca	1.292	0.067	19.204	0.000

Ejemplo. La importancia de corregir por comparaciones múltiples

Vamos a ver de una manera práctica la importancia de corregir por comparaciones múltiples. Simularemos 100 variables explicativas y lanzaremos tantos modelos de regresión simple como variables.

```
> set.seed(54321)
> nvar <- 100
> nobs <- 100
> dat <- matrix(rnorm(101 * 100), ncol = nvar + 1, nrow = nobs)
> colnames(dat) <- c("y", paste("X", 1:100, sep = ""))
> dat <- data.frame(dat)
> # Usaremos el código anterior para lanzar 100 modelos
> vary <- colnames(dat)[-1]
> sol <- matrix(0, nrow = length(vary), ncol = 3)
> colnames(sol) <- c("lower", "upper", "p.beta")
> rownames(sol) <- vary
> # Bucle que ajusta modelos variable por variable y recoge
> # IC y p valor
> for (i in 1:length(vary)) {
+   m <- lm(formula(paste("y~", vary[i], sep = "")), data = dat)
+   sol[i, 1:2] <- confint(m)[2, ]
+   sol[i, 3] <- summary(m)$coefficients[2, 4]
+ }
> rownames(sol)[sol[, 3] < 0.05]

[1] "X5" "X18" "X19" "X30" "X35" "X72" "X81"
```

A pesar de haber generado aleatoriamente todas las variables sin ninguna relación entre ellas, aparecen 7 variables *significativas*, aproximadamente $\alpha \times 100 \sim 5$, falsos positivos.

4.2. Selección de variables. Modelo más parsimonioso

Las estimación por mínimos cuadrados presenta dos inconvenientes.

El primero de ellos es la precisión en las predicciones. Las estimas de mínimos cuadrados suelen tener poco sesgo pero la variabilidad es muy alta

El segundo es la interpretación cuando aparecen muchas variables, lo que hace que para entender el fenómeno a estudio se prefiera quedarse con las variables que tienen un mayor efecto sacrificando aquellas que aportan poco en aras de la fácil comprensión del fenómeno.

El mejor modelo, *best subset selection* busca sobre un conjunto de $\{0, 1, 2, 3 \dots, p\}$ variables, aquel subconjunto k de ellas que produce la menor suma de cuadrados residual posible. Pero esa k puede tomar los valores $1, 2, \dots, p$

Por ejemplo un modelo con sólo una variable explicativa produciría $\binom{p}{2} = p(p-1)/2$ combinaciones de las p variables. Aunque este método es conceptualmente sencillo se enfrenta a problemas de computación, ya que el número de modelo que deben ser considerados aumenta dramáticamente según p crece. Con valores de $p > 40$ este método se hace computacionalmente inabordable

Existen otros métodos llegar a ese conjunto k , los dos más utilizados son el *forward*– y *backward*–

- *forward*, comienza "hacia adelante", sólo con el término independiente y va añadiendo variables una a una. Va creando una serie de modelos anidados
- *backward*, "hacia atrás", comienza con el modelo saturado con todas las variables, p , sólo es posible cuando $N > p$, y va eliminando variables una a una. La variable candidata a salir es aquella con menor $Z - score$

La función **step** realiza ambos métodos a la vez quedándose con el mejor de las dos opciones en términos de AIC, que se ve penalizado según se aumenta el número de variables. Cuando menor es el AIC, más parsimonioso es el modelo, es decir explica más con menos variables.

