Apuntes de Econometria aplicada a la Administración y Dirección de Empresas

Jose Luis Gallego y Francisco Parra

Primera Parte: Modelo Lineal General

Modelo de Regresión Lineal

Introducción

El concepto de regresión es uno de los pilares de la estadística, y data al menos de principio de 1800 con los trabajos de Legendre, Gauss y Laplace. Es posible que el término *regresión* sea debido a Francis Galton, quien acuñó el término "regresión hacia la media" para describir la observación de que los hijos de padres muy altos tienden a ser algo más bajos que sus progenitores, y por el contrario los hijos de padres muy bajos suelen ser algo más altos, y por lo tanto acercarse en ambos casos más a la media de la población. Este fenómeno que se produce en muchas más circunstancias en la naturaleza, queda muy bien explicado por Stephen M. Stiegler con el siguiente ejemplo. Supongamos que efectuamos en dos momentos diferentes de tiempo un examen sobre una materia concreta a un alumno, y evaluamos primero uno de ellos, observando que obtiene una nota mucho más alta que la media de sus compañeros de clase. ¿Cómo de buena esperamos que sea la puntuación en el segundo examen? Probablemente alta, pero también probablemente no tan alta como en la primera ocasión, ya que probablemente el gran éxito en la primera ocasión se deba a dos componentes: por un lado la capacidad del alumno (componente estable o permanente) y por otro un cierto grado de suerte (componente transitorio y en cierta medida aleatorio). El coeficiente que medía esa regresión hacia la media pasó desde entonces a indicarse con la letra r.

La **regresión lineal** es la técnica básica del análisis econométrico. Mediante dicha técnica tratamos de determinar relaciones de dependencia de tipo lineal entre una variable dependiente o endógena, respecto de una o varias variables explicativas o exógenas. Gujarati (1975), define el análisis de regresión como el estudio de la dependencia de la variable dependiente, sobre una o más variables explicativas, con el objeto de estimar o predecir el valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en medias muestrales repetidas) de las últimas.

- Ejemplo 1.1 El precio de la vivienda. Se desea explicar el precio de los pisos en el distrito de Puerto Chico de Santander. Para ello se consulta un portal inmobiliario y se obtiene una relación de todas las ofertas. Para la mayoría de los pisos se dispone de las siguientes variables: precio, metros cuadrados, antigüedad, necesita reforma y vistas a la bahía. Las dos últimas variables se codifican con unos y ceros para indicar si presentan o no presentan la característica correspondiente. ¿Cuál es el tipo de relación (positiva, negativa o nula) que cabe esperar entre la variable dependiente y las variables explicativas?
- Ejemplo 1.2 El precio de un coche usado. Se desea explicar el precio de los coches Seat Ibiza en España. Para ello se consulta un portal de coches de segunda mano y se encuentra una gran variedad de ofertas con la siguiente información: precio, kilometraje, año de matriculación, combustible, cilindrada, número de puertas. ¿Cuál es el tipo de relación (positiva, negativa o nula) que cabe esperar entre la variable dependiente y las variables explicativas?

El modelo de regresión lineal general expresa una variable dependiente como una función lineal de varias variables independientes más un término de error aleatorio,

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i, i = 1, 2, ..., n$$

(1)

en donde

- Y_i es la observación i-ésima de la variable dependiente, explicada o respuesta;
- X_{ji} es la observación i-ésima de la j-ésima variable independiente, explicativa o entrada;
- β_1 es el término constante, también llamado, ordenada;
- $\beta_2, ..., \beta_k$ son los coeficientes de regresión (pendientes) asociados a las variables explicativas;
- $\beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki}$ es la función lineal de las variables explicativas o hiperplano de de regresión poblacional;
- y u_i es un término de error aleatorio, también llamado perturbación o innovación aleatoria.

La ecuación de regresión puede escribirse como:

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ji} + u_i$$

donde $X_{1i} = 1$ es una variable constante.

Otra forma común de escribir el modelo de regresión lineal general es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_i, i = 1, 2, ..., n,$$

que incluye k variables explicativas más un término constante, en total $k + 1$ parámetros.

El resultado del análisis de regresión es el *modelo de regresión estimado*

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki} + \hat{u}_i, i = 1, 2, ..., n,$$

que se obtiene reemplazando los parámetros desconocidos $\beta_j (j = 1, ..., k)$ y los errores $u_i (i = 1, ..., n)$ del modelo de regresión poblacional (1) por las estimaciones $\hat{\beta}_j (j = 1, ..., k)$ y $\hat{u}_i (i = 1, ..., n)$.

El modelo de regresión estimado puede usarse para contrastar teorías económicas, simular políticas económicas y predecir variables económicas, que son tres objetivos fundamentales de la Econometría.

Modelo de Regresión Lineal Simple

El modelo de regresión lineal simple (RLS) es

- $Y_t = \beta_1 + \beta_2 X_t + u_t, t = 1, ..., n$ u_t iid $N(0, \sigma^2)$ (2)

donde

- $Y_t : t = 1, ..., n$ es una muestra aleatoria de la variable dependiente Y, siendo Y_t la observación t-ésima de la muestra;
- $X_t : t = 1, ..., n$ es una muestra fija de la variable independiente X, siendo X_t la observación t-ésima de X;
- β_1 y β_2 son, respectivamente, la ordenada y la pendiente de la recta de regresión poblacional $\beta_1 + \beta_2 X_t$;
- $u_t : t = 1, ..., n$ es un conjunto de errores aleatorios mutuamente independientes e idénticamente distribuidos según una normal con media cero y varianza constante σ^2, $u_t = iidN(0, \sigma^2)$.

El modelo R.L.S. parte de la existencia de una relación lineal entre la variable endógena (Y_t) y la variable exógena (X_t), de manera que el objetivo del investigador consiste en estimar los dos parámetros del modelo (2) a partir de los datos muestrales de los que disponemos. Para ello se utiliza el método de los **Mínimos**

Cuadrados Ordinarios (MCO), que requiere plantear ciertas hipótesis sobre el comportamiento de las variables que integran el modelo.

La variable u_t que se denomina término de perturbación o error, recoge todos aquellos factores que pueden influir a la hora de explicar el comportamiento de la variable Y_t y que, sin embargo, no están reflejados en la variables explicativa, X_t. Estos factores deberían ser poco importantes, ya que no debería existir ninguna variable explicativa relevante omitida en el modelo de regresión. En caso contrario se estaría incurriendo en lo que se conoce como un error de especificación del modelo. El término de perturbación también recogería los posibles errores de medida de la variable dependiente, Y_t.

De lo anterior se desprende que, a la hora de estimar los parámetros del modelo, resultará de vital importancia que dicho término de error no ejerza ninguna influencia determinante en la explicación del comportamiento de la variable dependiente. Por ello, si el modelo está bien especificado, cuando se aplica el método de Mínimos Cuadrados Ordinarios, cabe realizar las siguientes hipótesis de comportamiento sobre el término de error:

1. La esperanza matemática de u_t es cero, tal que $E(u_t) = 0$. Es decir, el comportamiento del término de error no presenta un sesgo sistemático en ninguna dirección determinada. Por ejemplo, si estamos realizando un experimento en el cual tenemos que medir la longitud de un determinado objeto, a veces al medir dicha longitud cometeremos un error de medida por exceso y otras por defecto, pero en media los errores estarán compensados.

2. La covarianza entre u_t y u_s es nula, y por tanto, $E(u_t u_s) = 0$. Ello quiere decir que el error cometido en un momento determinado, t, no debe estar correlacionado con el error cometido en otro momento del tiempo, s, o dicho de otro modo, los errores no ejercen influencia unos sobre otros. En caso de existir este tipo de influencia o correlación, nos encontraríamos ante el problema de la autocorrelación en los residuos, el cual impide realizar una estimación por Mínimos Cuadrados válida.

3. La matriz de varianzas y covarianzas del término de error debe ser escalar, tal que $Var(u_t) = \sigma^2$, $i = 1, \ldots, n$. Dado que siempre que medimos una variable, se produce un cierto error, resulta deseable que los errores que cometamos en momentos diferentes del tiempo sean similares en cuantía. Esta condición es lo que se conoce como supuesto de homocedasticidad que, en caso de no verificarse, impediría un uso correcto de la estimación lineal por Mínimos Cuadrados.

Estas hipótesis implican que los errores siguen una distribución Normal de media cero y varianza constante por lo que, dado su carácter aleatorio, hace que los errores sean por naturaleza impredecibles.

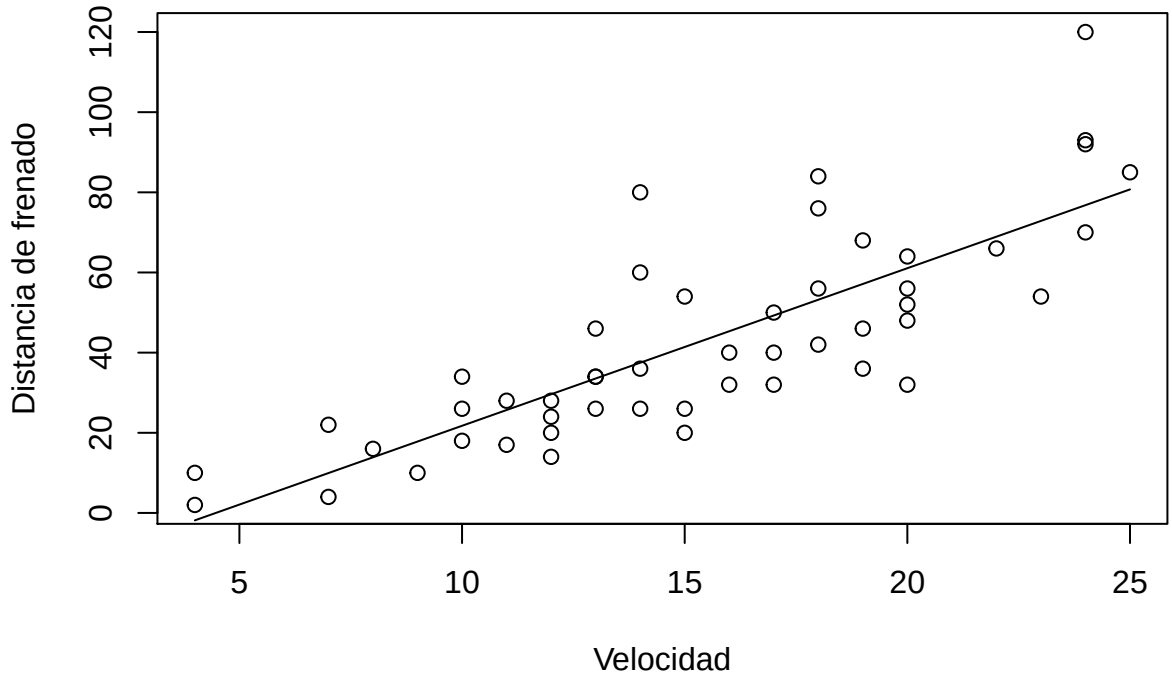
Asimismo, las variables incluidas en el modelo deben verificar que:

1. La variable dependiente Y_t se ajusta al modelo lineal durante todo el periodo muestral, es decir, no se produce un cambio importante en la estructura de comportamiento de Y_t a lo largo de la muestra considerada. Por tanto, se distribuirá como una normal con $\mathrm{E}(Y_t) = \beta_1 + \beta_2 X_t$ y $\mathrm{Var}(Y_t) = \mathrm{Var}(u_t)$.

2. La variable explicativa, X_t, es no estocástica, es decir, es considerada fija en muestreos repetidos.

Si suponemos que se verifican los supuestos anteriores, la estimación mínimo cuadrática de los parámetros β_1 y β_2, dará como resultado gráfico una recta que se ajusta lo máximo posible a la nube de puntos definida por todos los pares de valores muestrales (X_t, Y_t), tal y como se puede apreciar en el Figura 1.

Figura 1. Línea de regresión ajustada. Distancia de frenado ~ Velocidad



El término de error, u_t, puede ser entendido, a la vista del gráfico anterior, como la distancia que existe entre el valor observado, Y_t, y el correspondiente valor estimado, que sería la imagen de X_t en el eje de ordenadas. El objetivo de la estimación por Mínimos Cuadrados Ordinarios es, precisamente, minimizar el sumatorio de todas esas distancias al cuadrado; es decir:

$$Min \sum_{t=1}^{n} (u_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t)^2$$

Derivando esta expresión respecto a los coeficientes $\hat{\beta}_1$ y $\hat{\beta}_2$ e igualando a cero obtenemos el *sistema de ecuaciones normales*:

$$\begin{array}{c} \sum_{t=1}^{n} Y_i = n\hat{\beta}_1 + \sum_{t=1}^{n} \hat{\beta}_2 X_t \\ \sum (Y_t X_t) = \sum_{t=1}^{n} \hat{\beta}_1 X_t + \sum_{t=1}^{n} \hat{\beta}_1 (X_t)^2 \end{array}$$

donde n representa el tamaño muestral.

Resolviendo dicho sistema de ecuaciones obtenemos la solución para los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$:

$$\begin{array}{c} \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \\ \hat{\beta}_2 = \frac{\sum_{t=1}^{n} ((Y_t - \bar{Y})(X_t - \bar{X}))}{\sum (X_t - \bar{X})^2} \end{array}$$

La pendiente $\hat{\beta}_2 = \frac{S_{XY}}{SXX}$, puede calcularse como:

$\hat{\beta}_2 = \frac{cov(XY)}{\sigma_X^2}$

$\hat{\beta}_2 = \frac{\sum_{t=1}^{n} ((X_t - \bar{X})Y_t)}{\sum (X_t - \bar{X})X_t}$

$\hat{\beta}_2 = \frac{\sum_{t=1}^{n} X_t Y_t - n\bar{X}\bar{Y}}{\sum X_t^2 - \bar{X}^2}$

Las propiedades numéricas del método de mínimos cuadrados son:

1. La suma de los residuos es cero,

$\sum_{t=1}^{n} \hat{u}_t = 0$

2. Los residuos y la variable explicativa son ortogonales,

$\sum_{t=1}^{n} \hat{u}_t x_t = 0$

Conviene notar que estas dos propiedades numéricas implican que los valores ajustados y los residuos son ortogonales:

$\sum_{t=1}^{n} \hat{u}_t \hat{Y}_t = 0$

Las dos propiedades numéricas,$\sum_{t=1}^{n} \hat{u}_t = 0$ y $\sum_{t=1}^{n} \hat{u}_t x_t = 0$ son dos restricciones que permiten calcular los residuos a partir de los $n-2$ restantes.En otras palabras,hay $n-2$ residuos que pueden variar libremente.

En consecuencia los grados de libertad de la suma de cuadrados de los residuos son $n-2$. Lo que hace razonable el estimar la varianza del error, $\hat{\sigma}^2$ dividiendo la suma de cuadrados de los residuos por sus grados de libertad:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n} \hat{u}^2}{n-2}$$

Modelo de Regresión Lineal Múltiple

Pasamos a continuación a generalizar el modelo anterior al caso de un modelo con varias variables exógenas, de tal forma que se trata de determinar la relación que existe entre la variable endógena Y y las variables exógenas: $X_1, X_2, \ldots, X_k$. Dicho modelo se puede formular matricialmente de la siguiente manera:

$$y = \beta X + u \Rightarrow Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i, i = 1, 2, \ldots, n$$

donde:

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix}$$

es el vector de observaciones de la variable endógena

$$X = \begin{bmatrix} X_{11} & X_{21} & X_{21} & ... & X_{k1} \\ X_{12} & X_{22} & X_{32} & ... & X_{k2} \\ . & . & . & ... & . \\ X_{1n} & X_{2n} & X_{3n} & ... & X_{kn} \end{bmatrix}$$

es la matriz de observaciones de las variables exogenas

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix}$$

es el vector de los coeficientes que pretendemos estimar

$$u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_n \end{bmatrix}$$

es el vector de términos de error.

La primera columna de la matriz X, denominada **matriz de diseño** corresponde a un vector de n unos, siendo éste la variable que se considera asociada a la estimación del parámetro β_1, y el resto, de la segunda en adelante, corresponde a cada una de las variables explicativas consideradas.

$$X = \begin{bmatrix} 1 & X_{21} & X_{21} & ... & X_{k1} \\ 1 & X_{22} & X_{32} & ... & X_{k2} \\ . & . & . & ... & . \\ 1 & X_{2n} & X_{3n} & ... & X_{kn} \end{bmatrix}$$

Suponiendo que se verifican las hipótesis que veíamos antes, el problema a resolver nuevamente es la minimización de la suma de los cuadrados de los términos de error tal que:

$$\hat{u}'\hat{u} = Min \sum(\hat{u}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_k X_{ki})^2 = (y - X\hat{\beta})'(y - X\hat{\beta})$$

Desarrollando dicho cuadrado y derivando respecto a cada $\hat{\beta}_i$ obtenemos el *sistema de ecuaciones normales*:

$$\begin{aligned} n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{ki} &= \sum_{i=1}^{n} Y_i \\ \hat{\beta}_1 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{2i}X_{ki} &= \sum_{i=1}^{n} Y_i X_{2i} \\ &.... \\ \hat{\beta}_1 \sum_{i=1}^{n} X_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} X_{ki}X_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2 &= \sum_{i=1}^{n} Y_i X_{ki} \end{aligned}$$

Notese que si el modelo no tiene termino constante:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i, i = 1, 2, \ldots, n$$

El sistema de de ecuaciones normales queda :

$$\begin{aligned} \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}X_{1i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}X_{ki} &= \sum_{i=1}^{n} Y_i X_{1i} \\ \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{2i}X_{ki} &= \sum_{i=1}^{n} Y_i X_{2i} \\ &.... \\ \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}X_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} X_{ki}X_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2 &= \sum_{i=1}^{n} Y_i X_{ki} \end{aligned}$$

El sistema de ecuaciones normales expresado en notación matricial queda:

$$X'X\hat{\beta} = X'y$$

en donde basta con despejar $\hat{\beta}$ premultiplicando ambos miembros por la inversa de la matriz $(X'X)$ para obtener la estimación de los parámetros del modelo tal que:

$$\hat{\beta} = (X'X)^{-1}X'y$$

siendo:

$$X'X = \begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & ... & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum(X_{2i}X_{3i}) & ... & \sum(X_{1i}X_{ki}) \\ . & . & . & ... & . \\ \sum X_{ki} & \sum(X_{ki}X_{2i}) & \sum(X_{ki}X_{3i}) & ... & \sum X_{ki}^2 \end{bmatrix}$$

$$X'y = \begin{bmatrix} \sum Y_i \\ \sum(Y_iX_{2i}) \\ . \\ \sum(Y_iX_{ki}) \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ \hat{\beta}_k \end{bmatrix}$$

Utilizando el calculo matricial el proceso de minimización se realiza así:

$$\begin{aligned} \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta}) = (y' - \hat{\beta}'X')(y - X\hat{\beta}) = \\ = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \beta'X'X\hat{\beta} = \\ = y'y - 2\hat{\beta}'X'y + \beta'X'X\hat{\beta} \end{aligned}$$

La condición necesaria para que la función $\hat{u}'\hat{u}$ tenga un mínimo en β requiere que el vector de primeras derivadas sea un vector nulo,

$$\frac{\partial(\hat{u}'\hat{u})}{\partial\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0_k$$

de donde se obtiene $X'X\hat{\beta} = X'y$.

La condición suficiente para que la función $\hat{u}'\hat{u}$ tenga un mínimo en β requiere que la matriz hessiana o matriz de segundas derivadas sea definida positiva,

$$\frac{\partial^2(\hat{u}'\hat{u})}{\partial\hat{\beta}\partial\hat{\beta}} = 2X'X$$

La matriz cuadrada $X'X$ de orden k será definida positiva cuando las columnas de X sean linealmente independientes.

Notese que el vector $\hat{\beta}$ no tiene solución cuando las columnas de X son linealmente dependientes. En este caso, no existirá el estimador de mínimos cuadrados porque la matriz $(X'X)$ es singular o no-invertible. Este problema se denomina multicolinealidad exacta.

La propiedad numérica fundamental del método de mínimos cuadrados se deriva directamente de la condición necesaria de extremo:

$$X'\hat{u} = \begin{bmatrix} \sum_{i=1}^{n} \hat{u}_i \\ \sum_{i=1}^{n} \hat{u}_i X_{2i} \\ . \\ \sum_{i=1}^{n} \hat{u}_i X_{ki} \end{bmatrix} = 0_k$$

Notese que si el modelo no tiene termino constante, el resultado de la multiplicación matricial anterior queda:

$$X'\hat{u} = \begin{bmatrix} \sum_{i=1}^{n} \hat{u}_i X_{1i} \\ \sum_{i=1}^{n} \hat{u}_i X_{2i} \\ . \\ \sum_{i=1}^{n} \hat{u}_i X_{ki} \end{bmatrix} = 0_k$$

que no implica la propiedad numérica de que $\sum_{i=1}^{n} \hat{u}_i = 0$

Las k propiedades numericas son restricciones que permiten calcular los residuos a partir de los $n-k$ restantes.En otras palabras, hay $n-k$ residuos que pueden variar libremente.

En consecuencia los grados de libertad de la suma de cuadrados de los residuos son $n-k$. Lo que hace razonable el estimar la varianza del error, $\hat{\sigma}^2$ dividiendo la suma de cuadrados de los residuos por sus grados de libertad:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n} \hat{u}^2}{n-k}$$

Coeficiente de determinación

Una vez estimada la ecuación de regresión lineal tiene interés determinar la exactitud del ajuste realizado. Para ello hay que analizar la variación que experimenta esta variable dependiente y, dentro de esta variación, se estudia qué parte está siendo explicada por el modelo de regresión y qué parte es debida a los errores o residuos.

La forma de realizar dicho análisis es a partir de la siguiente expresión:

$$SCT = SCE + SCR$$

(3)

donde:

- SCT: es la Suma de Cuadrados Totales y representa una medida de la variación de la variable dependiente.
- SCE es la Suma de Cuadrados Explicados por el modelo de regresión.
- SCR es la Suma de Cuadrados de los Errores

En los modelos sin término constante no se cumple la proposición (3)

Cuando el modelo tiene término independiente, cada una de estas sumas viene dada por:

$$SCT = y'y - n\bar{Y}^2 = \sum_{t=1}^{n} Y_i^2 - n\bar{Y}^2$$

$$SCE = (\hat{\beta}' X')(X\hat{\beta}) - n\bar{Y}^2 = \sum_{t=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2$$

$$SCR = y'y - (\hat{\beta}' X')(X\hat{\beta}) = \sum_{t=1}^{n} Y_i^2 - \sum \hat{Y}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum_{t=1}^{n} \hat{u}_i^2$$

Los grados de libertad de la SCT son $n-1$, los de la SCR, $n-k$, y los de la SCE, $k-1$.

A partir de las expresiones anteriores es posible obtener una medida estadística acerca de la bondad de ajuste del modelo mediante lo que se conoce como coeficiente de determinación (R^2), que se define como:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Si el modelo incluye término constante, entonces $0 \leq R^2 \leq 1$.

Pueden darse tres situaciones:

1. $R^2 = 1$, relación exacta entre Y y $X_2, ..., X_k$: los residuos serán nulos, por lo que $SCR = 0$ y $R^2 = 1 - \frac{SCR}{SCT} = 1$. Se dice que el ajuste es perfecto.
2. $R^2 = 0$, inexistencia de relación exacta entre Y y $X_2, ..., X_k$: las pendientes $\hat{\beta}_2, ..., \hat{\beta}_k$ serán nulas, por lo que $SCE = 0$ y $R^2 = \frac{SCE}{SCT} = 0$.
3. $0 < R^2 < 1$, relación aproximada entre Y y $X_2, ..., X_k$:

a) $R^2 = \frac{SCE}{SCT} > 0$ porque las dos sumas de cuadrado son positivas;

b) $R^2 = \frac{SCE}{SCT} < 1$ porque $SCE < SCT$.

En los modelos sin término constante el coeficiente de determinación no esta acotado entre 0 y 1.

Dado que en una regresión sin ordenada el R^2 no esta acotado, algunos autores sugieren calcular el R^2 de una regresión sin ordenada como

$$R^2 = \frac{\sum_{i=1}^{n} \hat{Y}_i^2}{\sum_{i=1}^{n} Y_i^2}$$

Si el modelo incluye termino constante, el coeficiente de determinación coincide con el coeficiente de correlación simple entre Y y $\hat{Y}$, $R^2 = r_{Y\hat{Y}}^2$

Mediante este coeficiente es posible seleccionar el mejor modelo de entre varios que tengan el mismo número de variables exógenas, ya que la capacidad explicativa de un modelo es mayor cuanto más elevado sea el valor que tome este coeficiente. Sin embargo, hay que tener cierto cuidado a la hora de trabajar con modelos que presenten un R^2 muy cercano a 1 pues, aunque podría parecer que estamos ante el modelo "perfecto", en realidad podría encubrir ciertos problemas de índole estadística como la multicolinealidad que veremos en el siguiente apartado.

Por otra parte, el valor del coeficiente de determinación aumenta con el número de variables exógenas del modelo por lo que, si los modelos que se comparan tienen distinto número de variables exógenas, no puede establecerse comparación entre sus R^2. En este caso debe emplearse el coeficiente de determinación corregido ($\bar{R}^2$) , el cual depura el incremento que experimenta el coeficiente de determinación cuando el número de variables exógenas es mayor.

El coeficiente de determinación ajustado, $\bar{R}^2$, es el coeficiente de determinación corregido por los grados de libertad:

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{n-k}}{\frac{SCT}{n-1}} = 1 - \frac{n-1}{n-k}(1-R^2)$$

El R^2 puede calcularse a partir de las varianzas muestrales de Y , $\hat{Y}$ y $\hat{u}$ como:

$$R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$$

Transformaciones lineales

La transformación lineal de la variable Y es una función de la forma $T(Y_i) = a + bY_i$, donde a y b son números reales conocidos que definen el cambio de origen y el cambio de escala, respectivamente. Las dos transformaciones lineales que comúnmente se aplican en el análisis de regresión son los datos centrados y los datos tipificados.

Regresión con datos centrados

Los datos centrados o en desviaciones respecto a la media son:

$\tilde{Y}_i = Y_i - \bar{Y}$, $\tilde{X}_{2i} = X_{2i} - \bar{X}_2$,...,$\tilde{X}_{ki} = X_{ki} - \bar{X}_k$, $i = 1, 2, \ldots, n$

La regresión minimocuadrática con datos centrados:

$$\tilde{Y}_i = \hat{\beta}_2\tilde{X}_{2i} + \hat{\beta}_3\tilde{X}_{3i} + ... + \hat{\beta}_k\tilde{X}_{ki} + \hat{u}_i$$

proporciona las mismas pendientes y los mismos residuos que la regresión con datos crudos (1.2). Las $k-1$ pendientes se estiman resolviendo el sistema de $k-1$ ecuaciones normales.

Regresión con datos tipificados

Los datos tipificados son:

$Y^*_i = \frac{Y_i - \bar{Y}}{s_Y}$,$X^*_{2i} = \frac{X_{2i} - \bar{X}_2}{s_2}$,..., $X^*_{ki} = \frac{X_{ki} - \bar{X}_k}{s_k}$, $i = 1, 2, \ldots, n$

donde:

$s_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$,$s_2 = \sqrt{\frac{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}{n-1}}$, ...,$s_k = \sqrt{\frac{\sum_i^n (X_{ki} - \bar{X}_k)^2}{n-1}}$

Las pendientes estimadas del modelo de regresión lineal general miden la respuesta de la variable dependiente a cambios unitarios en las variables explicativas. El tamaño de estas estimaciones depende de las unidades en que se midan las variables. De aquí, para apreciar la importancia relativa de cada variable explicativa se usan datos tipificados.

En la regresión minimocuadrática con datos tipificados:

$$Y^*_i = \hat{\beta}^*_2 X^*_{2i} + \hat{\beta}^*_3 X^*_{3i} + ... + \hat{\beta}^*_k X^*_{ki} + \hat{u}^*_i$$

las pendientes y los residuos guardan las siguientes relaciones con las correspondientes estimaciones basadas en datos crudos (1.2):

$\hat{\beta}^*_j = \hat{\beta}_j \frac{s_j}{s_Y}$ y $\hat{u}^*_i = \frac{\hat{u}_i}{S_Y}$

Regresiones especiales

Todos los resultados establecidos para el modelo de regresión lineal general se particularizan fácilmente para los casos especiales del mismo.

Regresión sobre la constante

El modelo de regresión más sencillo es el que no incluye variables explicativas, $k = 1$. Se utilizará más adelante para realizar un análisis descriptivo de datos unidimensionales.

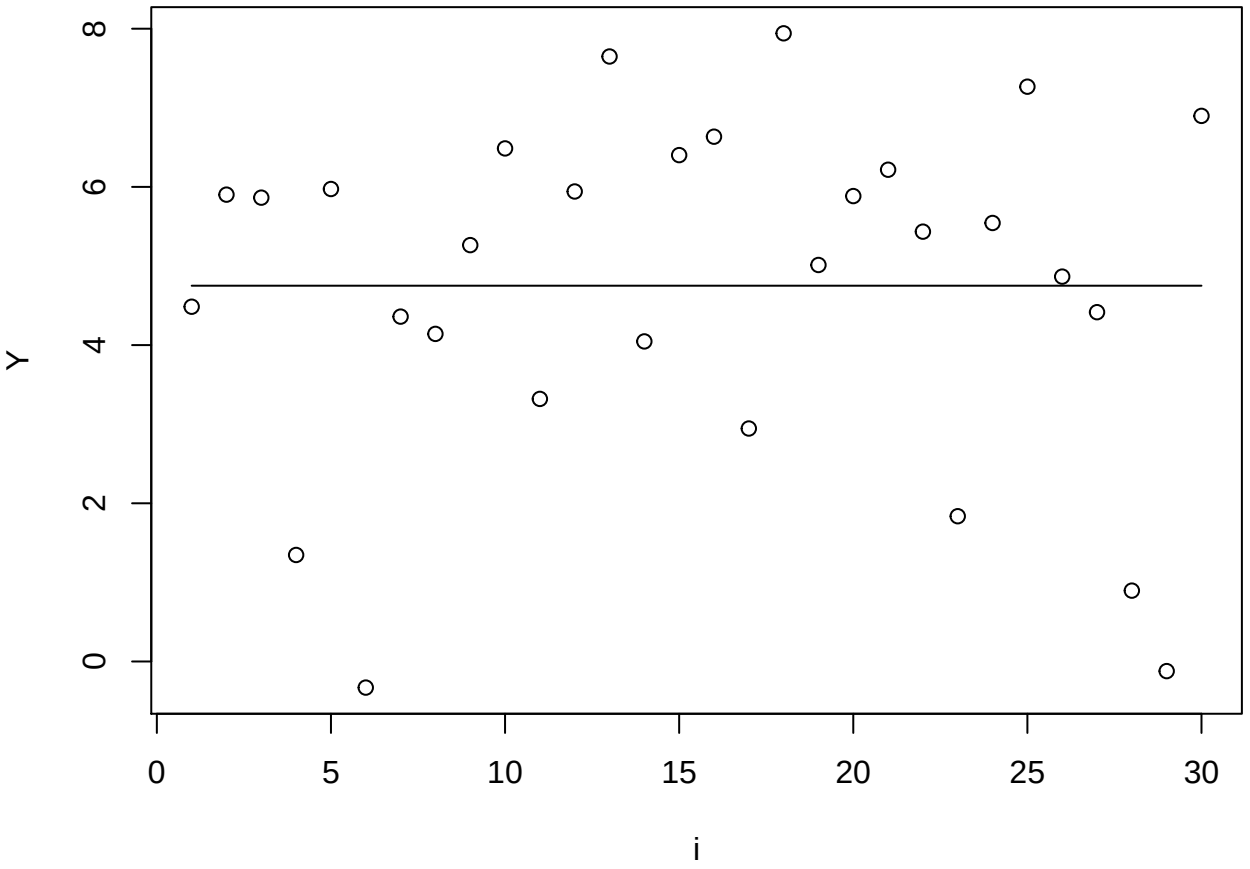
El modelo de regresión sobre una constante o sin predictores es:

$Y_i = \beta_1 + u_i,\ i = 1, ..., n$

en donde

- Y_i es un valor de la variable dependiente;
- β_i es una constante;
- y u_i es un error aleatorio.

Figura 2. Regresión sobre una constante



La estimación de mínimos cuadrados de la ecuación de regresión sobre una constante proporciona los siguientes resultados:

$\hat{\beta}_1 = \bar{Y}$, $\hat{Y}_i = \bar{Y}$ y $\hat{u}_i = Y_i - \bar{Y}$

Regresión sobre el orígen

Si la ordenada del modelo de regresión lineal simple se fija en cero, entonces la recta de regresión pasa por el origen de coordenadas.

El modelo de regresión simple a través del origen:

$Y_i = \beta X_i + u_i,\ i = 1, ..., n$

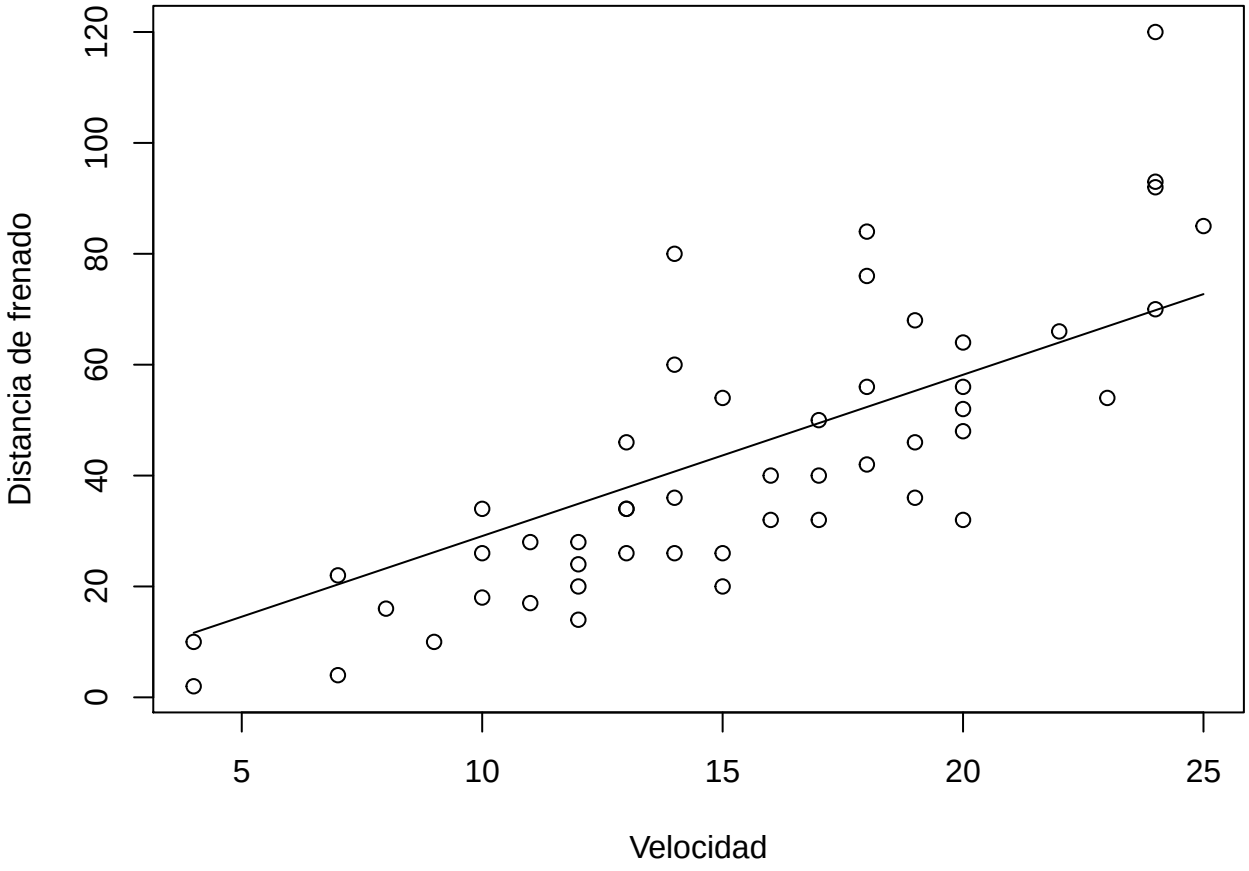
en donde

- Y_i es un valor de la variable dependiente,
- X_i es un valor de la variable independiente,
- βX_i es la recta de regresión poblacional que pasa por el origen de coordenadas,
- β es la pendiente de la recta de regresión,
- y u_i es un error aleatorio.

En la regresión a través del origen, el estimador minimocuadrático de β es:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

Figura 3. Regresión sobre el orígen



Regresión polinomial

El análisis de regresión con datos bivariantes puede extenderse ajustando a la nube de puntos un polinomio en X de orden r.

La regresión polinomial de orden r es

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_r X_i^r + u_i, i = 1, ..., n$$

en donde

- Y_i es un valor de la variable dependiente Y,
- X_i es un valor de la variable independiente X,
- $\beta_0 + \beta_1 X_i + ... + \beta_r X_i^r$ es el polinomio de regresión poblacional,
- $\beta_0, \beta_1, ..., \beta_r$ son los parámetros de regresión,
- y u_i es un error aleatorio.

En la regresión polinomial minimocuadrática

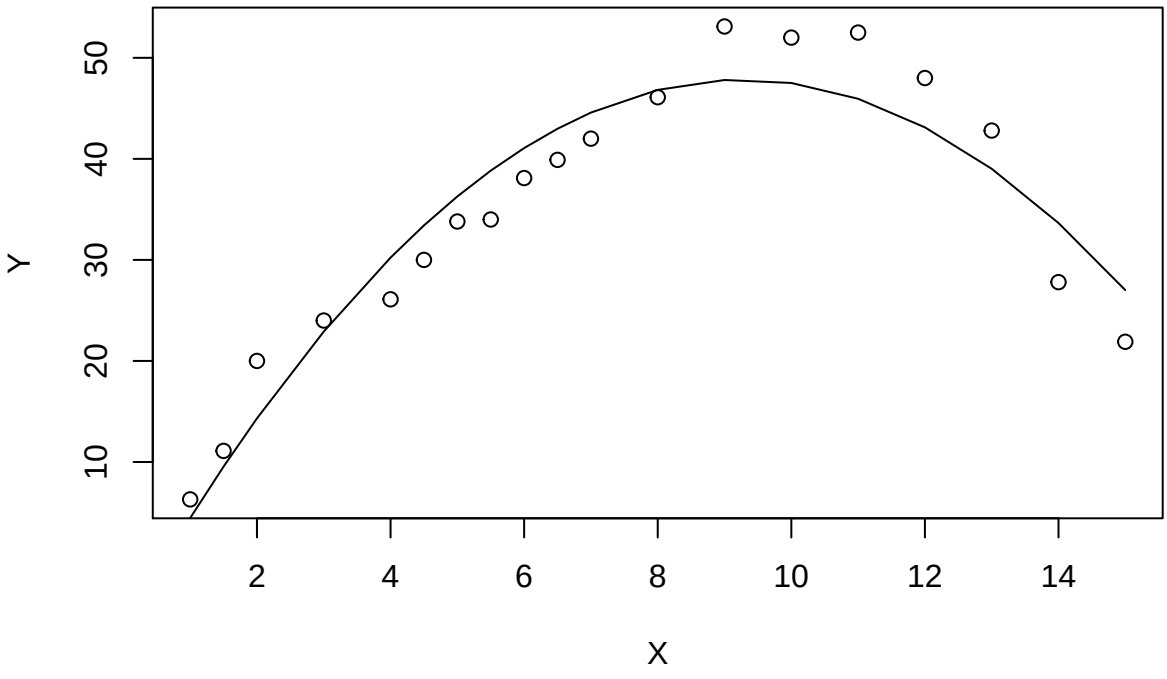
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \hat{u}_i$$

El sistema de ecuaciones normales es:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} Y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^3 &= \sum_{i=1}^{n} Y_i X_i \\ &.... \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i^2 + \hat{\beta}_1 \sum_{i=1}^{n} X_i^3 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^4 &= \sum_{i=1}^{n} Y_i X_i^2 \end{aligned}$$

La Figura 4 ilustra el ajuste de un polinomio cuadrático:

Figura 4. Regresión polinómica cuadrática



Regresión con variables ficticias

En un modelo econométrico, las variables representan a los conceptos u operaciones económicas que queremos analizar. Normalmente utilizamos variables cuantitativas, es decir, aquellas cuyos valores vienen expresados de

forma numérica; sin embargo, también existe la posibilidad de incluir en el modelo econométrico información cualitativa, siempre que esta pueda expresarse de esa forma.

Las variables cualitativas expresan cualidades o atributos de los agentes o individuos (sexo, religión, nacionalidad, nivel de estudios, etc.) y también recogen acontecimientos extraordinarios como guerras, terremotos, climatologías adversas, huelgas, cambios políticos, etc.

No cabe duda de que una forma de recoger factores de este tipo sería la utilización de variables proxy o aproximadas a las variables utilizadas. Por ejemplo, si quiero utilizar una variable que mida el nivel cultural de un país (variable cualitativa) puedo utilizar como variable proxy el número de bibliotecas existentes en un país, o representar una climatología adversa a partir de las temperaturas medias o precipitaciones. Sin embargo, no siempre es posible encontrar este tipo de variables y, en cualquier caso, debemos de ser conscientes de la posible existencia de errores en la definición de la variable.

Puesto que las variables cualitativas normalmente recogen aspectos de la presencia o no de determinado atributo (ser hombre o mujer, tener estudios universitarios o no tenerlos, etc.…) se utilizan variables construidas artificialmente, llamadas también ficticias o dummy, que generalmente toman dos valores, 1 ó 0, según se dé o no cierta cualidad o atributo. Habitualmente a la variable ficticia se le asigna el valor 1 en presencia de la cualidad y 0 en caso contrario. Las variables que toman valores 1 y 0, también reciben el nombre de variables dicotómicas o binarias.

Las variables dicotómicas pueden combinarse para caracterizar variables definidas por su pertenencia o no a un grupo. Si incluyo una variable cualitativa que me define la pertenencia o no de un país a un grupo, por ejemplo renta alta, media y baja, introduciré tres variables cualitativas en el modelo asociadas a la pertenencia o no a cada grupo; la primera caracterizaría a los individuos con renta alta, la segunda a los individuos con renta media, y la tercera a los individuos con renta baja.

Los modelos que utilizan variables cualitativas como regresores se diferencian en dos grupos, los modelos de Análisis de la Varianza o modelos ANOVA, que únicamente incluyen variables cualitativas como regresores; y los modelos de Análisis de la Covarianza o modelos ANCOVA que incluyen tanto variables cualitativas como cuantitativas. Los modelos ANOVA son muy utilizados en Sociología, Psicología, Educación, etc.; en Economía son más comunes los modelos ANCOVA.