Ejemplo. Cristales

Recordemos que nuestro modelo ha sido $RI \sim Na + Al + Si + K + Ca$, y lo hemos usado desde el principio sin preguntarnos nada más acerca de él. Veamos si este es el mejor modelo posible para explicar RI o no. Para ello ajustaremos el modelo con todas las variables posibles usando la instrucción $RI \sim .$ y sobre ese modelo aplicaremos la función **step**

```

> modT <- lm(RI ~ ., data = datos)
> k <- step(modT, trace = 0) # Probad sin trace=0
> # El mejor modelo esta guardado en k$terms
> modTf <- lm(formula(k$terms), data = datos)
> summary(modTf)

Call:
lm(formula = formula(k$terms), data = datos)

Residuals:
    Min       1Q   Median       3Q      Max
-4.887 -0.452 -0.028  0.420  4.366

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.8023     2.2695  -21.50  < 2e-16 ***
Na            1.2472     0.1159   10.76  < 2e-16 ***
Mg            1.7174     0.0809   21.22  < 2e-16 ***
K             1.2076     0.1360    8.88  3.1e-16 ***
Ca            2.9856     0.0810   36.85  < 2e-16 ***
Ba            2.8130     0.1803   15.60  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.998 on 208 degrees of freedom
Multiple R-squared:  0.894, Adjusted R-squared:  0.892
F-statistic: 352 on 5 and 208 DF, p-value: <2e-16

> ModReg1 <- lm(RI ~ Na + Al + Si + K + Ca, data = datos)

```

Comprobemos cual de los dos modelos tiene mejor habilidad predictiva, ya sabemos generar un esquema de validación cruzada

```

> ind <- sample(c(1, 2, 3), replace = TRUE, size = 214)
> niter <- 100
> Solmod1 <- matrix(0, ncol = 3, nrow = niter)
> Solmod2 <- matrix(0, ncol = 3, nrow = niter)
> for (i in 1:niter) {
+   ind <- sample(c(1, 2, 3), replace = TRUE, size = 79)
+   for (j in 1:3) {
+     datf <- datos[ind == j, ]
+     datp <- datos[ind != j, ]
+     fit1 <- lm(RI ~ Na + Al + Si + K + Ca, data = datos)
+     fit2 <- lm(RI ~ Na + Mg + K + Ca + Ba, data = datos)
+     pre1 <- predict(fit1, datp)
+     pre2 <- predict(fit2, datp)
+     Solmod1[i, j] <- sqrt(sum((datp$RI - pre1)^2)/length(pre1))
+     Solmod2[i, j] <- sqrt(sum((datp$RI - pre2)^2)/length(pre2))
+   }
+ }
> mod.bruto <- round(mean(apply(Solmod1, 1, mean)), 3)
> mod.AIC <- round(mean(apply(Solmod2, 1, mean)), 3)
> mod.bruto

[1] 1.111

> mod.AIC

[1] 0.983

```

El modelo calculado con mínimo AIC tiene mejor habilidad predictiva ya que en el proceso de cálculo va teniendo en cuenta el efecto de la inclusión de otras variables en el modelo.