Por ejemplo, al explicar el salario de los trabajadores, la variable cualitativa sexo es una variable explicativa a tener en cuenta. Si definimos dos variables ficticias: una, para indicar que la persona i pertenece al grupo de hombres (H_i); y otra, para indicar que pertenece al grupo de mujeres (M_i), es decir:

$$H_i = \begin{cases} 1 & \text{si } i \in \text{Hombres} \\ 0 & \text{si } i \notin \text{Hombres} \end{cases}$$

$$M_i = \begin{cases} 1 & \text{si } i \in \text{Mujeres} \\ 0 & \text{si } i \notin \text{Mujeres} \end{cases}$$

En la regresión del salario Y sobre las dos variables ficticias H y M,

$$\hat{Y}_i = \hat{\beta}_1 H_i + \hat{\beta}_2 M_i + \hat{u}_i, i = 1, ..., n$$

Los estimadores de mínimos cuadrados son:
$\beta_1 = \frac{\sum_{i=1}^{n} Y_i H_i}{\sum_{i=1}^{n} H_i^2} = \bar{Y}_1$ y $\beta_2 = \frac{\sum_{i=1}^{n} Y_i M_i}{\sum_{i=1}^{n} M_i^2} = \bar{Y}_2$

donde en donde $\bar{Y}_1$ es el salario medio de los hombres e $\bar{Y}_2$ es el salario medio de las mujeres.

Modelos linealizables

Los modelos de regresión no siempre son lineales, es decir, no siempre vienen expresados por la ecuación de una recta. Existen también modificaciones de esta ecuación de tal manera que se pueden practicar análisis de regresión cuadrática, cúbica, logarítmica,logística, etc.

Los principales modelos no lineales que se utilizan en estadística son:

Función potencial

$$Y_i = aX_i^b$$

En estas funciones puede tomarse logaritmos de forma que:

$$ln(Y)_i = log(a) + b \times log(X)_i$$

Si en esta transformación se hace un cambio de variables tal que $Y_i^* = log(Y)_i$ y $X_i^* = log(X)_i$ se estará ante una regresión lineal del tipo: $Y_i^* = a^* + b^* X_i^*$. Una vez determinadas a^* y b^*, se calcula a y b tomando antilogaritmos.

Función exponencial

El modelo es el siguiente:

$$Y_t = ab^{X_t}$$

Operando como en el caso anterior, se transforma en:

$$log(Y)_t = log(a) + x_t log(b)$$

Haciendo, en este caso, $Y_t^* = log(Y)_t$ calculamos a^* y b^* de modo que $Y_t^* = a^* + b^* X_t$. Una vez calculados obtenemos los parámetros originales a y b.

Función logarítmica

La ecuación es la siguiente:

$$Y_i = a + b \times log(X)_i$$

Basta con hacer el cambio $X_i^* = log(X)_i$ y tratarlo como una ecuación lineal.

Elasticidades y semielasticidades.

a) Elasticidad constante (log-log).

La relación entre Y y X se establece en términos de incrementos relativos:

$$log(Y)_t = \beta_1 + \beta_2 log(X)_t + u_t$$

β_2 es la elasticidad de Y respecto a X. Así:

$$\Delta\%Y = \beta_2 \Delta\%X$$

Según este modelo, se estima que por cada incremento de un 1 en X se produce un incremento de un $\beta_2\%$ en Y.

b) Semielasticidad (log-level).

El cambio se produce en términos porcentuales:

$$log(Y)_t = \beta_1 + \beta_2 X_t + u_t$$

$100.\beta_2$ es la semielasticidad de Y respecto a X. Así:

$$\Delta\%Y = (100.\beta_2)\Delta X$$

Según este modelo, se estima que por cada incremento de un $1u.m$, en X se produce un incremento de un $(100.\beta_2)\%$ en Y.

a) Semielasticidad constante (level-log).

La relación entre Y y X Se controla por incrementos relativos de X:

$$Y_t = \beta_1 + \beta_2 log(X)_t + u_t$$

$$\Delta Y = \frac{\beta_2}{100}\Delta\%X$$

Según este modelo, se estima que por cada incremento de un 1 en X se produce un incremento de un $\frac{\beta_2}{100}u.m.$ en Y.

Segunda Parte: Inferencia en el Modelo de Regresión General

Distribuciones muestrales

La distribución muestral de un estimador describe el conjunto de estimaciones de un parámetro obtenidas en las distintas muestras que pueden extraerse de una población. La distribución muestral es útil para (1) establecer las propiedades estadísticas del estimador, (2) realizar contrastes de hipótesis y (3) calcular intervalos de confianza.

Distribución de los valores ajustados

Cada uno de los coeficientes estimados,$\hat{\beta}_i$, son una estimación insesgada del verdadero parámetro del modelo y representan la variación que experimenta la variable dependiente Y_i cuando una variable independiente X_i varía en una unidad y todas las demás permanecen constantes (supuesto ceteris paribus). Dichos coeficientes poseen propiedades estadísticas muy interesantes ya que, si se verifican los supuestos antes comentados, son insesgados, eficientes y óptimos.

La valores estimados de la variable dependiente en expresión matricial vienen dados por:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

Y la distribución estadística de la variable dependiente, se expresa en forma matricial:

$$\mathrm{E}(y) = X\beta$$

$$\mathrm{Var}(y) = \sigma^2 I$$

Distribución de los parametros

En el modelo de regresión la distribución muestral de los estimadores $\hat{\beta}_j$, vienen determindas por la forma funcional de los parámetros, así como por los supuestos básicos del modelo.

En el modelo de Regresión lineal simple los estimadores de mínimos cuadrados $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$ y $\hat{\beta}_2 = \frac{S_{XY}}{S_{XX}}$, son variables aleatorias linealmente distribuidas,

$$\hat{\beta}_1 \sim N\Big(\beta_1, \sigma^2\Big[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\Big]\Big)$$

$$\hat{\beta}_2 \sim N\Big(\beta_2, \frac{\sigma^2}{S_{XX}}\Big)$$

con covarianza muestral

$$cov(\beta_1, \beta_2) = \frac{-\sigma^2\bar{X}}{S_{XX}}$$

En la regresión lineal multiple, la distribución de probabilidad del estimador MCO $\hat{\beta}$ será una distribución Normal multivariante con vector de medias β y matriz de varianzas y covarianzas $\sigma^2(X'X)^{-1}$.

La esperanza matemática del estimador MCO se demuestra a partir de:

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}((X'X)^{-1}X'y)\\ &= (X'X)^{-1}X'\mathrm{E}(y)\\ &= (X'X)^{-1}X'X\beta = \beta\end{aligned}$$

De la definición de matriz de varianzas y covarianzas, se tiene que:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}((X'X)^{-1}X'y)\\ &= (X'X)^{-1}X'\mathrm{Var}(y)((X'X)^{-1}X')'\\ &= (X'X)^{-1}X'\mathrm{Var}(y)X(X'X)^{-1}\\ &= (X'X)^{-1}X'\sigma^2IX(X'X)^{-1}\\ &= (X'X)^{-1}\sigma^2\end{aligned}$$

El estimador $\hat{\beta}$ del parámetro β es insesgado porque su esperanza matemática coincide con el verdadero valor del parámetro $\mathrm{E}(\hat{\beta}) = \beta$.

Propiedades estadísticas del estimador MCO

El estimador $\hat{\beta}$ cumple tres propiedades estadísticas deseables:

1. Insesgadez: $E(\hat{\beta}) = \beta$.
2. Eficiencia: el teorema de Gauss-Markov afirma que, dentro de la clase de estimadores lineales e insesgados, el estimador minimocuadrático $\hat{\beta}$ es el que tiene menor varianza.
3. Consistencia: se dice que $\hat{\beta}_n$ es un estimador consistente de β si converge en probabilidad a β cuando $n \to \infty$, $\hat{\beta}_n \to \beta$.

Distribución de la varianza estimada del error

A diferencia de $\hat{\beta}_j$, que son estimadores lineales, $\hat{\sigma}^2$ es una función cuadrática de las observaciones $Y_1, ..., Y_n$. Su distribución muestral viene determinada por esta forma de relación no lineal, así como por los supuestos básicos del modelo de la regresión lineal.

En el modelo clásico, el estadístico $\frac{\hat{u}'\hat{u}}{\sigma_u^2} = \frac{(n-k)\hat{\sigma}^2}{\sigma_u^2}$ sigue una distribución muestral ji-cuadrado con $n-k$ grados de libertad,

$$\frac{(n-k)\hat{\sigma}^2}{\sigma_u^2} \rightarrow \chi^2_{n-k}$$

Entonces $\hat{\sigma}_u^2 = \frac{\hat{u}'\hat{u}}{n-k}$ es un estimador insesgado de σ_u^2 con varianza $\frac{2\sigma_u^4}{n-k}$

La esperanza matemática de una variable aleatoria z con distribución ji-cuadrado con m grados de libertad es igual a los grados de libertad m, $E(z) = m$. Por tanto,

$$E\left(\frac{\hat{u}'\hat{u}}{\sigma_u^2}\right) = (n-k)$$

De aquí $E(\hat{u}'\hat{u}) = (n-k)\sigma_u^2$ y

$$E(\hat{\sigma}_u^2) = E\left(\frac{\hat{u}'\hat{u}}{n-k}\right) = \sigma_u^2$$

La varianza de $z \sim \chi^2_m$ es igual a dos veces los grados de libertad, $var(z) = 2m$. Por tanto,

$$var\left(\frac{\hat{u}'\hat{u}}{\sigma_u^2}\right) = 2(n-k)$$

De aquí $var(\hat{u}'\hat{u}) = 2(n-k)\sigma_u^4$ y

$$var(\hat{\sigma}_u^2) = var\left(\frac{\hat{u}'\hat{u}}{(n-k)^2}\right) = \frac{2\sigma_u^4}{n-k}$$

La inferencia en el modelo de Regresión lineal General.

Introducción

Hasta el momento hemos visto como la estimación por MCO permite obtener estimaciones puntuales de los parámetros del modelo. La inferencia acerca de los mismos permite completar dicha estimación puntual, mediante la estimación por intervalos y el contraste de hipótesis de hipotesis sobre los parametros o coeficientes.

Un contraste de hipótesis es un procedimiento para decidir si los datos de una muestra apoyan o contradicen una determinada hipótesis sobre la población de la que se han extraído.

Una hipótesis estadística es una conjetura sobre un parámetro poblacional, o sobre la distribución de la población, que puede ser verdadera o falsa.

Una hipótesis es una afirmación que está sujeta a verificación o comprobación. Hay que tener presente que una hipótesis no es un hecho establecido o firme, las hipótesis están basadas en la experiencia, en la observación, en la experimentación o en la intuición del sujeto que las formula.

Cuando las hipótesis se plantean de tal modo que se pueden comprobar por medio de métodos estadísticos reciben el nombre de hipótesis estadísticas. Estas hipótesis son afirmaciones que se efectúan sobre uno o más

parámetros de una o más poblaciones. Las hipótesis estadísticas son de dos tipos: hipótesis nula e hipótesis alternativa. La hipótesis nula, o que no se verifique dicha afirmación, simbolizada por H_0, es la hipótesis que se debe comprobar.

Para contrastar una hipótesis nula examinamos los datos de la muestra tomados de la población y determinamos si son o no compatibles con dicha hipótesis. Si son compatibles entonces H_0 se acepta, en caso contrario se rechaza. Si se acepta la hipótesis nula afirmamos que los datos de esa muestra en concreto no dan suficiente evidencia para que concluyamos que la hipótesis nula sea falsa; si se rechaza decimos que los datos particulares de la muestra ponen de manifiesto que la hipótesis nula es falsa, entonces la hipótesis alternativa,H_1, es considerada verdadera.

El criterio que permite decidir si rechazamos o no la hipótesis nula es siempre el mismo. Definimos un estadístico de prueba, y unos límites que dividen el espacio muestral en una región en donde se rechaza la hipótesis establecida, y otra región en la que no se rechaza, llamada región de aceptación. A la región donde se rechaza la hipótesis nula se le llama región crítica. Esta región es un subconjunto del espacio muestral, y si el valor del estadístico de prueba pertenece a él se rechaza la hipótesis nula.

El límite entre la región crítica y la región de aceptación viene determinado por la información previa relativa a la distribución del estadístico de prueba.

Señalar que un estadístico de prueba es una fórmula que nos dice como confrontar la hipótesis nula con la información de la muestra y es, por tanto, una variable aleatoria cuyo valor cambia de muestra a muestra.

Otra de las consideraciones a realizar en el contraste de hipótesis es fijar la probabilidad del error de rechazar la prueba siendo cierta. A este error se le denomina **nivel de significación**. Por ejemplo, si se utiliza un nivel de significación de $\alpha = 0,05$, equivale a decir que si para realizar un contraste tomáramos infinitas muestras de la población, rechazaríamos la hipótesis nula de forma incorrecta un 5 por ciento de las veces.

En un contraste de hipótesis pueden darse cuatro resultados que se derivan de cruzar las dos posibles decisiones que podemos tomar sobre H_0 (aceptar o rechazar H_0) y las dos posibles situaciones en que se encuentra H_0 (verdadera o falsa). Estos cuatro resultados se muestran en el cuadro siguinete:

Decisión/Situación	H_0 es verdadera	H_0 es falsa
Aceptar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

Tomamos una decisión correcta al aceptar H_0 cuando es verdadera o al rechazar H_0 cuando es falsa. En cambio, cometemos un error al rechazar H_0 cuando es verdadera, que se denomina error de tipo I, o al aceptar H_0 cuando es falsa, denominado error de tipo II.

El contraste de hipótesis se apoya en distribuciones estadisticas:

Distribución normal

Se denomina *distribución normal estándar* a la distribución normal que tiene media 0 y varianza 1. Escrito sucintamente como $Z \sim N(0, 1)$.

En R la distribución normal se genera con:

Densidad: dnorm(x, mean = 0, sd = 1, log = FALSE)

Función de distribución: pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

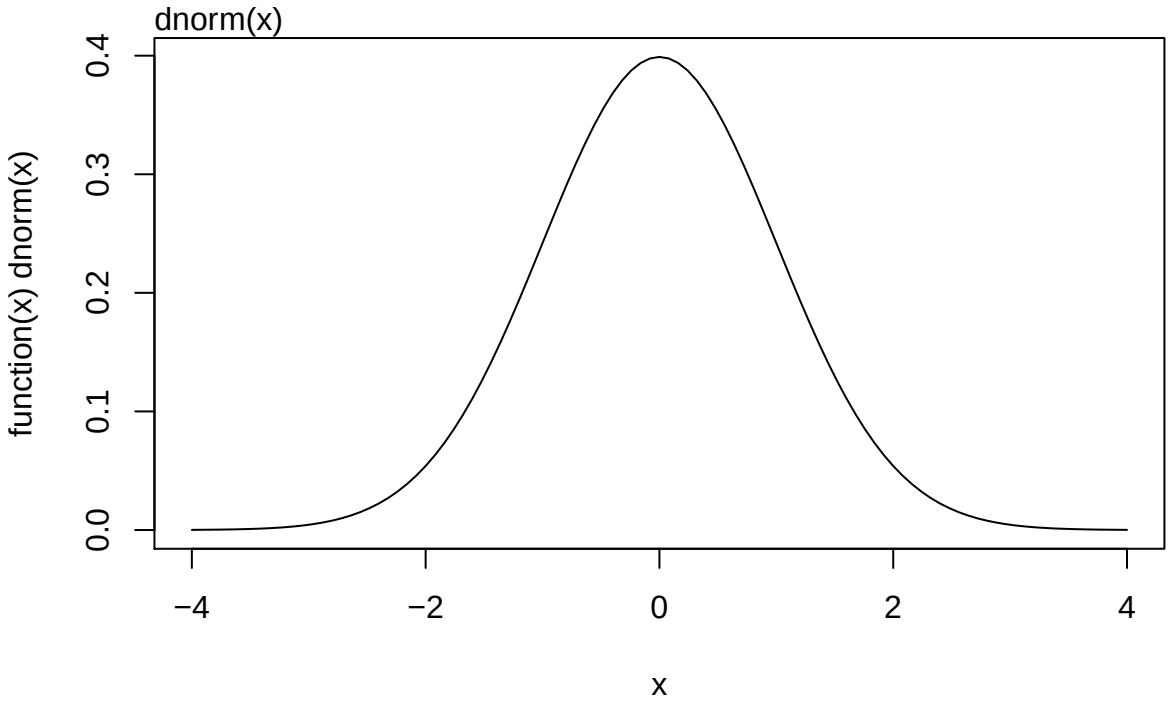
Función inversa; qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

Generación de numeros aleatorios: rnorm(n, mean = 0, sd = 1)

Función de densidad y de densidad acumulada:

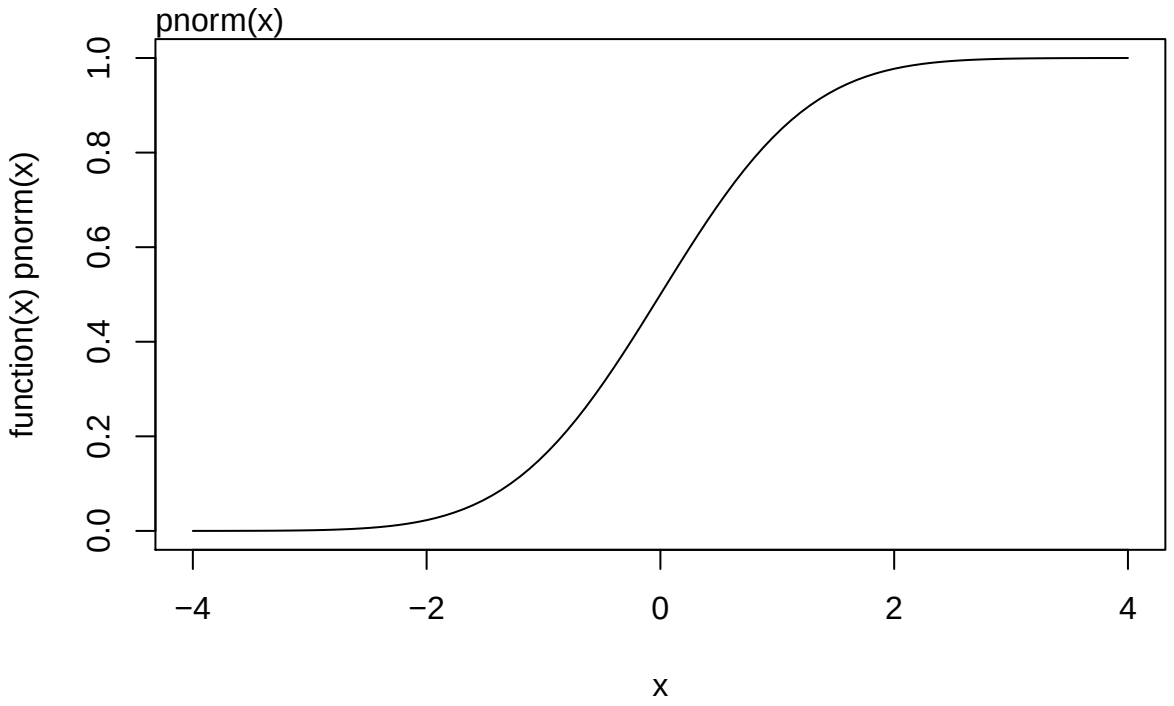
```r
require(graphics)

## Grafico funcion densidad
plot(function(x) dnorm(x), -4, 4,
     main = "Figura 5 Función de densidad de una normal estándar")
mtext("dnorm(x)", adj = 0)
```

Figura 5 Función de densidad de una normal estándar




```r
## Gráfico función de densidad acumulada

plot(function(x) pnorm(x), -4, 4,
     main = "Figura 6 Función de densidad acumulada de una normal estándar")
mtext("pnorm(x)", adj = 0)
```

Figura 6 Función de densidad acumulada de una normal estándar



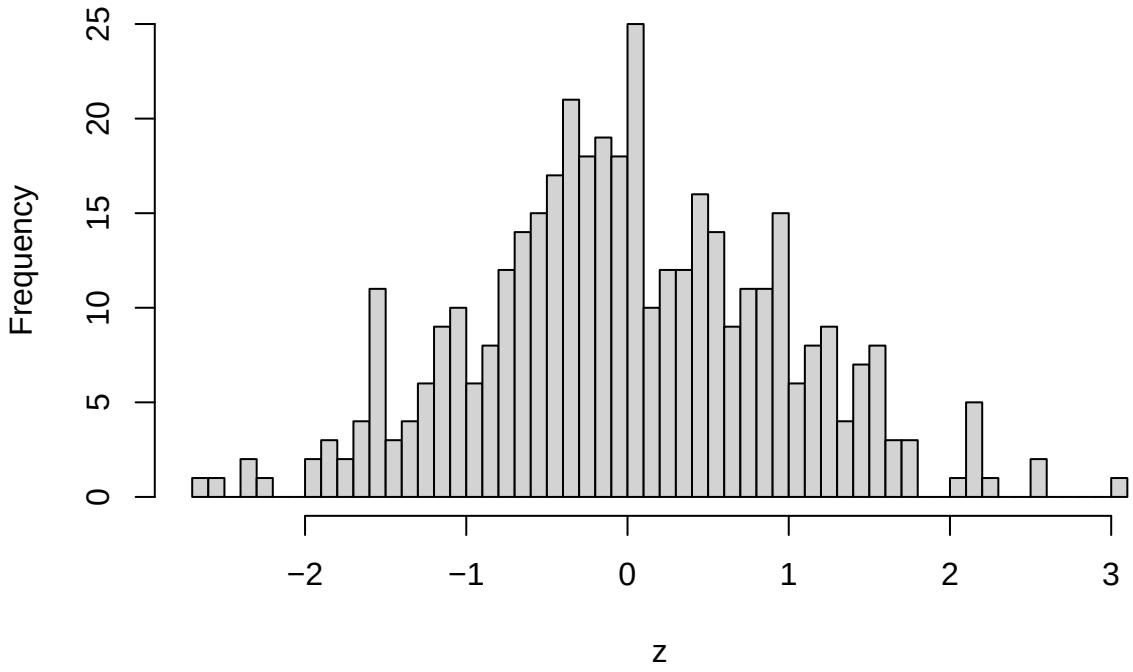
Generamos 400 números aleatorios para una normal estándar:

```
z= rnorm(400, mean = 0, sd = 1)
summary(z)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.647194 -0.593366 -0.045114  0.004971  0.629481  3.088361
```

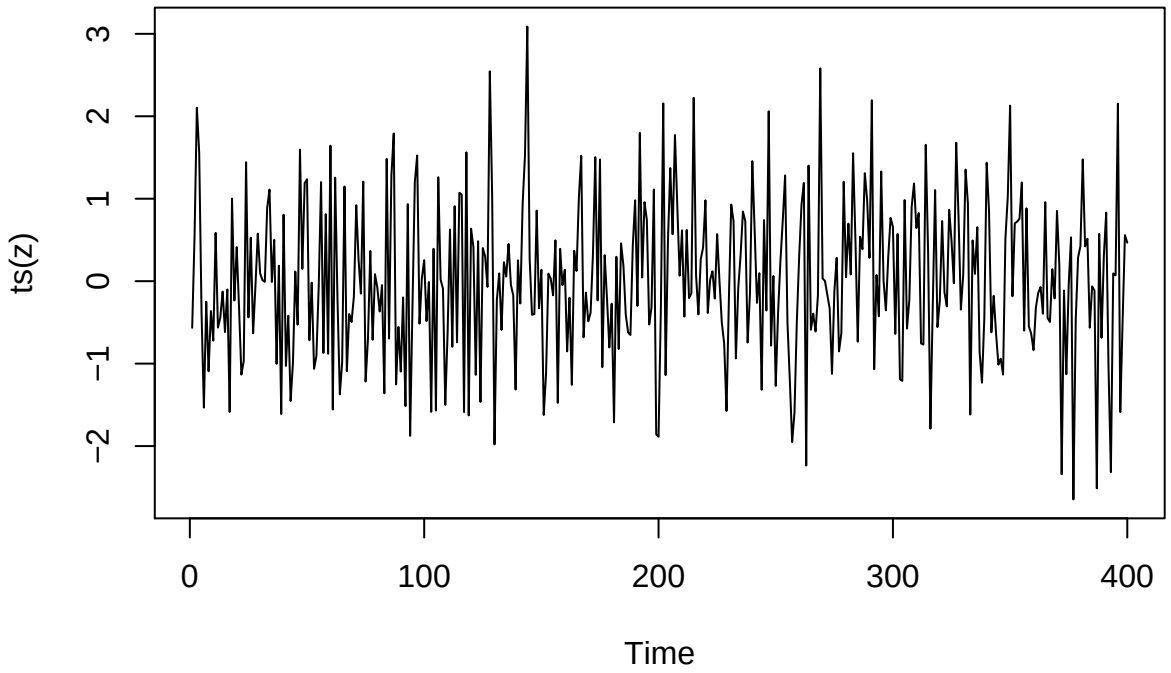
```
hist(z,nclass=50,main="Figura 7 Histograma normal estándar")
```

Figura 7 Histograma normal estándar




```
plot(ts(z), main="Figura 8 Secuencia temporal de una normal estándar")
```

Figura 8 Secuencia temporal de una normal estándar



Una distrución normal estándar, Z , puede transformarse en un distribución normal, X , de media μ y varianza σ^2, denotándose como $X \sim N(\mu, \sigma^2$, aplicando la siguiente transformación:

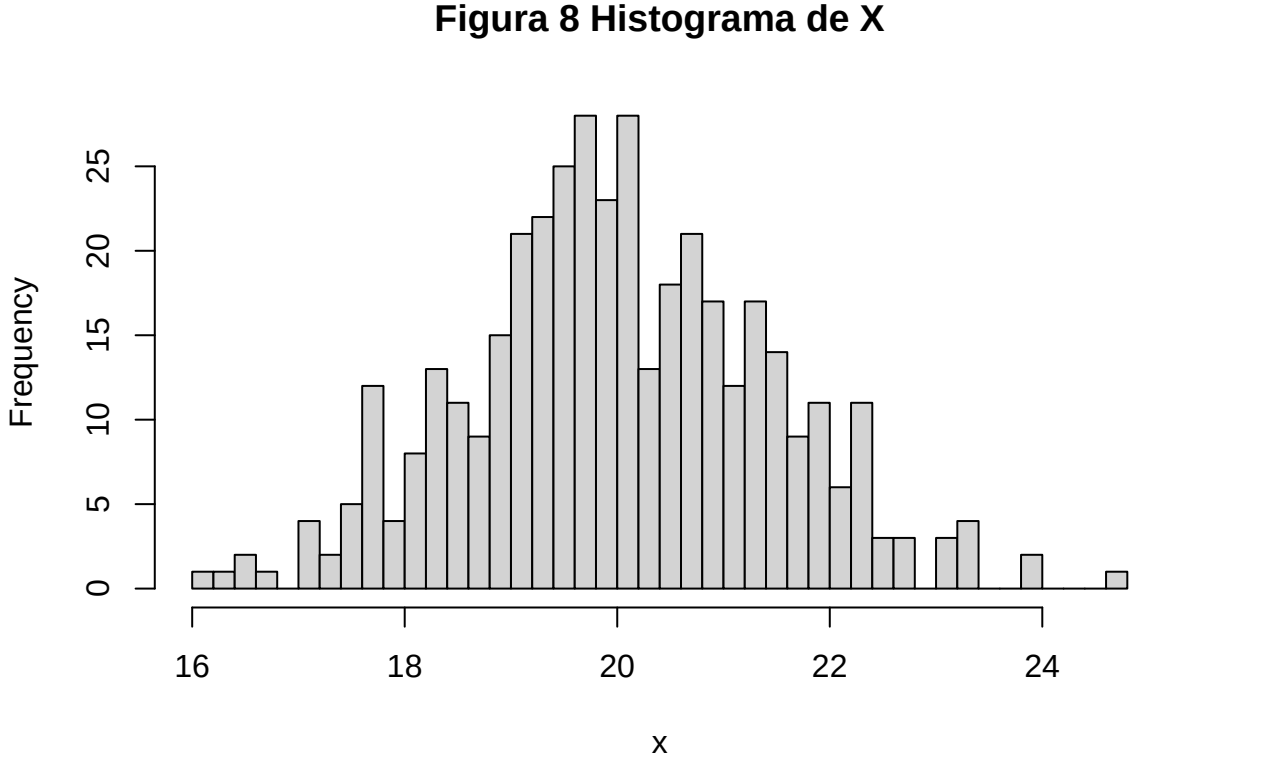
$$X = \mu + \sigma Z$$

Ejemplo para $\mu = 20$ y $\sigma = 1.5$

```
x=20+1.5*z
summary(x)
```

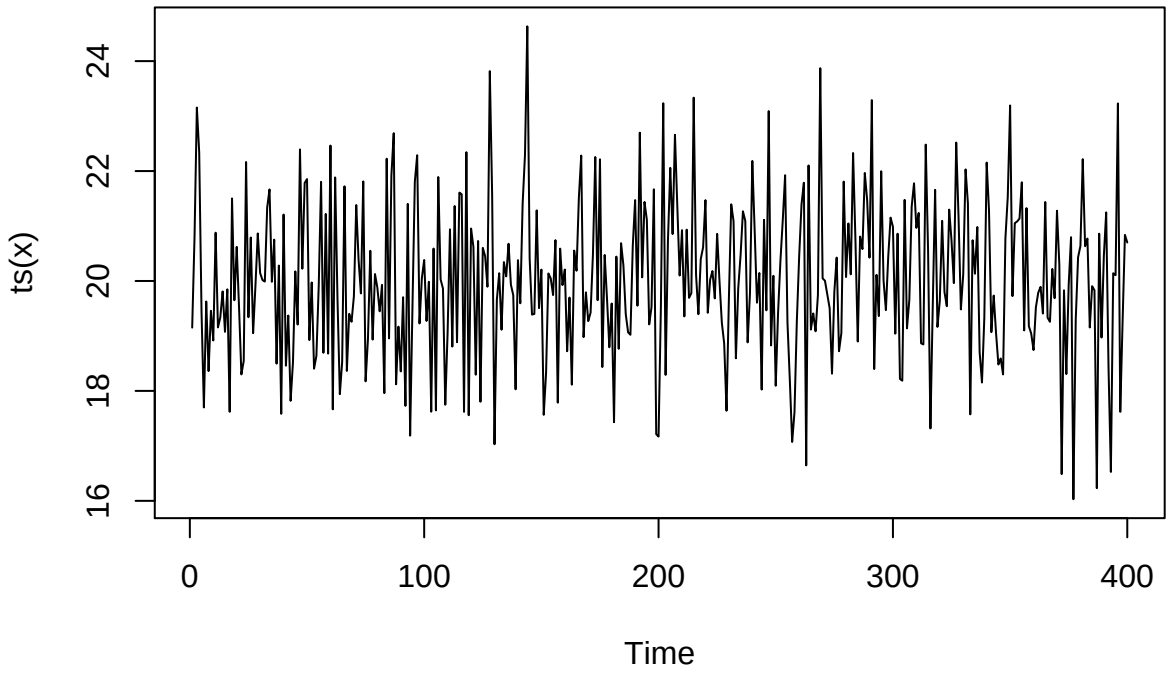
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.03   19.11   19.93   20.01   20.94   24.63
```

```
hist(x,nclass=50,main="Figura 8 Histograma de X")
```




```
plot(ts(x), main="Figura 9 Secuencia temporal de X")
```

Figura 9 Secuencia temporal de X



Las distribuciones asociadas a la distribución normal que se van a utilizar en este curso son:

Distribución t-student

Sea Z una variable aleatoria con distribución normal estándar y sea Y una variable aleatoria con distribución χ^2 con k grados de libertad, siendo z e Y independientes. Entonces la variable aleatoria $T = \frac{z}{\sqrt{\frac{Y}{k}}}$, tiene una distribución t de Student con k grados de libertad (Gosset 1908).

$$T = \frac{Z}{\sqrt{\frac{Y}{k}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2}{k}}} \simeq t_k$$

En R la distribución t de student se genera con:

dt(x, df, ncp, log = FALSE)

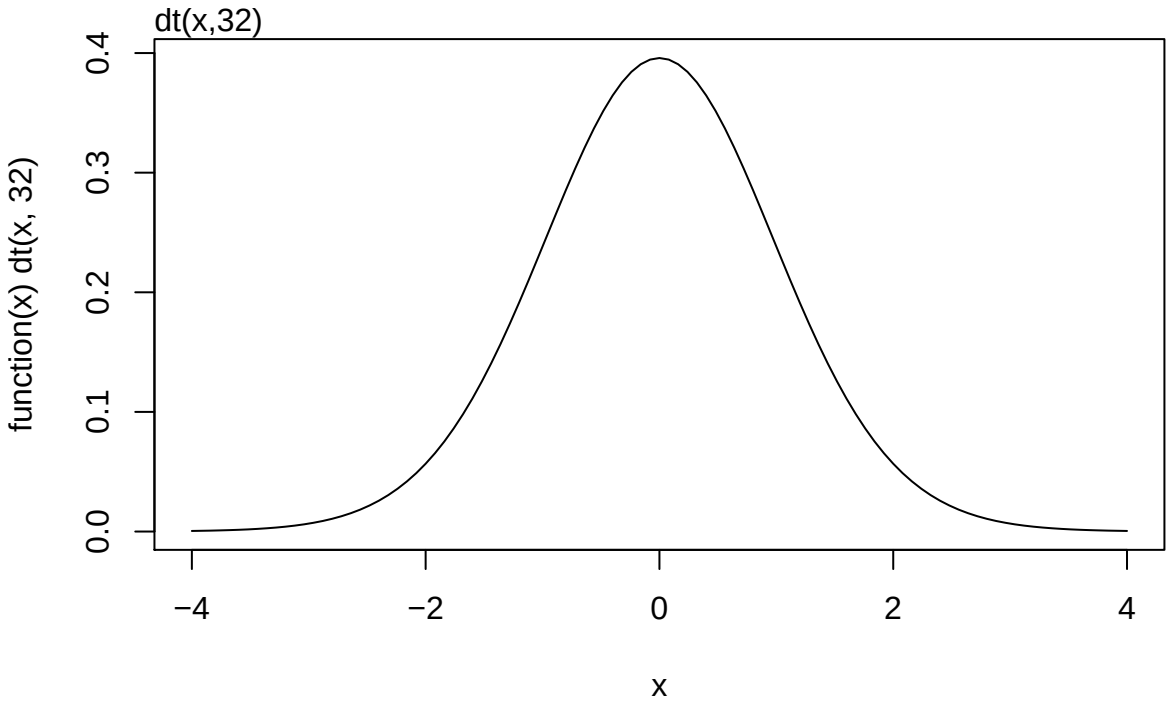
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)

qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)

rt(n, df, ncp)

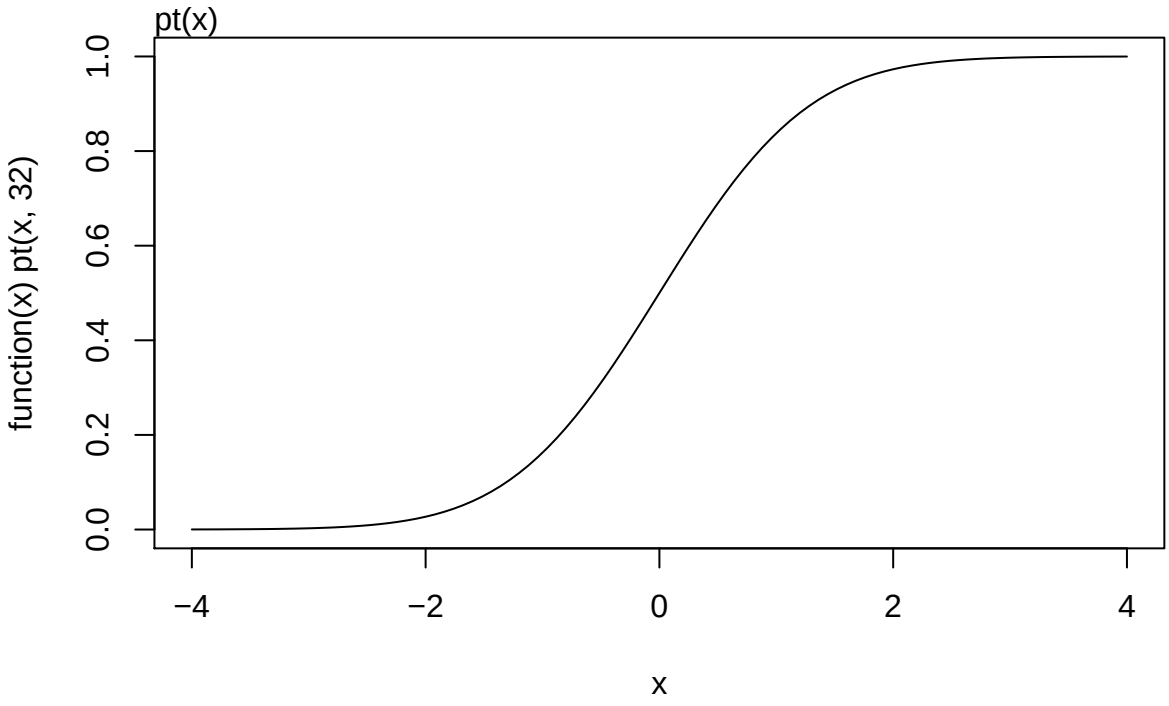
```
# Grafico funcion densidad
plot(function(x) dt(x,32), 4, -4,
     main = "Figura 11 Función de densidad de una t de student con k=32")
mtext("dt(x,32)", adj = 0)
```

Figura 11 Función de densidad de una t de student con k=32




```
# Gráfico función de densidad acumulada

plot(function(x) pt(x,32), -4, 4,
     main = "Figura 12 Función de densidad acumulada de una t de student con k=32")
mtext("pt(x)", adj = 0)
```

Figura 12 Función de densidad acumulada de una t de student con k=



Distribución Ji-Cuadrado.

Definición. Si $Z_1, Z_2, ..., Z_k$ son k variables aleatorias independientes distribuidas bajo una normal estándar, $Z \sim N(0,1)$, entonces la suma de sus cuadrados sigue una distribución ji-cuadrado con k grados de libertad.

$$\sum_{i=1}^{k} Z_i \sim \chi_k^2$$

En R la distribución ji-cuadrado se genera con:

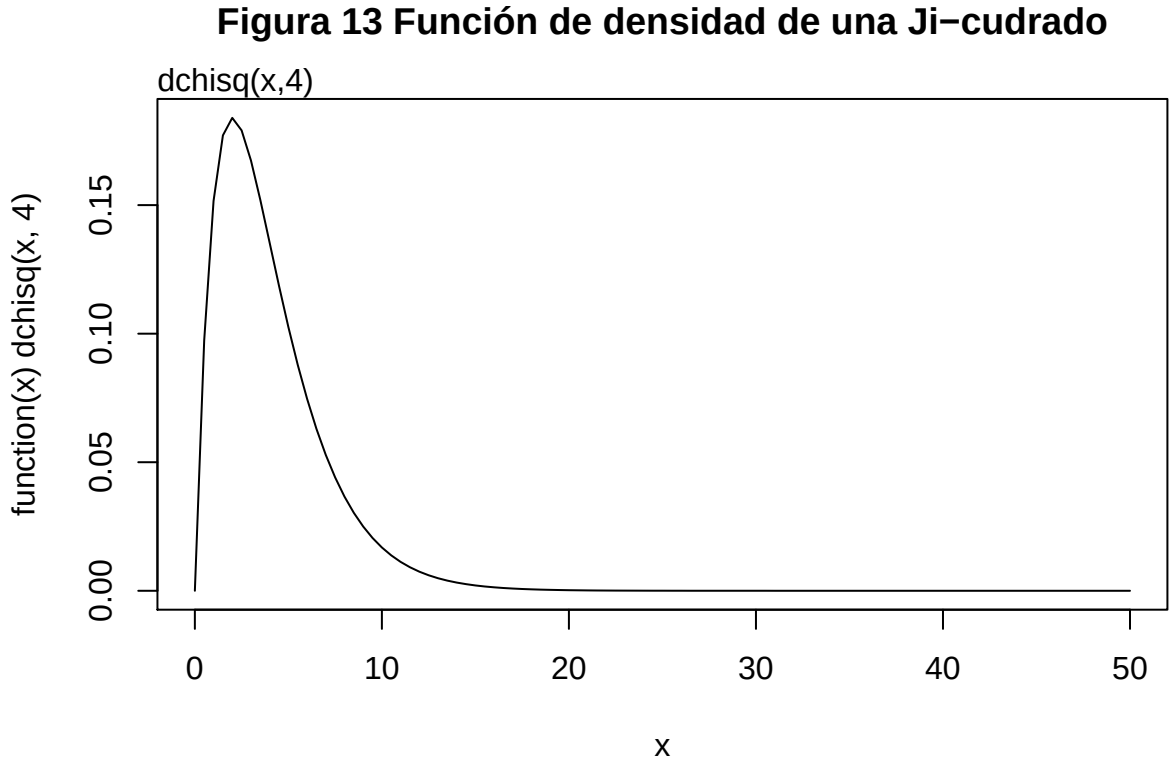
dchisq(x, df, ncp = 0, log = FALSE)

pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)

qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)

rchisq(n, df, ncp = 0)