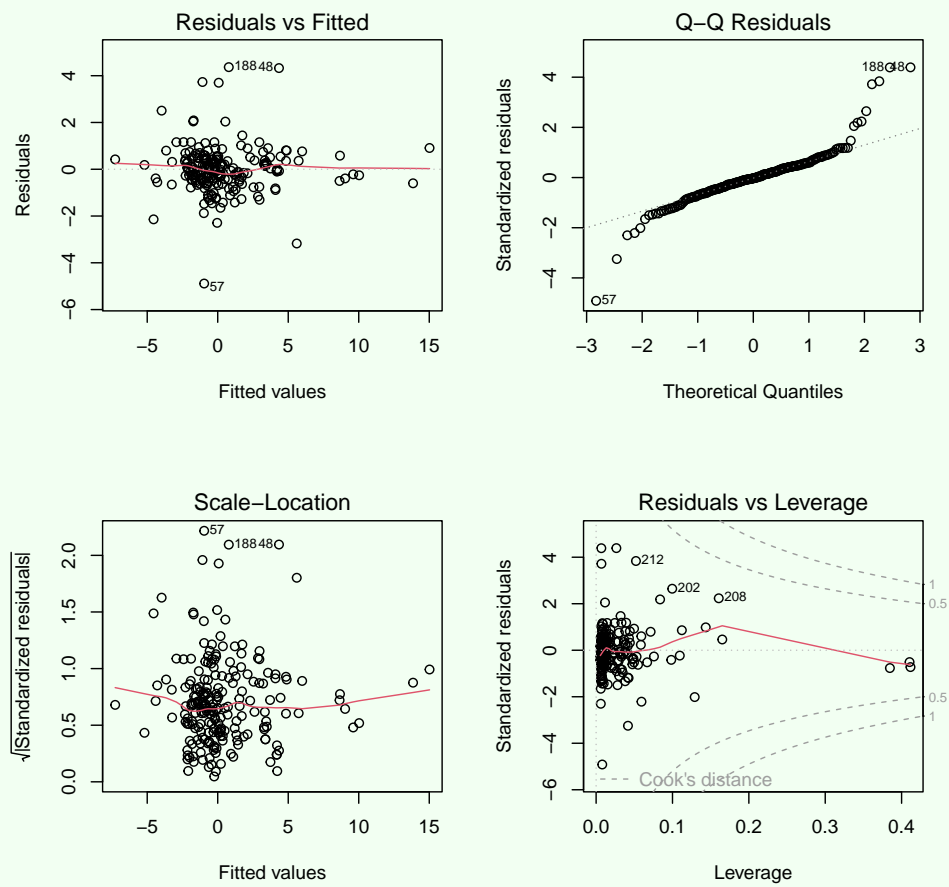


Figura 12: Gráfico de residuos en el modelo ModTf estimado

4.3. Inclusión de variables cualitativas en el modelo

Las variables dependientes que pretendemos estudiar mediante un modelo de regresión múltiple podrían estar influenciadas por variables cualitativas. Por ejemplo en un variable cualitativa, x_i con dos niveles, *categorías*, se transforma en una *dummy variable*, que toma dos posibles valores dependiendo de la pertenencia o no a la categoría.

$$x_i = \begin{cases} 1 & \text{pertenece} \\ 0 & \text{no pertenece} \end{cases} \quad (9)$$

y por tanto el modelo se formulará de la siguiente manera,

$$y_i = \beta_0 + \beta_1 x_i + e_i = \begin{cases} \beta_0 + \beta_1 + e_i & \text{pertenece} \\ \beta_0 + e_i & \text{no pertenece} \end{cases} \quad (10)$$

Variables cualitativas de uso frecuente en estudios de regresión son el sexo, el nivel de estudios, variables que indiquen área geográfica ó lugar, tratamiento, padecer una enfermedad, y otras muchas que dependerán del tipo de variable a estudiar y el diseño estadístico.

Si tenemos una variable categórica con k niveles, tendríamos que crear k variables *dummies* con el esquema descrito en 9

Supongamos ahora un modelo con una variable continua, X_1 , una variable cualitativa con dos niveles, X_2 . Sabemos que,

$$x_2 = \begin{cases} 1 & \text{pertenece} \\ 0 & \text{no pertenece} \end{cases} \quad (11)$$

y el modelo quedará formulado de la siguiente forma,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \beta_2 + e_i & \text{pertenece} \\ \beta_0 + \beta_1 x_{1i} + e_i & \text{no pertenece} \end{cases} \quad (12)$$

Cuando la observación i pertenece a la categoría, tenemos $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 + e_i$ frente a $\beta_0 + \beta_1 x_{1i} + e_i$ si la observación no pertenece a la categoría. Es decir que cuando i pertenece a la categoría añadimos *algo más* al termino independiente, resultando en dos rectas paralelas, una para cada categoría.