```
## Grafico funcion densidad
plot(function(x) dchisq(x,4), 0, 50,
     main = "Figura 13 Función de densidad de una Ji-cudrado")
mtext("dchisq(x,4)", adj = 0)
```



Distribución F de Fisher.

Sean U y V dos variables aleatorias independientes siguiendo distribuciones ji-cuadrado con m y n grados de libertad, respectivamente, $U \sim \chi_m^2$ y $V \sim \chi_n^2$. Entonces la ratio de las dos variables corregidas por sus respectivos grados de libertad sigue una distribución F con m grados de libertad en el numerador y n en el denominador.

$$T = \frac{\frac{U}{m}}{\frac{V}{n}} \sim F$$

df(x, df1, df2, ncp, log = FALSE)

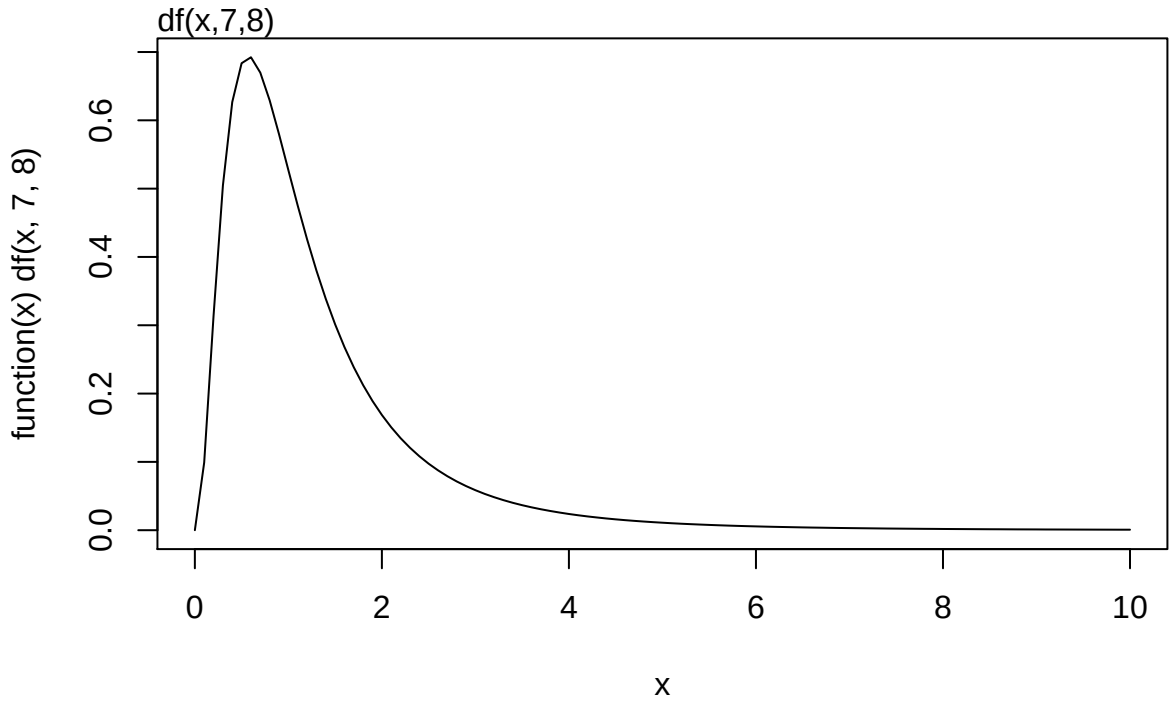
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)

qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)

rf(n, df1, df2, ncp)

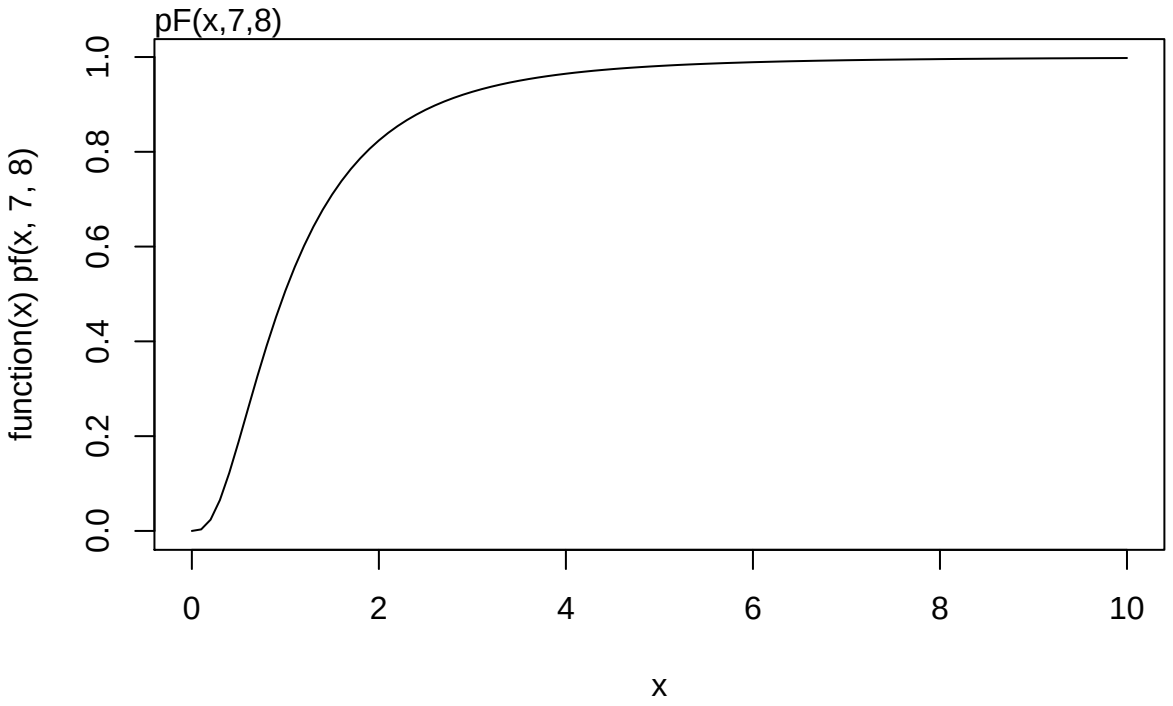
```
require(graphics)

## Grafico funcion densidad
plot(function(x) df(x,7,8), 0, 10,
     main = "Figura 14 Función de densidad de una distribución F")
mtext("df(x,7,8)", adj = 0)
```

Figura 14 Función de densidad de una distribución F




```
## Gráfico función de densidad acumulada

plot(function(x) pf(x,7,8), 0, 10,
     main = "Figura 15 Función de densidad acumulada de una distribución F")
mtext("pF(x,7,8)", adj = 0)
```

Figura 15 Función de densidad acumulada de una distribución F



Fases del contraste de Hipótesis

En la formalización del procedimiento de contrastación podemos distinguir cinco pasos principales:

1. Formular las hipótesis H_0 y H_1 a contrastar;
2. Proponer un estadístico de contraste;
3. Elegir un nivel de significación aceptable;
4. Determinar la región crítica y/o calcular el p-valor;
5. Tomar la decisión de aceptar o rechazar la hipótesis nula.

1. Formular las hipótesis

Las hipótesis simples seran:

a) Bilaterales:

El contraste bilateral sitúa la región de rechazo en los dos extremos (colas) de la distribución muestral. En cambio, el contraste unilateral sitúa la región de rechazo en uno de los dos extremos (colas) de la distribución muestral. El contraste bilateral (o de dos colas) se utiliza cuando la Hipótesis Alternativa asigna al parámetro cualquier valor diferente al establecido en la Hipótesis Nula.

En el contraste bilateral la Hipótesis Nula es

$$H_0 : \beta_i = \beta_i^0$$

siendo la Hipótesis Alternativa,

$$H_0 : \beta_i \neq \beta_i^0$$

La Hipótesis Alternativa establece que, caso de rechazar la Hipótesis Nula, decidimos que la proporción de la población a que pertenece la muestra no es β_i^0.

b) Unilaterales

En el contraste unilateral la Hipótesis Nula se formula:

$$H_0 : \beta_i = \beta_i^0$$

en donde si la Hipótesis Alternativa se expresa como:

$$H_0 : \beta_i > \beta_i^0$$

establece que, caso de rechazar la Hipótesis Nula, decidimos que la proporción de la población a que pertenece la muestra es superior a β_i^0.

2. Calcular el estadístico de contraste

Los datos son una representación de las observaciones, y contienen la evidencia según la cual se modifica o no las hipótesis iniciales. Habitualmente son transformados en estadísticos que cuantifican las características de interés de las muestras.

Una manera de obtener el grado de coherencia de la evidencia observada con la Hipótesis Nula consiste en obtener la probabilidad de ocurrencia de lo observado si la Hipótesis Nula fuera verdadera. El procedimiento consiste en obtener probabilidades de ocurrencia de los estadísticos en las distribuciones muestrales definidas en las hipótesis nulas.

Puesto que las probabilidades de ocurrencia tienen que obtenerse de una distribución estadística teórica, los estadisticos de contraste tienen que estar vinculados a una distribución estadística concreta.

Este sería un ejemplo de un estadístico teórico vinculado a la distribución t de student.

$$t_i = \frac{\hat{\beta}_i - \beta_i^0}{S_{\hat{\beta}_i}}$$

3. Elegir el nivel de significación

Cuando se toma la decisión de rechazar o no la Hipótesis Nula podemos acertar o cometer errores. En el trabajo real no sabemos qué ocurre porque no sabemos si la Hipótesis Nula es verdadera o no. Sin embargo, dados ciertos supuestos podemos obtener las probabilidades de cometer errores de tipo I y de tipo II.

La probabilidad de cometer errores de tipo I, que se simboliza α, es la probabilidad de ocurrencia de los valores del estadístico en la región de rechazo cuando la Hipótesis Nula es verdadera. El valor de α, también denominado nivel de significación, es definido por el investigador antes de recoger los datos, y la costumbre es hacer $\alpha = 0.05$ o $\alpha = 0.01$.

La probabilidad de cometer errores de tipo II se simboliza β y depende de varias circunstancias como la distancia que separa el valor asignado al parámetro en la Hipótesis Nula de su valor real, el tamaño muestral y el valor asignado a alfa.

4. Determinar la región crítica y/o calcular el p-valor

La potencia del contraste es la probabilidad de acertar cuando la Hipótesis Nula es falsa. Esta probabilidad es la de que el conjunto de valores muestrales del estadístico de contraste se sitúe en la región de rechazo bajo el supuesto de que la Hipótesis Alternativa sea verdadera.

Par ello hay que obtener en las tablas de la distribución de referencia:

a) Un valor crítico que diferencie las areas de la distribución en las que se va a aceptar o rechazar la Hipótesis nula.

Un ejemplo de valores críticos en una *distribución t-student* para un contraste bilateral (2 colas), sería el valor de $t_{n-k,0.975}$ ó $t_{n-k,0.025}$. Es decir el valor que delimita el 2.5 % de los valores de la distribución t-student más grandes, o el valor que delimita el 2.5 % de los valores de la distribución t-student más pequeños.

b) el p-valor asociado al valor de la distribución teórica correspondiente al estadistico de contraste.

El p-valor se define como la probabilidad de obtener un estadístico de contraste al menos tan extremo como el que evalúo.

c) El intervalo de confianza dentro del cual, con un determinado nivel de significación, oscilará el verdadero valor de un parámetro o de un pronóstico.

El intervalo de confianza para un parámetro que se construye en base a la probabilidad de que su valor desconocido esté comprendida entre dos valores β_a y β_b equidistantes a ese parámetro:

$$\mathrm{P}(\beta_a \leq \beta_i \leq \beta_b) = 1 - \alpha$$

siendo $1 - \alpha$ el nivel o grado de confianza asociado a dicho intervalo.

En términos generales, los intervalos de confianza para los estadísticos muestrales se expresan como:

```
Estimador ± (Factor de Fiabilidad)*(Error Típico del Estimador)
```

5. Tomar la decisión de aceptar o rechazar la hipótesis nula

Se trata de establecer la regla de decisión para aceptar la hipotesis nula H_0. En el ejemplo de un contraste simple bilateral si $\mid t_i \mid < t_{n-k,1-\frac{\alpha}{2}}$ acepto H_0 (dado que por la simetría de la función, $-t_{n-k,\frac{\alpha}{2}} = t_{n-k,1-\frac{\alpha}{2}}$).

En el caso de que obtenga el p-valor del estadístico de contraste, este debe ser inferior al nivel de significación α elegido.

p-valor (t_i) < p-valor $(|t_{\frac{\alpha}{2},n-2}|)$

En el caso de que utilice el intervalo de confianza el criterio de decisión que rechaza la hipótesis nula es que el valor β_i^0 no se encuentre dentro del intervalo de confianza.

Constraste de hipótesis en el modelo lineal general.

Estadístico de referencia

La distribución tipificada de $\hat{\beta}_j$ sigue una distribución normal estándar,

$$s_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}} \approx N(1,0)$$

donde a_{jj} es el elemento de la fila j y la columna j de la matriz $(X'X)^{-1}$.

Esta distribución es útil para contrastar hipótesis sobre el parámetro verdadero β_j , pero depende de un parámetro desconocido σ^2. Si reemplazamos σ^2 por su estimación de $\hat{\sigma}^2$, la distribución del estadístico tipificado se altera en muestras finitas.

En el modelo de RLG, la ratio $t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$ sigue una distribución t-student con $n - k$ grados de libertad.

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}} \approx t_{n-k}$$

Contraste de significación individual

Se denomina constraste de significación individual al contraste de hipótesis de que el parámetro $\beta_i = 0$.

El contraste de hipótesis más cómun en el análisis de regresión es el contraste de significación, en el que se formulan las hipótesis:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

En el modelo de RLG, la hipótesis nula $H_0 : \beta_j = 0$ se rechaza frente a la hipótesis alternativa $H_1 : \beta_j \neq 0$ al nivel de significación α cuando

$$\left| \frac{\hat{\beta}_i}{\sqrt{\hat{V}(\hat{\beta}_j)}} \right| > |t_{n-k}|$$

donde $\sqrt{\hat{V}(\hat{\beta}_j)}$ es la desviación típica estimada de $\hat{\beta}_j$ y t_{n-k} es el valor crítico tal que $Pr(t_{n-k} > t_{\frac{\alpha}{2},n-k}) = \frac{\alpha}{2}$.

Se rechaza H_0 cuando $|t_j|$ toma valores grandes, es decir, cuando $|t_j|$ es mayor que un determinado valor crítico c. Ahora bien, bajo este criterio, la probabilidad del error de tipo I (probabilidad de rechazar H_0 cuando es verdadera) será:

$$Pr(|t_{n-k}|) > c$$

Por tanto, si fijamos un nivel de significación α, el valor crítico $c = t_{\frac{\alpha}{2},n-k}$

El contraste de significación también puede basarse en el p-valor y en el intervalo de confianza.

El p-valor de la ratio t_j, definido como:

$Pr(|t_{n-k}| > |t_j|)$ o $2Pr(t_{n-k} > |t_j|)$

rechaza H_0 frente a H_1 al nivel de significación α cuando:

p-valor de $t_j < \alpha$

El intervalo de confianza del $100(1-\alpha)\%$ para el coeficiente β_j, definido como

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-k}\sqrt{\hat{V}(\hat{\beta}_j)}$$

rechaza $H_0 : \beta_j = 0$ frente a $H_1 : \beta_j \neq 0$ al nivel de significación α cuando los dos límites del intervalo tienen el mismo signo, es decir, cuando el intervalo no contiene el valor cero.

A veces, en lugar de formular una hipótesis alternativa bilateral $H_1 : \beta_j \neq 0$, formulamos una alternativa unilateral, que puede ser de lado izquierdo $H_1 : \beta_j < 0$ o de lado derecho $H_1 : \beta_j > 0$. En este caso, asignamos todo el nivel de significación α a la cola izquierda o a la cola derecha de la distribución t_{n-k}, y usamos $-t_{\alpha,n-k}$ o $t_{\alpha,n-k}$ como valores críticos.

En el modelo de RLS, el contraste de la hipótesis nula $H_0 : \beta_j = \beta_j^0$ frente a una determinada alternativa uni o bilateral puede realizarse con la ratio-t, el p-valor o el intervalo de confianza:

Alternativa	Región crítica	p-valor	Intervalo de confianza
$H_1 : \beta_j \neq \beta_j^0$	$\|t_j\| > t_{\frac{\alpha}{2},n-k}$	$Pr(\|t_{n-k}\| > \|t_j\|)$	$\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-k}\sqrt{\hat{V}(\hat{\beta}_j)}$
$H_1 : \beta_j < \beta_j^0$	$t_j < -t_{\alpha,n-k}$	$Pr(t_{n-k} < t_j)$	$\hat{\beta}_j - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{V}(\hat{\beta}_j)}; \infty$

Alternativa	Región crítica	p-valor	Intervalo de confianza
$H_1 : \beta_j > \beta_j^0$	$t_j > t_{\alpha,n-k}$	$Pr(t_{n-k} > t_j)$	$-\infty; \hat{\beta}_j + t_{\frac{\alpha}{2},n-k}\sqrt{\hat{V}(\hat{\beta}_j)}$

Hipótesis sobre la varianza del error

Las hipótesis que formulamos sobre σ^2 son:

$$H_0 : \sigma^2 = \sigma_0^2$$

$H_1 : \sigma^2 \neq \sigma_0^2$ ($H_1 : \sigma^2 < \sigma_0^2$ o $H_1 : \sigma^2 > \sigma_0^2$).

En el modelo de RLG, la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$ se rechaza frente a la hipótesis alternativa $H_1 : \sigma^2 \neq \sigma_0^2$ al nivel de significación α cuando:

$\frac{SCR}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2},n-k}$ o $\frac{SCR}{\sigma_0^2} > \chi^2_{\frac{\alpha}{2},n-k}$

donde $\chi^2_{p,n-k}$ es el valor crítico tal que $Pr(\chi^2_{n-k} < \chi^2_{p,n-k}) = p$ con $p = 1 - \frac{\alpha}{2}$ o $p = \frac{\alpha}{2}$.

El p-valor del estadístico χ^2 se define como:

$$P - valor = 2min\left[Pr\left(\frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2},n-k}\right); Pr\left(\frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} > \chi^2_{\frac{\alpha}{2},n-k}\right)\right]$$

y rechaza H_0 frente a H_1 al nivel de significación α cuando

$p - valor$ de $\sigma^2 < \alpha$

El intervalo de confianza del $100(1-\alpha)\%$ para σ^2, definido como

$$\left(\frac{SCR}{\chi^2_{\frac{\alpha}{2},n-k}}; \frac{SCR}{\chi^2_{1-\frac{\alpha}{2},n-k}}\right)$$

rechaza $H_0 : \sigma^2 = \sigma_0^2$ frente a $H_1 : \sigma^2 \neq \sigma_0^2$ al nivel de significación α cuando no contiene al parámetro σ_0^2.

El contraste F

Se desea contrastar la hipótesis nula de que un subvector de s coeficientes β_s es igual a β_s^0 frente a la hipótesis alternativa de que β_s es distinto de β_s^0.

$H_0 : \beta_s = \beta_s^0$ versus $H_0 : \beta_s \neq \beta_s^0$

La hipótesis H_0 se rechaza al nivel de significación α si:

$$F \approx \frac{(\hat{\beta}_s - \beta_s^0)Var(\hat{\beta}_s)^{-1}(\hat{\beta}_s - \beta_s^0)}{s} > c$$

donde c es el valor crítico para el cual $Pr(F_{s,n-k} > c) = \alpha$.

Contraste de significación conjunta

Un caso especial del contraste F es

$H_0 : \beta_s = 0_s$ versus $H_0 : \beta_s \neq 0_s$

donde la hipótesis nula $H0 : \beta_s = 0_s$ conlleva la eliminación de s variables explicativas de la ecuación de regresión.

La hipótesis de no significación conjunta $H_0 : \beta_s = 0_s$ se rechaza al nivel de significación α si:

$$F \approx \frac{\hat{\beta}_s' Var(\hat{\beta}_s)^{-1} \hat{\beta}_s}{s} > c$$

donde c es el valor crítico para el cual $Pr(F_{s,n-k} > c) = \alpha$.

Hay un procedimiento más conveniente de realizar el contraste de significación conjunta basado en sumas de cuadrados de residuos.

La hipótesis de no significación conjunta $H_0 : \beta_s = 0_s$ se rechaza al nivel de significación α si:

$$F \approx \frac{\frac{(\hat{u}_r' \hat{u}_r - \hat{u}'\hat{u})}{s}}{\frac{\hat{u}'\hat{u}}{n-k}} > c$$

donde c es el valor crítico para el cual $Pr(F_{s,n-k} > c) = \alpha$. $\hat{u}'\hat{u}$ es la suma de cuadrados de los residuos en la regresión de y sobre X y $\hat{u}_r'\hat{u}_r$ es la suma de cuadrados de los residuos en la regresión de y sobre X_r.

Para realizar el contraste de significación conjunta seguimos los siguientes pasos:

1. estimamos el modelo de regresión con todas las variables de interés,

$$y = \beta X + u = \beta_s X_s + \beta_r X_r + u$$

que nos proporciona la suma de los residuos $\hat{u}'\hat{u}$.

2. estimamos el modelo de regresión bajo $H_0 : \beta_s = 0_s$

$$y = \beta_r X_r + u_r$$

que nos proporciona la suma de los residuos $\hat{u}_r'\hat{u}_r$.

3. Calculamos el estadístico de contraste

$$F = \frac{\frac{(\hat{u}_r' \hat{u}_r - \hat{u}'\hat{u})}{s}}{\frac{\hat{u}'\hat{u}}{n-k}}$$

4. y, finalmente, comparamos F con el valor crítico c de la distribución $F_{s,n-k}$ alnivel de significación α. Si $F < c$, aceptamos H_0; en caso contrario, rechazamos H_0.

Contraste de significación global (Tabla ANOVA)

El caso de significación conjunta más relevante es

$H_0 : \beta_s = 0$ versus $H_0 : \beta_s \neq 0$

en donde el subvector $\beta_s = (\beta_1, \beta_2, ..., \beta_k)'$ incluye todas los coeficientes (pendientes) del modelo salvo el término constante: $s = k - 1$.

La hipótesis de no significación global $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ se rechaza al nivel de significación α construyendo el estadístico experimental:

$$F = \frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$$

y la regla de decisión que rechaza la hipótesis H_0 ocurre cuando $F_{exp} > F_{k-1,n-k,\alpha}$.

El contraste de significación global se resume en el cuadro siguiente, en donde la variación total de la variable dependiente (SCT) se descompone en la explicada por la regresión (SCE) y en la no explicada (SCR). Los grados de libertad de estas tres sumas de cuadrados son $n-1$, $k-1$ y $n-k$, respectivamente.

A partir de esta información muestral, podemos calcular el numerador y denominador del estadístico F mediante la Tabla ANOVA:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadistico F
Regresión	SCE	$k-1$	$\frac{SCE}{k-1}$	$\frac{\frac{SCE}{k}}{\frac{SCR}{n-k}}$
Residual	SCR	$n-k$	$\frac{SCR}{n-k}$	
Total	SCT	$n-1$		

El estadístico esperimental que rechaza la hipótesis $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$, tambien se calcula:

$$F = \frac{\frac{R^2}{k-1}}{\frac{(1-R^2)}{n-k}}$$

Contraste de hipótesis: la hipótesis lineal general

La hipótesis lineal general especifica un conjunto de relaciones lineales, también llamadas restricciones, entre los parámetros del modelo de regresión lineal.

El conjunto de restricciones lineales

$$R\beta = r$$

se denomina hipótesis lineal general, donde R es una matriz conocida de orden $q \times k$ y rango $q \leq k$, y r es un vector conocido de orden $q \times 1$.

Veamos algunos ejemplos de hipótesis lineales que se pueden formular como $R\beta = r$.

Usaremos como ilustración el modelo de regresión múltiple

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

.

1. Significación individual de β_2, $H_0 : \beta_2 = 0$.

En notación matricial la forma de la hipótesis es:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = 0$$

siendo $R = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}'$ y $r = 0$.

2. Igualdad de pendientes, $H_0 : \beta_2 = \beta_3$.

En notación matricial la forma de la hipótesis es:

$$\begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = 0$$

siendo $R = \begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}'$ y $r = 0$.

3. Significación conjunta de dos pendientes, $H_0 : \beta_2 = \beta_3 = 0$.

En notación matricial la forma de la hipótesis es:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

siendo $R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}'$ y $r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

4. Significación global, $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

En notación matricial la forma de la hipótesis es:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

siendo $R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}'$ y $r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Mínimos cuadrados restringidos

La estimación del modelo clásico sujeto a un conjunto de restricciones lineales puede llevarse a cabo de dos formas equivalentes:

(1) incorporando las restricciones en la ecuación y

(2) aplicando la fórmula general del estimador de mínimos cuadrados restringidos.

Mientras que la forma (1) es útil en aplicaciones prácticas cuando se utiliza un programa de ordenador con capacidad para el análisis de regresión, la forma (2) es interesante para derivar las propiedades estadísticas generales del estimador.

El modelo que se obtiene al incorporar la hipótesis lineal $H_0 : R\beta - r = 0$ en $y = X\beta + u$ se denomina modelo con restricciones o modelo restringido.

En el problema de contraste $H_0 : R\beta = r$ versus $H_1 : R\beta \neq r$, se rechaza H_0 al nivel de significación α si

$$F \cong \frac{[R\hat{\beta} - r]'[\hat{\sigma}^2 R(X'X)^{-1}R'][R\hat{\beta} - r]}{q} > c$$

donde c es el valor crítico para el cual $prob(F_{q,n-k} > c) = \alpha$.

En el problema de contraste $H_0 : R\beta = r$ versus $H_1 : R\beta \neq r$, se rechaza H_0 al nivel de significación α si

$$F \cong \frac{\frac{SCR_{CR} - SCR_{SR}}{GL_{CR} - GL_{SR}}}{\frac{SCR_{SR}}{GL_{SR}}} > c$$

donde c es el valor crítico para el cual $prob(F_{GL_{CR}-GL_{SR},GL_{SR}} > c) = \alpha$. SCR_{CR} y GL_{CR} son la suma de cuadrados de los residuos y los grados de libertad en el modelo con restricciones, respectivamente, y SCR_{SR} y GL_{SR} son estas magnitudes en el modelo sin restricciones.

Para realizar el contraste de la hipótesis lineal general se siguen siguientes pasos:

1. Se estima el modelo sin restricciones

$$y = X\hat{\beta} + \hat{u}$$

que proporciona la suma de cuadrados de los residuos, $SCR_{SR} = \sum \hat{u}'\hat{u}$, y los grados de libertad, $GL_{SR} = n - k$;

2. se estima el modelo con restricciones $y = X\hat{\beta}_* + \hat{u}_*$ y $R\hat{\beta}_* = r$,

que proporciona la suma de cuadrados de los residuos, $SCR_{SR} = \sum \hat{u}'_*\hat{u}_*$, y los grados de libertad, $GL_{SR} = n - (k - q)$;

3. se calcula el estadístico de contraste

$$F = \frac{\frac{SCR_{CR} - SCR_{SR}}{GL_{CR} - GL_{SR}}}{\frac{SCR_{SR}}{GL_{SR}}}$$

4. finalmente, se compara el valor del estadístico F con el valor crítico c para el cual $prob(F_{q,n-k} > c = \alpha$. Si $F < c$, se acepta H_0; en caso contrario, se rechaza H_0.

Predicción

Una vez estimado y validado el modelo, una de sus aplicaciones más importantes consiste en poder realizar predicciones acerca del valor que tomaría la variable endógena en el futuro o para una unidad extramuestral. Esta predicción se puede realizar tanto para un valor individual como para un valor medio, o esperado, de la variable endógena, siendo posible efectuar una predicción puntual o por intervalos. Su cálculo se realiza mediante las expresiones que figuran a continuación:

a) Predicción individual

Nos interesa predecir el valor y_0 de la variable dependiente asociado al vector de valores conocidos $x_0 = \begin{pmatrix} 1 & x_{02} \cdots & x_{0k} \end{pmatrix}'$ de las variables explicativas. Parece razonable predecir y_0 como:

$$\hat{y}_0 = x'_0\hat{\beta}$$

que se denomina predicción puntual de y_0.

Para derivar las propiedades estadísticas de la predicción puntual, necesitamos extender el marco del modelo clásico con los siguientes supuestos sobre la observación a predecir:

1. el valor y_0 es una realización del modelo lineal general, es decir, $y_0 = x'_0\beta + u_0$;
2. el vector $x_0 = \begin{pmatrix} 1 & x_{02} \cdots & x_{0k} \end{pmatrix}'$ asociado a y_0 es conocido;
3. el error u_0 es una variable aleatoria (normal) con media 0 y varianza σ_u^2 , siendo independiente de $u_i (i = 1, ..., n) : E(u_0) = 0$, $E(u_0)^2 = \sigma_u^2$ y $E(u_0 u_i) = 0$ para $i = 1, ..., n$.

El error de predicción es la diferencia entre el valor observado y_0 y su pronóstico $\hat{y}_0$,

$$e_0 = y_0 - \hat{y}_0$$

Bajo el supuesto 1, el error de predicción

$$e_0 = x'_0\beta + u_0 - x'_p\hat{\beta} = x'_0(\beta - \hat{\beta}) + u_0$$

es la suma de dos componentes: (1) el error en la estimación de los parámetros $x'_0(\beta - \hat{\beta})$ y (2) el error aleatorio inherente al modelo u_0.

El error de predicción e_0 sigue una distribución normal con media cero y varianza $\sigma_u^2(1 + X'_0(X'X)^{-1}X_0)$.

La predicción $\hat{y}_0 = x_0'\hat{\beta}$ es lineal, insesgada y óptima.

b) Intervalo de predicción. La predicción por intervalo o el intervalo de confianza para y_0 de nivel $100(1-\alpha)\%$ es

$$\hat{y}_0 \pm c\sqrt{\hat{var}(e_0)}$$

donde c es el valor crítico para el cual $prob(t_{n-k} > c) = \frac{\alpha}{2}$.

o bien

$$\hat{y}_0 \pm t_{n-k,\frac{\alpha}{2}}\hat{\sigma}_u\sqrt{1 + X_0'(X'X)^{-1}X_0}$$

Estimacion del modelo de regresión con R

Modelo de regresión lineal simple

Creamos en primer lugar un vector de reales mediante la función c y lo guardamos con el nombre "Cantidad".

```
Cantidad <-c(2.456,2.325,2.250,2.200,2.100,2.082,2.045,2.024)
```

Se crea ahora el vector de nombre "Precio".

```
Precio<-c(82,92,94,99,106,108,112,115)
```

Para obtener los estadísticos básicos del vector (Cantidad): media, desviación estandar, varianza y mediana, se utilizan las siguientes funciones R:

```
mean(Cantidad)
```

```
## [1] 2.18525
```

```
sd(Cantidad)
```

```
## [1] 0.1515847
```

```
var(Cantidad)
```

```
## [1] 0.02297793
```

```
median(Cantidad)
```

```
## [1] 2.15
```

Si se quiere tener un resumen sumario de estadísticas de una variable:

```
summary(Cantidad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.024   2.073   2.150   2.185   2.269   2.456
```

La función de R que nos permite estimar un modelo de regresión lineal es la función lm. La forma de invocar a la función para estimar un modelo de regresión lineal simple es lm(y~x). Se puede consultar la ayuda de la función para ver todas las posibilidades que ofrece.

En nuestro ejemplo, obtenemos:

```
lm(Cantidad~Precio)
```

```
##
## Call:
## lm(formula = Cantidad ~ Precio)
##
```

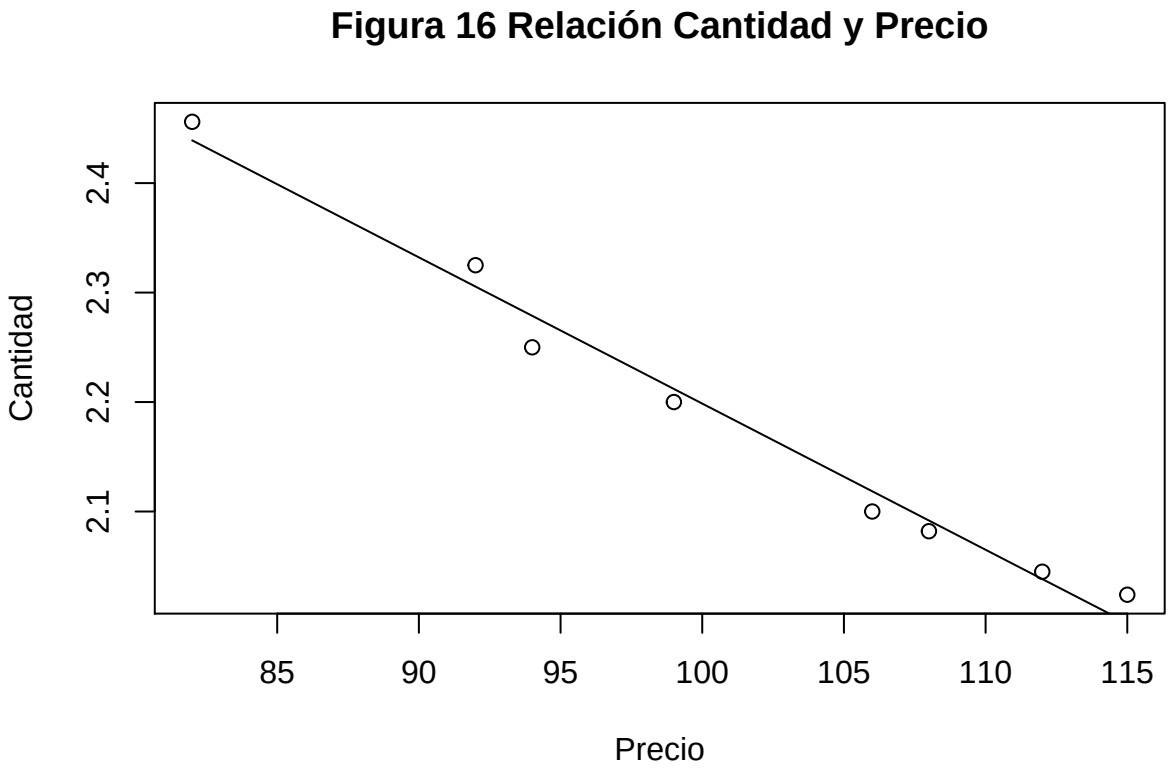
```
## Coefficients:
## (Intercept)       Precio
##     3.53427     -0.01336
```

En lugar de invocar simplemente la función podemos guardar su resultado en una variable y veremos así que obtenemos más información.

```
reg = lm(Cantidad~Precio)
```

Si queremos obtener una gráfica con los resultados de la regresión realizada:

```
plot(Cantidad~Precio, main="Figura 16 Relación Cantidad y Precio")
lines(reg$fitted~Precio)
```



Para realizar el análisis del modelo estimado utilizaremos la función summary. Así:

```
summary(reg)
```

```
##
## Call:
## lm(formula = Cantidad ~ Precio)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.02875 -0.01359 -0.00154  0.01762  0.02574
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5342726  0.0734707   48.10 5.41e-09 ***
## Precio      -0.0133567  0.0007235  -18.46 1.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.02154 on 6 degrees of freedom
## Multiple R-squared:  0.9827, Adjusted R-squared:  0.9798
## F-statistic: 340.8 on 1 and 6 DF,  p-value: 1.629e-06
```

Como puede observarse, el test de significación global, con un valor de $F = 340.8$ y $p - value = 0.000001629$ nos indica que el coeficiente β_1 (precio) tiene un valor significativamente distinto de 0. Al tratarse de una regresión lineal simple, este contraste es equivalente al derivado de la t de Student con $H_0 : \beta_1 = 0$, donde como vemos obtenemos el mismo p-value, siendo el valor del estadístico t la raiz cuadrada del estadístico F ($t^2 = -18.46^2 = 340.8$).

Por otra parte, el parámetro β_0 es también significativamente distinto de 0, con un valor del estadístico $t = 48.10$ y un $p - value = 5.41e - 09$.

Respecto a la bondad del ajuste, se obtiene un $R^2 = 0.9827$, indicando un alto grado de ajuste.

Finalmente, el modelo estimado ha sido:

$$Cantidad_i = 3.5342726 - 0.0133567 * Precio_i$$

Modelo de regresión lineal múltiple

Utilizando la base de datos "mtcars", con datos sobre automoviles y sus características que incluye el programa R, cuya estructura se lista con la función "str":

```
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Realizamos una regresión lineal multiple entre el consumo de gasolina y todas las caracteristicas que la base de datos incorpora para cada coche.

```
summary(lm(mpg~.,data=mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
```

```
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

El sumario de estadísticas sobre la regresión nos ofrece los resultados de los contrastes de significación individual sobre los parámetros y el de significación global. El hecho de que el contraste global sea significativo ($F = 13.93; p = 3.793e-07$) y ninguno de los individuales lo sea, es debido a la existencia de multicolinealidad en las variables independientes. Este problema tiene distintas soluciones, como pueden ser la selección de variables o la aplicación de técnicas de reducción de dimensiones en las variables explicativas, cuestiones que veremos más adelante.

Para realizar una predicción del consumo de gasolina (mpg) en el modelo lineal que utiliza como variables explicativas el peso del vehículo (wt) y los caballos de vapor (hp), en un coche que pese 3 (1.000 lbs) y tenga 120 hp hay que utilizar la función R "predict".

```
summary(lm(mpg~wt+hp,data=mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

```
newdatos=data.frame(wt=3,hp=120)
predict(lm(mpg~wt+hp,data=mtcars),newdatos,interval="prediction",type="response")
```

```
##        fit      lwr      upr
## 1 21.78102 16.38174 27.18031
```

Tercera parte: Extensiones al modelo de regresión lineal

Introducción

Como veíamos en el apartado anterior, el modelo de regresión lineal requiere que se cumplan las siguientes hipótesis sobre los términos de error:

- Media cero: $E(u_i) = 0\ i = 1, \ldots, n$
- Varianza constante: $Var(u_i) = \sigma^2 I\ i = 1, \ldots, n$
- Residuos incorrelacionados: $Cov(u_i, u_j) = 0$

El incumplimiento de alguna de dichas hipótesis, implica la no aleatoriedad de los residuos y, por tanto, la existencia de alguna estructura o relación de dependencia en los residuos que puede ser estimada, debiendo ser considerada en la especificación inicial del modelo. Los principales problemas asociados al incumplimiento de las hipótesis de normalidad de los residuos son, por un lado, la heteroscedasticidad, cuando la varianza de los mismos no es constante, y la autocorrelación o existencia de relación de dependencia o correlación entre los diferentes residuos, lo que violaría el supuesto de términos de error incorrelacionados.

Si se construye una gráfica de los resultados de una estimación mínimo cuadrática (en ordenadas) frente al valor absoluto de los residuos (en abscisas), cuando éstos últimos presentan una distribución aleatoria, es decir una distribución Normal de media cero y varianza constante, $N(0, \sigma^2)$, el resultado obtenido (véase Figura 2) muestra que el tamaño del error es independiente del tamaño de la variable estimada, ya que errores con valor elevado se corresponden con valores bajos y altos de la variable dependiente estimada; sin embargo, una distribución de residuos con problemas de heteroscedasticidad da lugar a una figura como la que puede observarse en la Figura 3, en donde se manifiesta una clara relación de dependencia entre la variable estimada y el tamaño del error. En este caso los errores de mayor tamaño se corresponden con los valores más altos de la variable estimada.

Figura 17 Residuos Normales

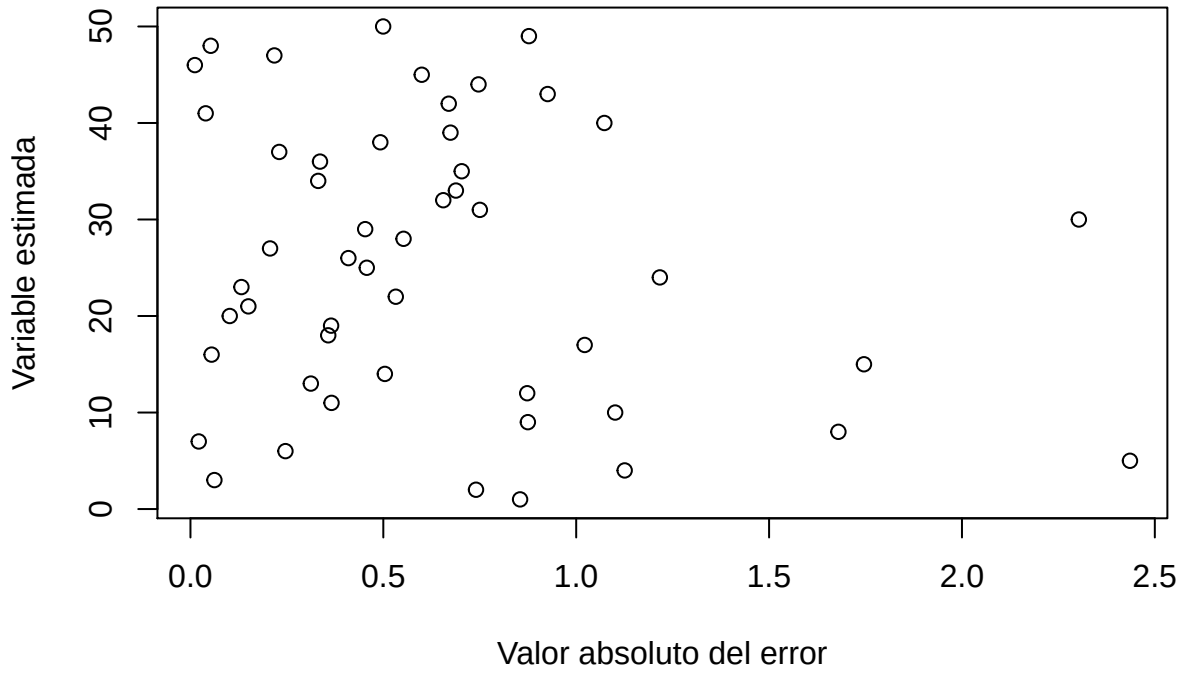
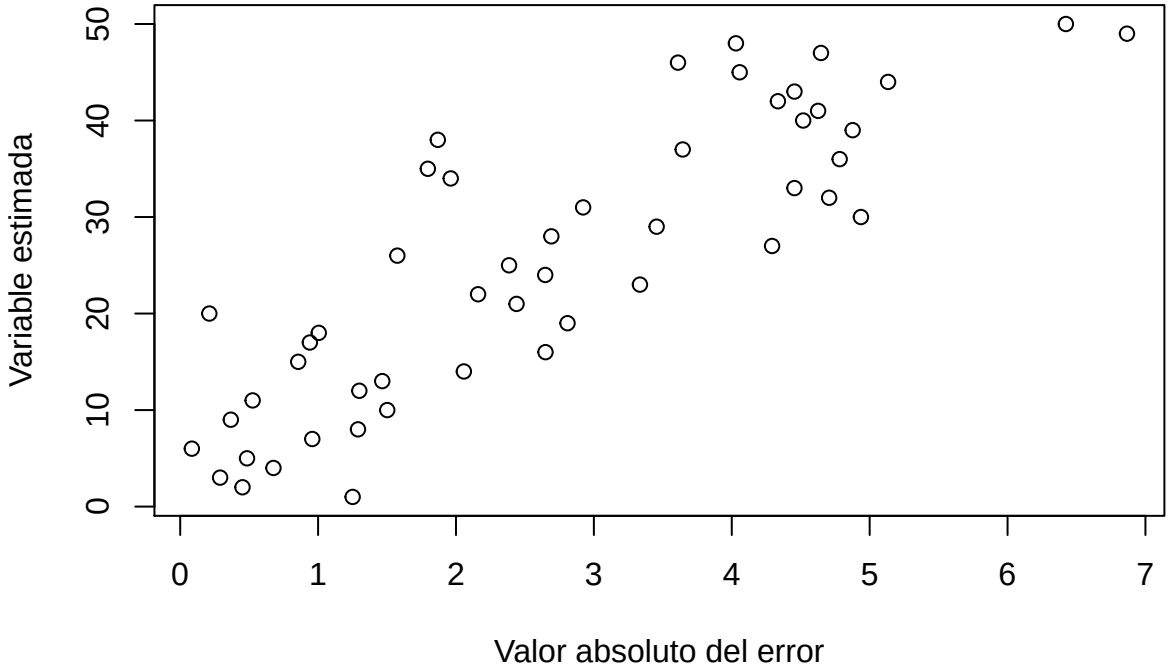
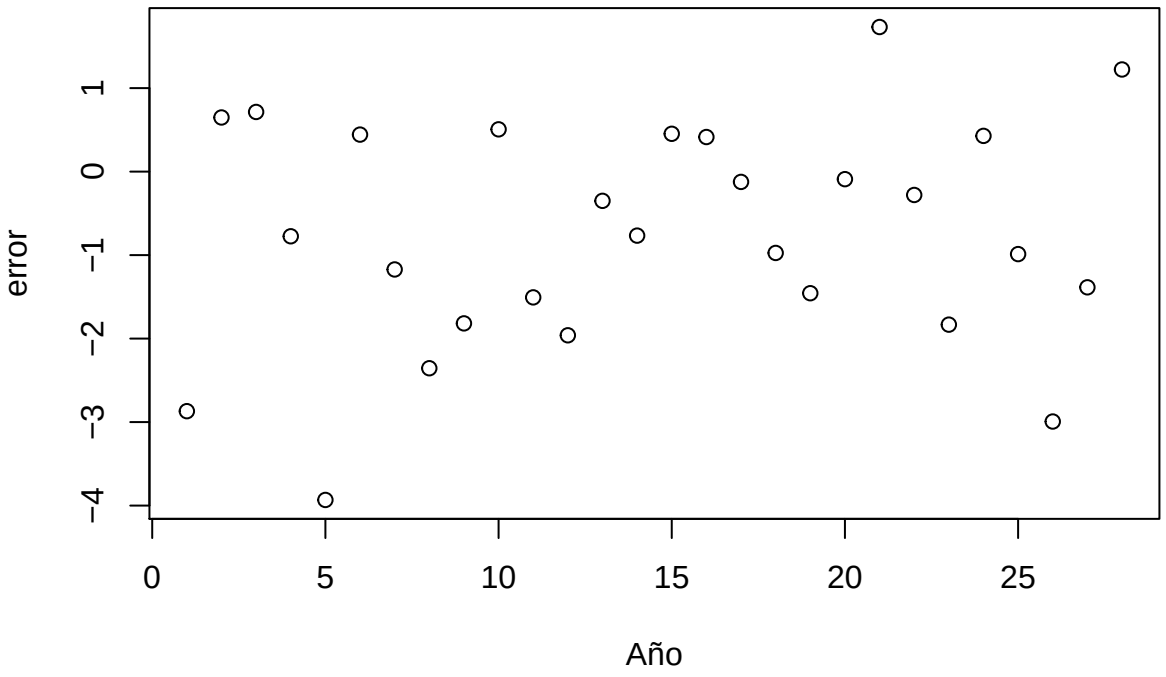