Si añadimos una interacción entre la variable cualitativa y cuantitativa, $X_1 X_2$, ahora tenemos,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \beta_2 + \beta_3 x_{1i} + e_i & \text{pertenece} \\ \beta_0 + \beta_1 x_{1i} + e_i & \text{no pertenece} \end{cases} \quad (13)$$

Para el caso con interacción, si la observación i pertenece a la categoría, tenemos $y_i = \beta_0 + (\beta_1 + \beta_3) x_{1i} + \beta_2 + e_i$ frente a $\beta_0 + \beta_1 x_{1i} + e_i$ si la observación no pertenece a la categoría. En este caso no se da un desplazamiento de las dos rectas de regresión y además una diferencia entre las pendientes de las rectas de regresión dependiendo de la pertenencia o no a la categoría, $(\beta_1 + \beta_3) \neq \beta_1$

Ejemplo. Cristales y variables cualitativas

En los datos originales `fgl` hay una columna que eliminamos al principio, llamada `type`, que es cualitativa. Si hacemos un gráfico de cajas y bigotes para cada nivel de la variable `type` podemos ver si hay diferencias entre cada nivel

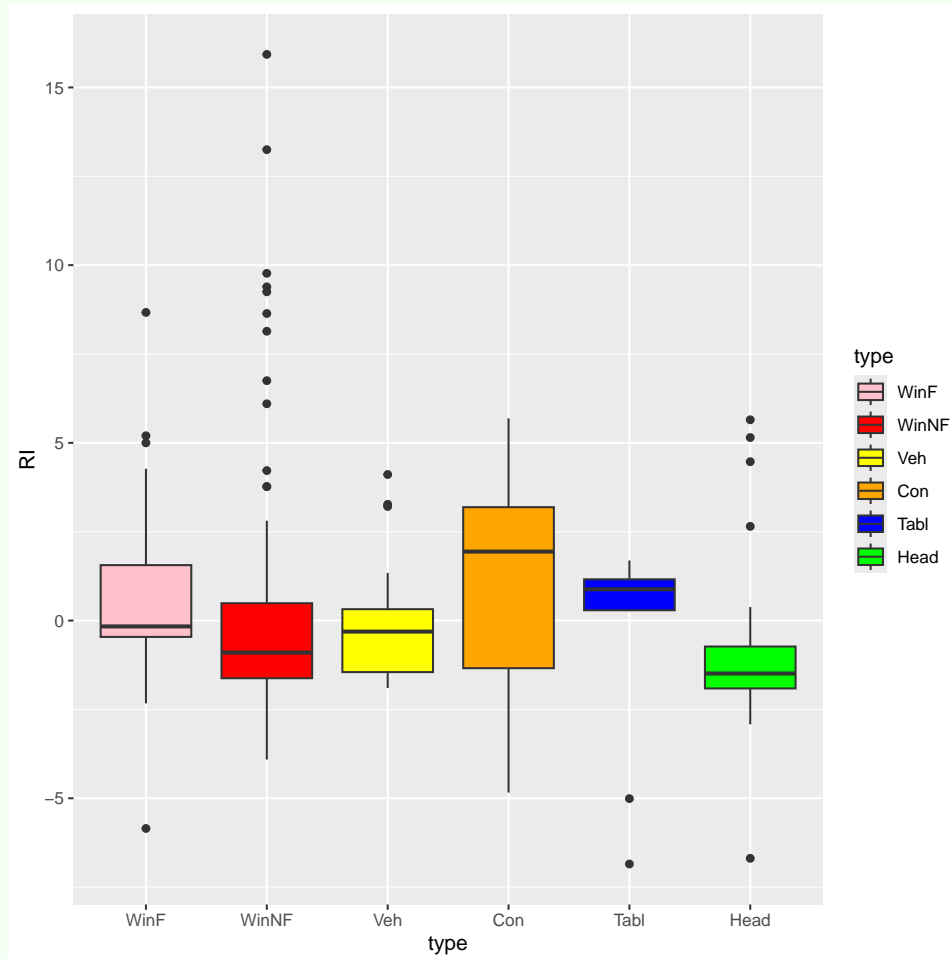


Figura 13: Gráfico de cajas y bigotes de `RI` vs `type`

Vamos a plantear un modelo para `RI` con la variable cuantitativa `Ca` y la cualitativa `type`. De manera matricial el modelo necesita la matriz **X**

```

> X <- model.matrix(~Ca + type, data = fgl)
> head(X)

  (Intercept)   Ca typeWinNF typeVeh typeCon typeTabl typeHead
1           1 8.75          0       0       0         0       0
2           1 7.83          0       0       0         0       0
3           1 7.78          0       0       0         0       0
4           1 8.22          0       0       0         0       0
5           1 8.07          0       0       0         0       0
6           1 8.07          0       0       0         0       0

> XX <- crossprod(X)
> XX

              (Intercept)          Ca typeWinNF typeVeh typeCon typeTabl typeHead
(Intercept)    214.0    1916.79        76.0    17.00    13.00     9.00    29.00
Ca              1916.8 17600.02       689.6   149.31   131.61    84.21   246.25
typeWinNF        76.0    689.60        76.0     0.00     0.00     0.00     0.00
typeVeh          17.0    149.31         0.0    17.00     0.00     0.00     0.00
typeCon          13.0    131.61         0.0     0.00    13.00     0.00     0.00
typeTabl         9.0     84.21         0.0     0.00     0.00     9.00     0.00
typeHead        29.0    246.25         0.0     0.00     0.00     0.00    29.00

> y <- fgl$RI
> b <- solve(XX) %*% t(X) %*% y
> b

              [,1]
(Intercept) -15.10944
Ca           1.79916
typeWinNF    -0.59702
typeVeh      -0.72895
typeCon      -2.17729
typeTabl     -2.26915
typeHead     -1.05170

```

Al construir la matriz $\mathbf{X}'\mathbf{X}$ vemos que uno de los niveles de la variable *type* no aparece. Eso sucede por que al incluir una variable cualitativa uno de los niveles ha de hacerse cero, de otro modo la matriz $\mathbf{X}'\mathbf{X}$ tendría determinante nulo.

```

> modCat <- lm(RI ~ Ca + type, data = fgl)
> summary(modCat)$coefficients

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.10944    0.765819 -19.7298 1.9051e-49
Ca           1.79916    0.083987  21.4220 2.0105e-54
typeWinNF    -0.59702    0.280147  -2.1311 3.4261e-02
typeVeh      -0.72895    0.455677  -1.5997 1.1119e-01
typeCon      -2.17729    0.521017  -4.1789 4.3196e-05
typeTabl     -2.26915    0.598625  -3.7906 1.9710e-04
typeHead     -1.05170    0.373054  -2.8192 5.2822e-03

> fgl$predlm = predict(modCat)

```

Aquí ajustamos una pendiente común para la variable Ca , $\beta_1 = 1.799$ y el efecto *type* hace que se modifique el término independiente, $\beta_0 = -15.109$, dando un patrón de rectas paralelas, una para cada nivel.