Figura 18 Residuos Heterocedásticos



La representación gráfica de los errores en forma de serie temporal, es decir, poniendo en el eje de ordenadas los errores y en abscisas el periodo temporal en que están datados, permite apreciar la ausencia o presencia de correlación, ya que a los residuos no correlacionados (Figura 4) les corresponde una representación gráfica en la que no se aprecia pauta temporal alguna, sucediéndose de forma impredecible o aleatoria, mientras que en los residuos con problemas de autocorrelación la pauta temporal es evidente, evidenciándose que cada residuo podría ser previsto en función de la sucesión de los errores correspondientes a periodos temporales pasados.

Figura 19 Residuos con autocorrelación



Estos problemas asociados a los errores pueden detectarse con tests estadísticos diseñados para ello.

Normalidad

Para decidir si una muestra aleatoria proviene de una determinada distribución normal podemos usar métodos visuales y contrastes de hipótesis.

Métodos visuales

Algunos gráficos que podemos visualizar para contrastar normalidad son los siguientes:

1. El histograma o gráfico de barras de la distribución de frecuencias de la muestra aleatoria. Se suele dibujar junto con la función de densidad de la distribución $N(\bar{Y}, s_Y^2)$ para comprobar mejor si tiene forma de campana, es simétrico y unimodal. Cuando se muestran en un mismo gráfico el histograma y la función de densidad, la frecuencia del eje de ordenadas se reemplaza por la **densidad** definida como:

$$densidad = \frac{frecuencia\ relativa}{ancho\ del\ intervalo}$$

2. La función de distribución (FDA) empírica, que se define como la proporción de datos menores que y:

$$\hat{F}(y) = \frac{nž\ de\ datos\ menores\ o\ iguales\ que\ y}{nž\ total\ de\ datos} = \frac{\sum_{t=1}^{n} I(Y_t \leq y)}{n}$$

donde $I(Y_t \leq y)$ es una función indicador que toma el valor 1 cuando la condición $Y_t \leq y$ es verdadera, y el valor O cuando dicha condición es falsa.

Denotando el conjunto de datos ordenados de menor a mayor como $\{Y_{(t)} : t = 1, \cdots, n\}$, se cumple que $Y_{(1)} < Y_{(2)} < \cdots Y_{(n)}$ siendo $Y_{(t)}$ el dato t-ésimo menor de la muestra. Usando esta notación, la FDA empírica es la función escalón:

$$\hat{F}(y) = \begin{cases} 0 \; si \; y < Y_{(1)} \\ \frac{t}{n} \; si \; Y_{(t)} \leq y > Y_{(t+1)} \\ 1 \; si \; y \geq Y_{(n)} \end{cases}$$

La gráfica de la FDA empírica se dibuja representando los pares $(Y_{(t)}, \frac{t}{n})$ en un diagrama XY. Después, se unen estos pares en forma de escalera, siendo 1/n la altura de cada escalón cuando no hay datos repetidos. Esta gráfica debería ser similar a la de la FDA de la distribución formulada bajo H_0, $N(\mu_0, \sigma_0^2)$.

3. El gráfico Q-Q muestra, en un diagrama XY, los cuantiles empíricos frente a los cuantiles teóricos de la distribución normal. El valor $Y_{(t)}$ suele asociarse al t/(n+1)-ésimo cuantil empírico; el t/(n+1)-ésimo cuantil teórico en una $N(\mu_0, \sigma_0^2)$ es $F^{-1}[t/(n+1)]$, es decir, el valor que deja a su izquierda una probabilidad t/(n+1). Bajo H_0, el conjunto de pares $(F^{-1}[t/(n+1)], Y_{(t)})$ debería situarse sobre la diagonal de 45^o. El gráfico Q-Q también puede construirse usando los pares $(F^{-1}[(t-0.5)/n], Y_{(t)})$.
4. El gráfico P-P muestra, en diagrama XY, la probabilidad acumulada empírica t/(n+1) frente a la probabilidad acumulada teórica $F(Y_{(t)})$ de la distribución $N(\mu_0, \sigma_0^2)$. Bajo H_0, el conjunto de pares $(F(Y_{(t)}), t/(n+1))$ debería situarse aproximadamente sobre la diagonal de 45^o. También suele utilizarse (t-0.5)/n como probabilidad empírica.

Partiendo de la base de datos cars, analizamos a continuación con estos cuatro gráficos los residuos de la regresión entre la distancia de frenado y la velocidad.

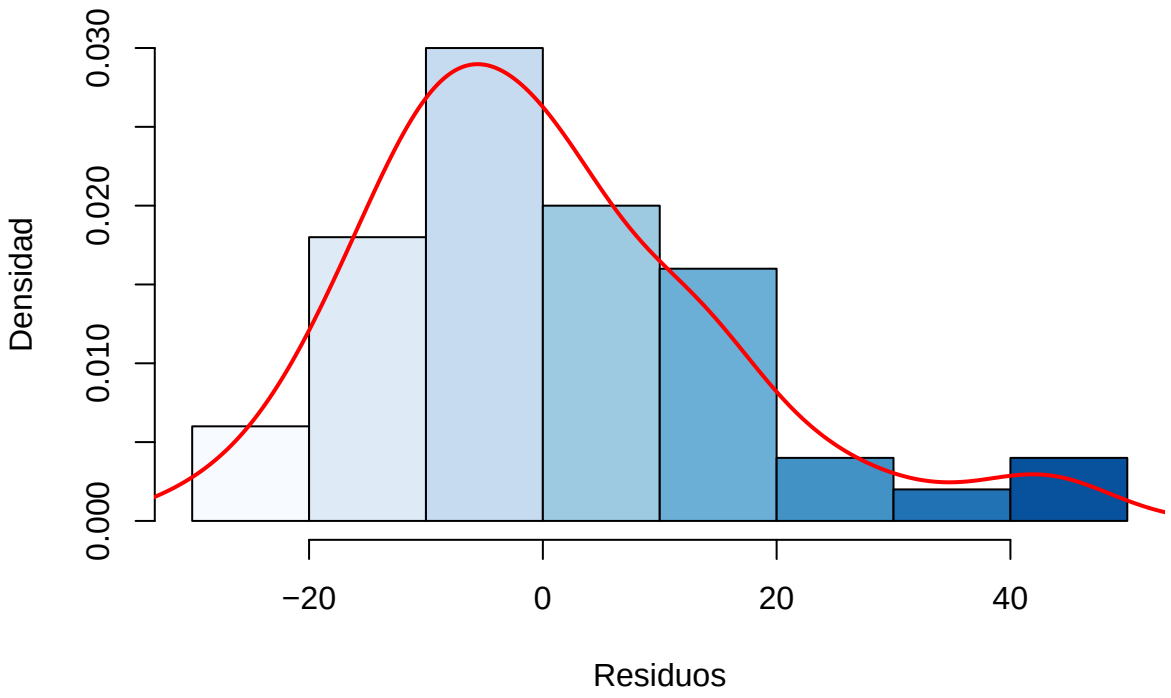
```
mod1=lm(dist~speed,cars)
summary(mod1)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
residuos <- mod1$residuals
minimo <- min(residuos)
maximo <- max(residuos)
```

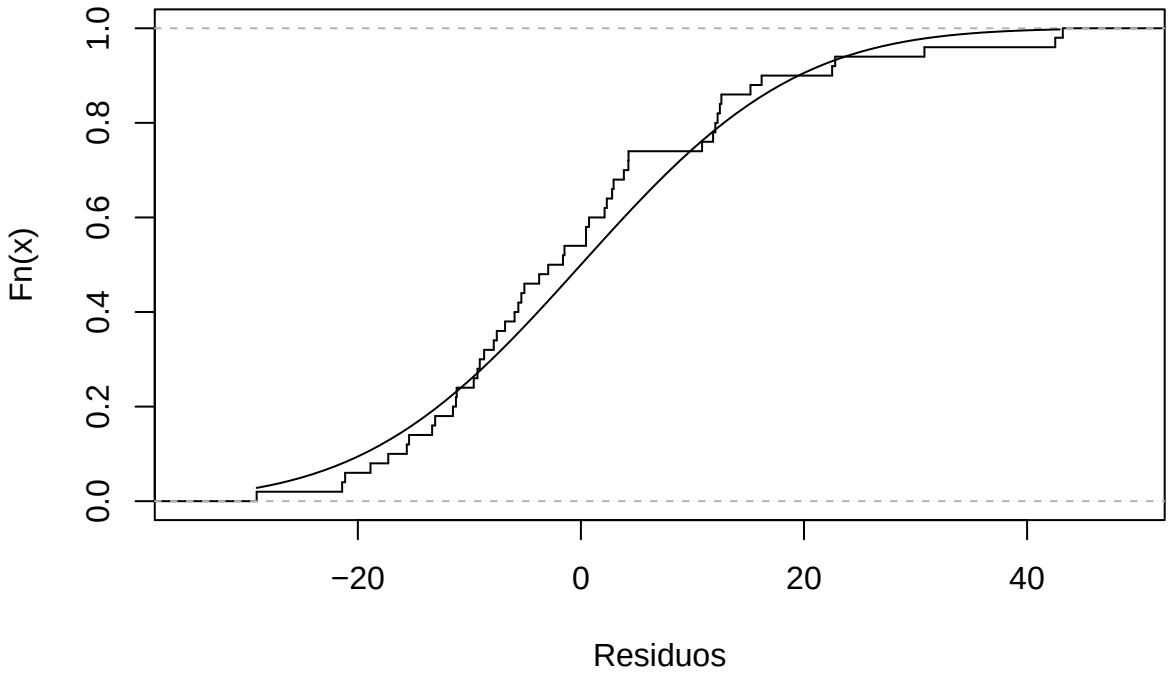
```
# Histograma
hist(residuos, freq = FALSE, main = "Figura 20 Histograma y función de densidad empírica y teórica", yla
# Calculamos la densidad
dx <- density(residuos)

# Añadimos la línea de densidad
lines(dx, lwd = 2, col = "red")
```

Figura 20 Histograma y función de densidad empírica y teórica




```
FDAemp = ecdf(residuos)
plot(FDAemp, verticals = TRUE, do.points = FALSE, main = "Figura 21 Comparación FDA empírica y teórica"
lines(minimo:maximo, pnorm(minimo:maximo, mean = mean(residuos), sd = sd(residuos)), type = "l")
```

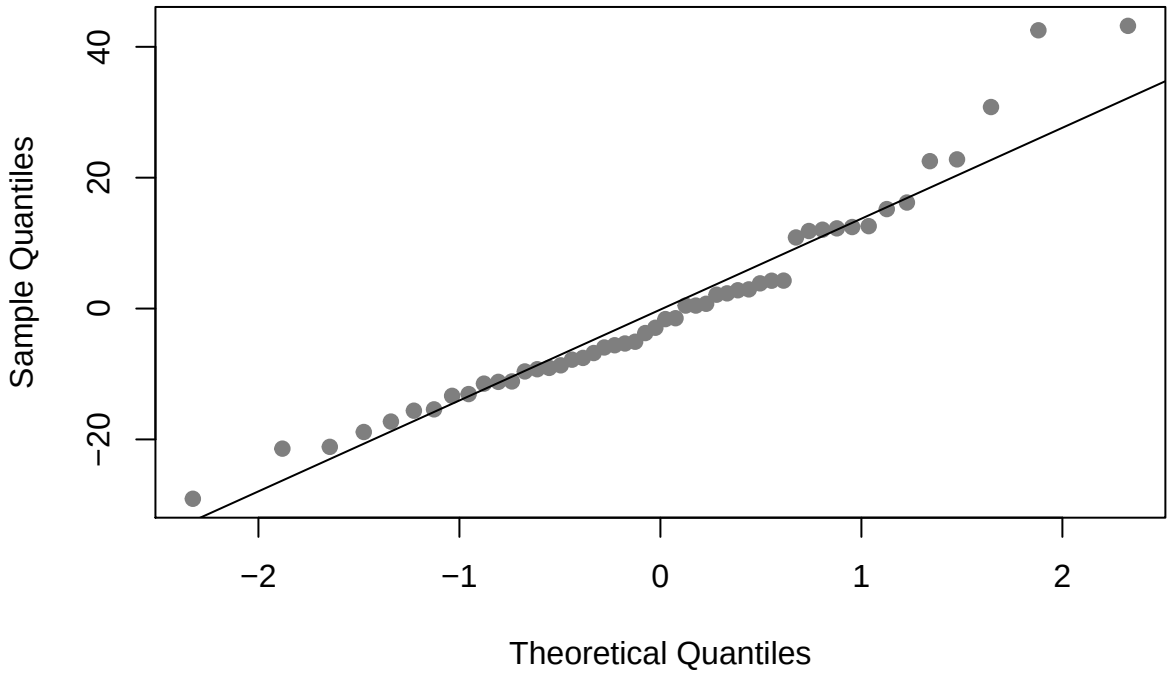
Figura 21 Comparación FDA empírica y teórica




```
qqnorm(residuos, pch = 19, col = "gray50", main = "Figura 22 Plot Q-Q")
```

```
qqline(residuos)
```

Figura 22 Plot Q−Q




```
# Gráfico P-P
library(StatDA)
```

```
## Loading required package: sgeostat
```

```
## Warning in fun(libname, pkgname): couldn't connect to display ":0"
```

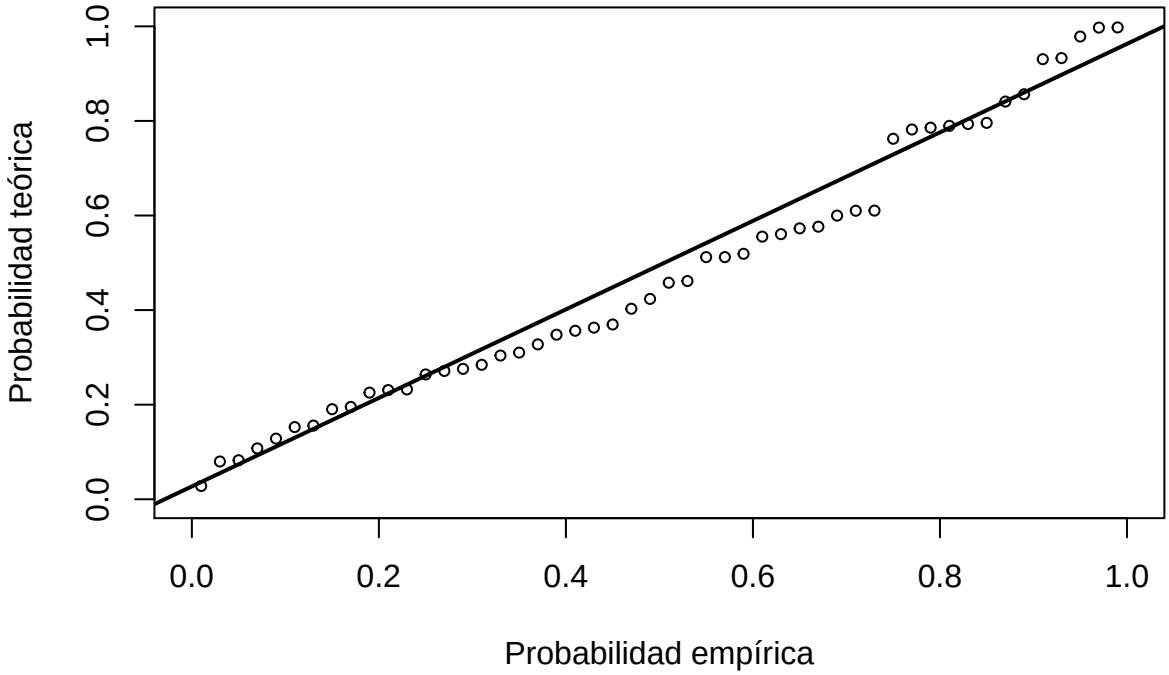
```
## Registered S3 method overwritten by 'geoR':
##   method         from
##   plot.variogram sgeostat
```

```
## Warning in rgl.init(initValue, onlyNULL): RGL: unable to open X11 display
```

```
## Warning: 'rgl.init' failed, running with 'rgl.useNULL = TRUE'.
```

```
ppplot.das(residuos, pdist = pnorm, ylab = "Probabilidad teórica", xlab = "Probabilidad empírica", pch=
title("Figura 23 Plot P-P")
```

Figura 23 Plot P−P



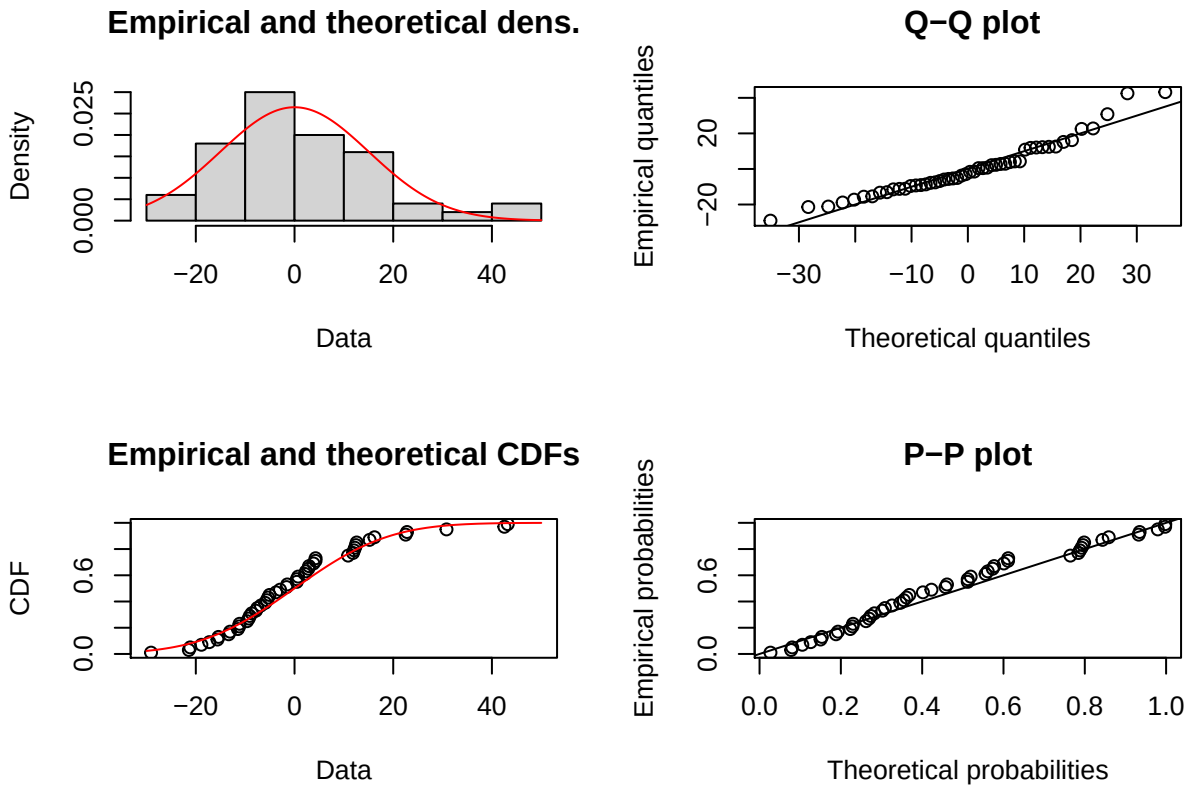
La librería **fitdistrplus** a través de la función *fitdist* nos permite realizar estos cuatro gráficos mediante una sola instrucción.

```
library(fitdistrplus)
```

```
## Loading required package: MASS
```

```
## Loading required package: survival
```

```
fit_normal <- fitdist(residuos, distr = "norm")
plot(fit_normal)
```



Contrastes de normalidad

Aunque los métodos gráficos son muy informativos ('una imagen vale más que...') a menudo se critican porque se prestan a interpretaciones subjetivas y no permiten calcular el nivel de significación. Los contrastes más utilizados serían:

Contrastes de Kolmogorov-Smirnov y Lilliefors

Un procedimiento más formal para contrastar normalidad es el contraste de bondad de ajuste de Kolmogorov-Stnirnov, si bien, señalar que en realidad es de carácter general, y está diseñado para determinar la bondad de ajuste de dos distribuciones de probabilidad entre sí.