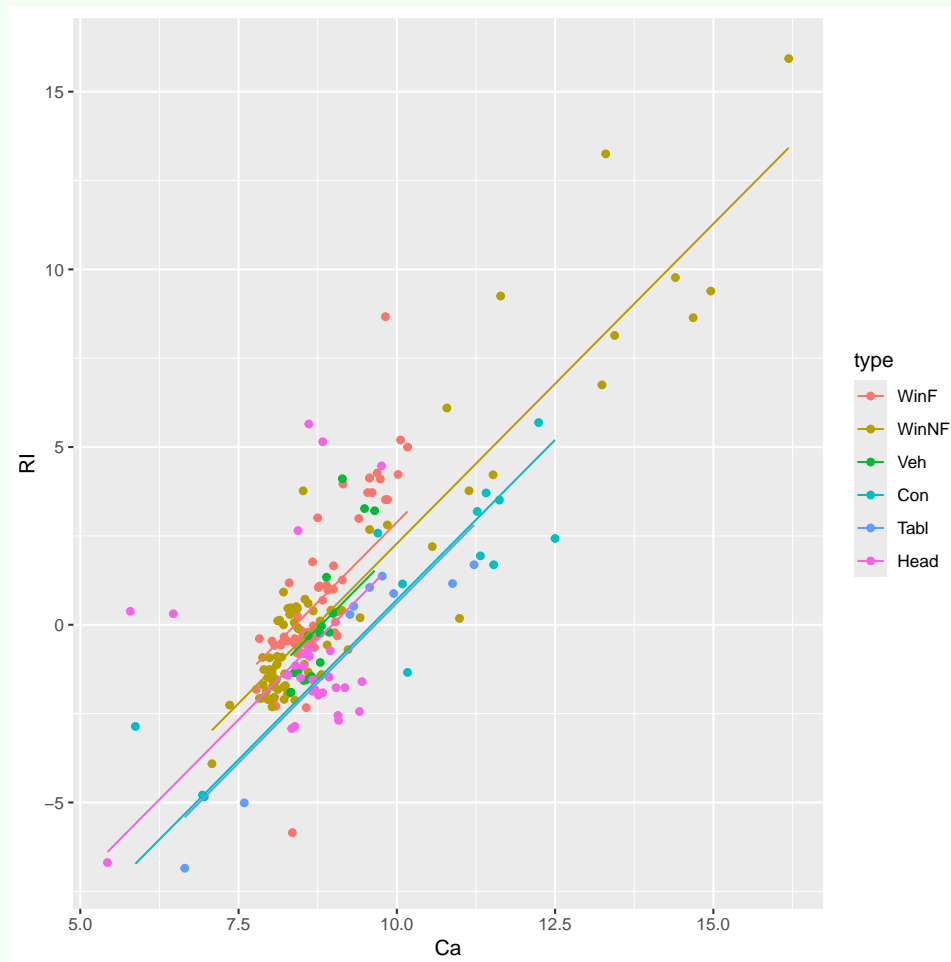


Figura 14: Ajuste de una recta de regresión con la variable Ca y *type*

Podemos ajustar una recta para cada nivel de la variable *type* pensando que existe interacción.

```
> modCat2 <- lm(RI ~ Ca + type + Ca:type, data = fgl)
> summary(modCat2)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-28.6495	2.80249	-10.2229	4.8795e-20
Ca	3.3383	0.31790	10.5012	7.2974e-21
typeWinNF	12.5241	2.92732	4.2783	2.9053e-05
typeVeh	-10.9772	9.21233	-1.1916	2.3483e-01
typeCon	15.8184	3.48674	4.5367	9.7921e-06
typeTabl	9.4276	4.48357	2.1027	3.6730e-02
typeHead	22.8131	3.76743	6.0553	6.7281e-09
Ca:typeWinNF	-1.4929	0.33072	-4.5142	1.0786e-05
Ca:typeVeh	1.1693	1.04769	1.1161	2.6570e-01
Ca:typeCon	-1.9792	0.37592	-5.2650	3.5717e-07
Ca:typeTabl	-1.3421	0.48789	-2.7508	6.4844e-03
Ca:typeHead	-2.7550	0.43345	-6.3560	1.3508e-09

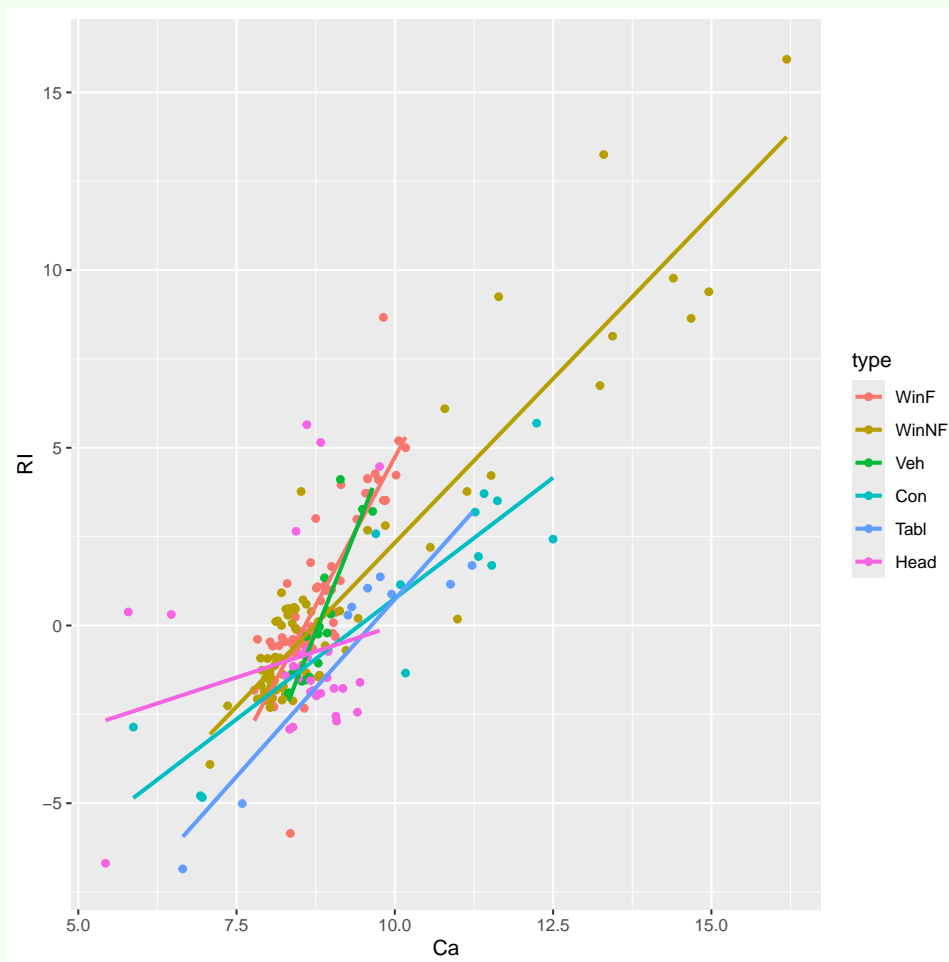


Figura 15: Ajuste de una recta de regresión con la variable *Ca* y *type*, donde se da una pendiente común para *Ca*, su incremento para cada uno de los niveles de *type*, y el incremento en la ordenada en el origen

El término independiente de este modelo, $\beta_0 = -28.65$, se ve modificado para cada nivel, por ejemplo para WinNF, $\beta_{0_{typeWinNF}} = 12.524$.

El modelo estima una pendiente general para la variable Ca , $\beta_1 = 3.338$, pero dicha pendiente es modificada para cada nivel de $type$. En el caso de $type = WinNF$, la pendiente pasa a ser $\beta_{1_{typeWinNF}} = (3.338 + -1.493) = 1.845$

Por último podemos preguntarnos que modelo de los dos planteados es mejor, lo haremos planteando una tabla ANOVA,

```
> anova(modCat, modCat2)
```

```
Analysis of Variance Table
```

```
Model 1: RI ~ Ca + type
```

```
Model 2: RI ~ Ca + type + Ca:type
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	207	588				
2	202	465	5	122	10.6	4.4e-09 ***

```
---
```

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

Se puede comprobar que se prefiere el modelo con la interacción, entre la variable cuantitativa y cualitativa. En términos de R^2 , tenemos que el modelo sin interacción $R^2 = 0.701$ y con interacción $R^2 = 0.763$

En general habrá que tener en cuenta el mecanismo generador de los datos ó plausibilidad para elegir uno de los tipos de modelos.