La hipótesis nula del contraste KS establece que la muestra proviene de una población continua con FDA $F_0(y)$; la alternativa, que la FDA es diferente.

$$H_0 : F(y) = F_0(y)$$

$$H_1 : F(y) \neq F_0(y)$$

El estadístico de contraste evalúa la diferencia máxima entre las funciones $F_0(y)$ y $\hat{F}_0(y)$.

$$KS = \max_{1 \leq t \leq n} \left| F_0(Y_t) - \frac{t}{n} \right|$$

cuya distribución muestral no es estándar y se tabula por simulación. Se rechaza H_0 frente a H_1 al nivel de significación α cuando $KS > c_{\alpha;n}$, donde el valor crítico depende también de n. Cuando $F_0(y)$ está completamente especificada, los valores críticos relevantes son $c_{0.01;n} = 1.63/\sqrt{n}$, $c_{0.05;n} = 1.36/\sqrt{n}$ y $c_{0.10;n} = 1.22/\sqrt{n}$ para n > 35. Sin embargo, estos valores críticos no son válidcs cuando se ha estimado algún parámetro de $F_0(y)$, siendo necesario calcularlos por simulación.

En el caso de que queramos verificar la normalidad de una distribución, la prueba de Lilliefors conlleva algunas mejoras con respecto a la de Kolmogórov-Smirnov. Para esta prueba las estimaciones de la media y la varianza de la población están basadas en los datos. Posteriormente, se calcula la discrepancia máxima entre la función de distribución empírica y la función de distribución acumulativa de la distribución normal con la media estimada y la varianza estimada. Al igual que en la prueba de Kolmogorov-Smirnov, esta será la estadística de prueba. Finalmente, se evalúa si la discrepancia máxima es lo suficientemente grande como para ser estadísticamente significativa, lo que requiere el rechazo de la hipótesis nula. Aquí es donde esta prueba se vuelve más complicada que la prueba de Kolmogorov-Smirnov. Debido a que la hipotética $F_0(y)$ se ha movido más cerca de los datos mediante una estimación basada en esos datos, la discrepancia máxima se ha hecho más pequeña de lo que hubiera sido si la hipótesis nula hubiera destacado solo una distribución normal. Por lo tanto, la "distribución nula" del estadístico de prueba, es decir, su distribución de probabilidad suponiendo que la hipótesis nula es cierta, es estocásticamente más pequeña que la distribución de Kolmogorov-Smirnov. Hasta la fecha, las tablas para esta distribución se han calculado solo mediante los métodos de Monte Carlo. Puede calcularse mediante la función lillie.test() del paquete **nortest**.

Test de Shapiro-Wilk

Se plantea como hipótesis nula que una muestra $x_1, \cdots, x_n$ proviene de una población normalmente distribuida. Fue publicado en 1965 por Samuel Shapiro y Martin Wilk. Se considera uno de los test más potentes para el contraste de normalidad. El estadístico del test es:

$$W = \frac{(\sum_{t=1}^{n} a_t Y_{(t)})^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

donde las variables a_t, que suelen aparecer tabuladas en los manuales, se calculan:

$$(a_1, \cdots, a_n) = \frac{m' V^{-1}}{(m' V^{-1} V^{-1} m)^{1/2}}$$

siendo $m_1, ..., m_n$ los valores medios del estadístico ordenado, de variables aleatorias independientes e idénticamente distribuidas, muestreadas de distribuciones normales y V es la matriz de covarianzas de ese estadístico de orden.

La hipótesis nula se rechazará si W es demasiado pequeño. El valor de W puede oscilar entre 0 y 1.

Interpretación: siendo la hipótesis nula que la población está distribuida normalmente, si el p-valor es menor a α (nivel de significación) entonces la hipótesis nula es rechazada (se concluye que los datos no vienen de una distribución normal). Si el p-valor es mayor a α, se concluye que no se puede rechazar dicha hipótesis.

La normalidad se verifica, pues, confrontando dos estimadores alternativos de la varianza σ^2:

• un estimador no paramétrico (numerador), y • un estimador paramétrico, varianza muestral (denominador).

Usaremos este estadístico en el caso de muestras pequeñas ($n \leq 50$).

Test Jarque-Bera

El **test de Jarque-Bera** no requiere estimaciones de los parámetros que caracterizan la normal. El estadístico de Jarque-Bera cuantifica que tanto se desvían los coeficientes de asimetría y curtosis de los esperados en una distribución normal. Su formulación es la siguiente:

$$JB = \frac{n}{6}\left(A^2 + \frac{(C-3)^2}{4}\right) \approx X_2^2$$

donde S es la asimetría de la muestra y C la curtosis definidas como:

$$A = \frac{m^3}{s^3} = \frac{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^3}{\left[\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2\right]^{3/2}}$$

$$C = \frac{m^4}{s^4} = \frac{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^4}{\left[\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2\right]^{2}}$$

El estadístico de Jarque-Bera se distribuye asintóticamente ($\approx$) como una distribución chi cuadrado con dos grados de libertad, y puede usarse para probar la hipótesis nula de que los datos pertenecen a una distribución normal. El hecho de que la relación sea aproximada implica que la distribución muestral del estadístico es conocida en muestras muy grandes (n>2000), pero desconocida en muestras pequeñas.

Como hipótesis nula se considera que la asimetría y el exceso de curtosis, conjuntamente, son nulos, es decir:

$$H_0 : A = 0; \quad C = 3$$

El estadítico JB rechaza la hipótesis de normalidad al nivel de significación α cuando $JB > c_\alpha$, donde c_α es el valor crítico tal que $Pr(X_2^2 > c_\alpha) = \alpha$. En muestras finitas, los valores críticos se calculan por simulación.

Puede calcularse mediante la función jarque.bera.test() del paquete **tseries**.

Los resultados de los cuatro test comentados con la serie de errores del modelo de regresión entre distancias de frenado y velocidades serían:

```
# Kolmogorov-Smirnov
ks.test(residuos, "pnorm", mean(residuos), sd(residuos))
```

```
## Warning in ks.test(residuos, "pnorm", mean(residuos), sd(residuos)): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuos
## D = 0.12957, p-value = 0.3708
## alternative hypothesis: two-sided
```

```
# Lilliefors
library(nortest)
lillie.test(residuos)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuos
## D = 0.12957, p-value = 0.03529
```

```
# Shapiro-Wilk
shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.94509, p-value = 0.02152
```

```
# Jarque-Bera
library("tseries")
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
jarque.bera.test(residuos)
```

```
##
##  Jarque Bera Test
##
## data:  residuos
## X-squared = 8.1888, df = 2, p-value = 0.01667
```

Existen muchos otros test de normalidad, entre ellos citar: Anderson-Darling, Cramer-von Mises, Pearson chi-square, Shapiro-Francia, Frosini, Geary, Hegazy-Green, Spiegelhalter, Weisberg-Bingham, Agostino, entre otros.

Heterocedasticidad

Si existe heterocedasticidad en los residuos de nuestro modelo, esto implicará que la precisión de la explicación del modelo no es constante. En el campo de la economía, la heterocedasticidad puede provenir de cambios estructurales, eventos especiales, etc.

La consecuencia que lleva asociada es que la estimación de la varianza de los parámetros no es correcta, es decir, $\sigma^2(X'X)^{-1}$ no es una estimación consistente de $Var(\hat{\beta})$, pudiendo llevar a conclusiones erróneas sobre los β_k. Sin embargo, la expresión $\hat{\beta} = (X'X)^{-1}X'Y$ sigue siendo una estimación consistente de β.

Para tratar este problema podemos optar por:

- Mejorar la especificación del modelo (añadir, quitar o transformar variables).
- Utilizar estimadores consistentes para $Var(\beta)$, como por ejemplo los propuestos por White (1980) o por Newey y West (1987).

La detección de la heteroscedasticidad se realiza a través de diversos contrastes paramétricos, entre los que cabe destacar el contraste de Bartlett (Mood, 1950), el constraste de Goldfeld-Quandt (1965) y el contraste de White (1980).

El contraste de White se basa en que, bajo la hipótesis nula de homocedasticidad, la matriz de varianzas y covarianzas de los estimadores MCO de $\hat{\beta}$ es: $\sigma^2(X'X)^{-1}$

Por el contrario, si existe heteroscedasticidad, la matriz de varianzas y covarianzas viene dada por:

$$(X'X)^{-1}X'\Omega X(X'X)^{-1}, \Omega = (\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$$

Por tanto, si tomamos la diferencia entre ambas queda:

$$(X'X)^{-1}X'\Omega X(X'X)^{-1} - \sigma^2(X'X)^{-1}$$

Por ello, basta con contrastar la hipótesis nula de que todas estas diferencias son iguales a cero, lo que equivale a contrastar que no hay heteroscedasticidad.

Los pasos a seguir para realizar el contraste de White son los siguientes:

1. Estimar el modelo original y obtener la serie de residuos estimados.
2. Realizar una regresión del cuadrado de la serie de residuos obtenidos en el paso anterior sobre una constante, las variables exógenas del modelo original, sus cuadrados y los productos cruzados de segundo orden (los productos resultantes de multiplicar cada variable exógena por cada una de las restantes). Es decir, se trata de estimar por MCO la relación:

$$e_i^2 = \alpha + \varphi_1 X_{1i} + ... + \varphi_k X_{ki} + \eta_1 X_{1i}^2 + ... + \eta_k X_{ki}^2 + \varpi_1 X_{1i} X_{2i} + ... + \varpi_k X_{1i} X_{ki} + ... + \rho X_{k-1i} X_{ki} + u_i$$

3. Al aumentar el tamaño muestral, el producto nR^2 (donde n es el número de observaciones y R^2 es el coeficiente de determinación de la última regresión) sigue una distribución Chi-cuadrado con $p-1$ grados de libertad, donde p es el número de variables exógenas utilizadas en la segunda regresión. Se aceptará la hipótesis de existencia de heteroscedasticidad cuando el valor del estadístico supere el valor crítico de la distribución Chi-cuadrado (c) al nivel de significación estadística fijado ($nR^2 > c$).

Para realizar en R el constraste de heterocedasticidad de White en el modelo estimado para las distancias y velocidades de la base de datos "cars" sobre distancias y velocidades de frenado de automóviles de los años 20 que incorpora R, se puede utilizar la función función bptest de la libreria "lmtest", especificando en el argumento varformula y la fórmula con los términos no lineales:

```
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(mod1, varformula = ~ speed  + I(speed^2),data=cars)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 3.2157, df = 2, p-value = 0.2003
```

En el caso particular de que la hipótesis de homogeneidad de varianzas sea rechazada debido a que la varianza cambia con la media, existe una familia de transformaciones que proporciona, en general, homogeneidad en varianzas. Esta familia es de la forma:

$$T(X) = \begin{bmatrix} X^p & si\ p \neq 0 \\ ln(X) & si\ p = 0 \end{bmatrix}$$

Dentro de esta familia está la denominada transformación de **Box-Cox**, cuya formulación es la siguiente:

$$z(\lambda) = \begin{bmatrix} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & si\ \lambda \neq 0 \\ \dot{y} ln(y) & si\ \lambda = 0 \end{bmatrix}$$

siendo $\dot{y}$ la media geométrica de los valores de y.

Haciendo el análisis de la varianza con las variables $z(\lambda)$, se procede a estimar λ por máxima verosimilitud, es decir:

$$L(\lambda) = \frac{n}{2} ln \sum \sum e_{ij}^2(\lambda)$$

En R se realiza la transformación box-cox con la función "boxcox"" de la libreria "EnvStats". Por defecto, selecciona el λ en base al criterio de Probability Plot Correlation Coefficient (PPCC).

```r
library(EnvStats)
```

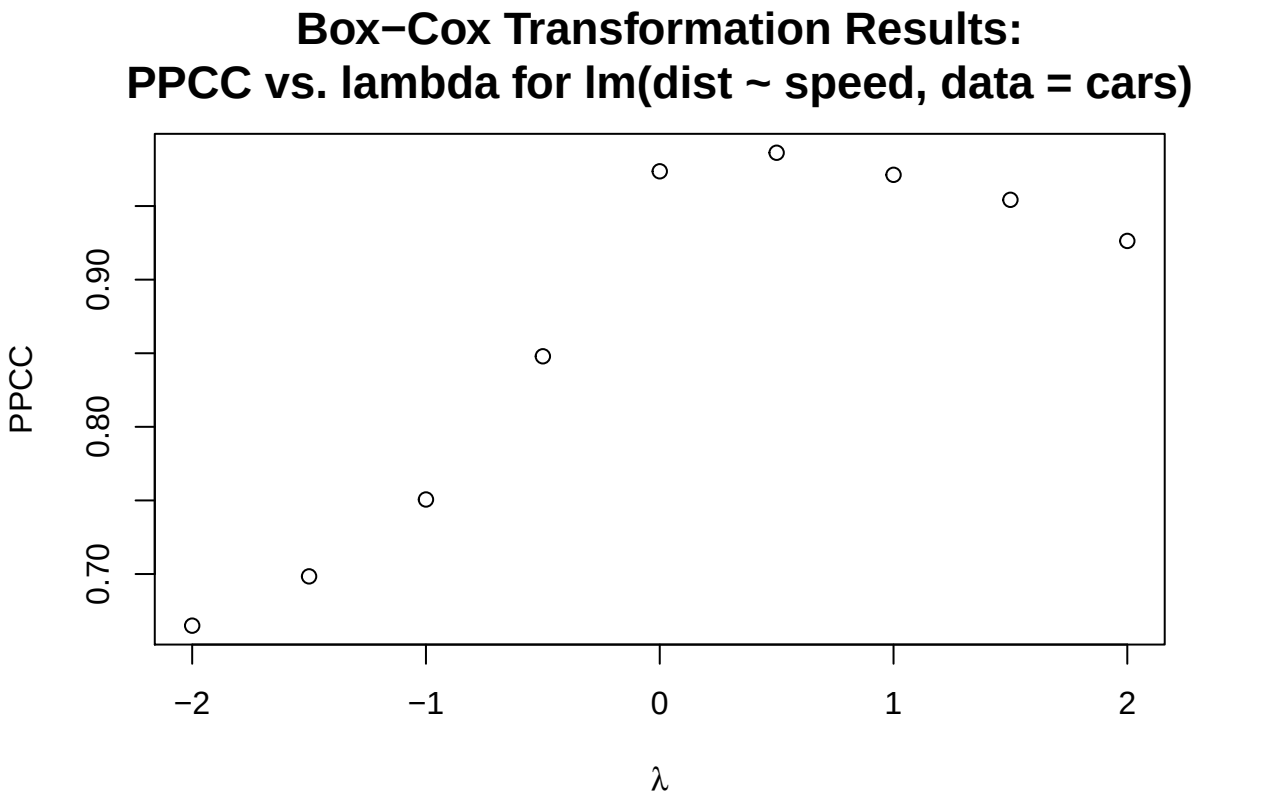
```
##
## Attaching package: 'EnvStats'
## The following object is masked from 'package:MASS':
##
##     boxcox
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
## The following object is masked from 'package:base':
##
##     print.default
```

```r
boxcox(lm(dist ~ speed, data = cars))
```

```
## $lambda
## [1] -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0
##
## $objective
## [1] 0.6649249 0.6983992 0.7506285 0.8479300 0.9736106 0.9862199 0.9712216
## [8] 0.9542555 0.9263064
##
## $objective.name
## [1] "PPCC"
##
## $optimize
## [1] FALSE
##
## $optimize.bounds
## lower upper
##    NA    NA
##
## $eps
## [1] 2.220446e-16
##
## $lm.obj
##
## Call:
## lm(formula = dist ~ speed, data = cars, y = TRUE, qr = TRUE)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
##
##
## $sample.size
## [1] 50
##
## $data.name
## [1] "lm(dist ~ speed, data = cars)"
##
## attr(,"class")
```

```
## [1] "boxcoxLm"
```

```
plot(boxcox(lm(dist ~ speed, data = cars)))
```




```
lambda <- boxcox(lm(dist ~ speed, data = cars), optimize = TRUE)$lambda
lambda
```

```
## [1] 0.2235103
```

Autocorrelación

Decimos que existe autocorrelación cuando el término de error de un modelo econométrico está correlacionado consigo mismo a través del tiempo, tal que $Cov(e_i, e_j) \neq 0$. Ello no significa que la correlación entre los errores se dé en todos los periodos, sino que puede darse tan sólo entre algunos de ellos.

La razón más frecuente es que uno o varios de los parámetros β_j no son constantes durante la muestra. En el campo de la economía, puede provenir de cambios estructurales, eventos especiales, etc.

Al igual que en el caso anterior, en presencia de autocorrelación, los estimadores MCO siguen siendo insesgados pero no poseen mínima varianza, debiéndose utilizar en su lugar el método de estimación de los Mínimos Cuadrados Generalizados (MCG). Lógicamente, si la falta de constancia en el valor de los parámetros es grave, nuestra estimación no servirá para nada.

Para tratar este problema podemos optar por:

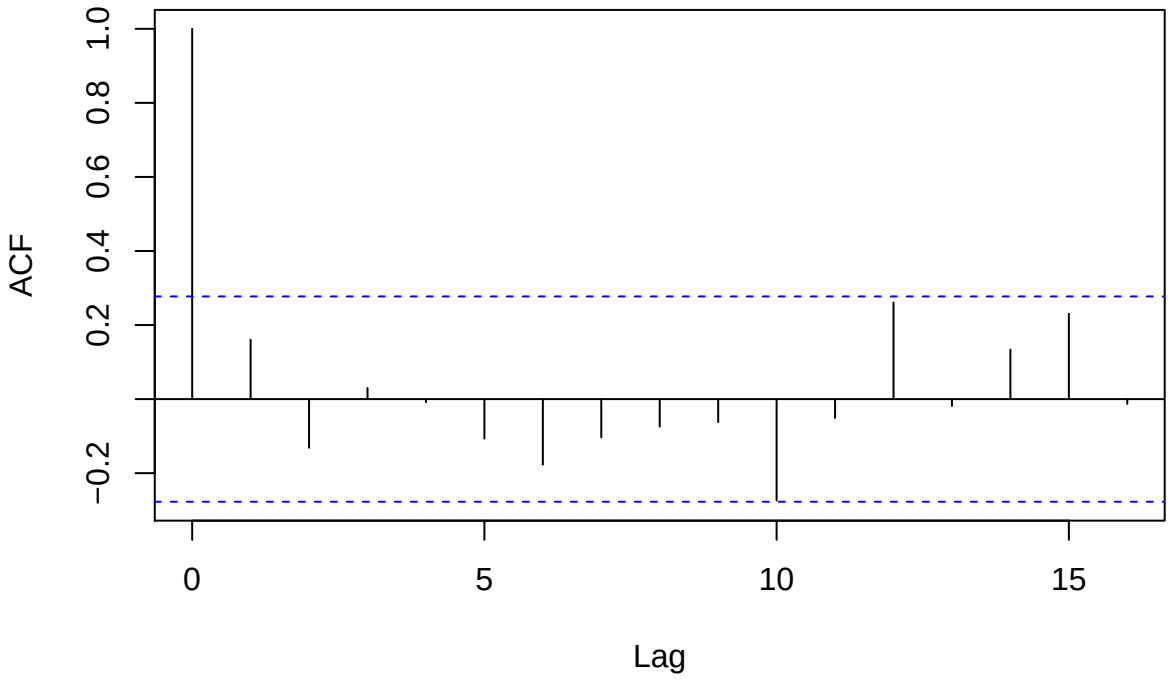
- Mejorar la especificación del modelo (añadir, quitar o transformar variables) o añadir los errores obtenidos como una variable mas para neutralizar la autocorrelación.
- Utilizar estimadores consistentes para $Var(\beta)$, como por ejemplo el de Newey y West (1987).

La existencia de autocorrelación en los residuos es fácilmente identificable obteniendo las funciones de autocorrelación (acf) y autocorrelación parcial (acp) de los errores mínimo-cuadráticos obtenidos en la estimación. Si dichas funciones corresponden a un ruido blanco, se constatará la ausencia de correlación entre

los residuos. Sin embargo, el mero examen visual de las funciones anteriores puede resultar confuso y poco objetivo, por lo que en la práctica econométrica se utilizan diversos contrastes para la autocorrelación, siendo el más utilizado el de Durbin-Watson (1950), que pasamos a ver seguidamente.

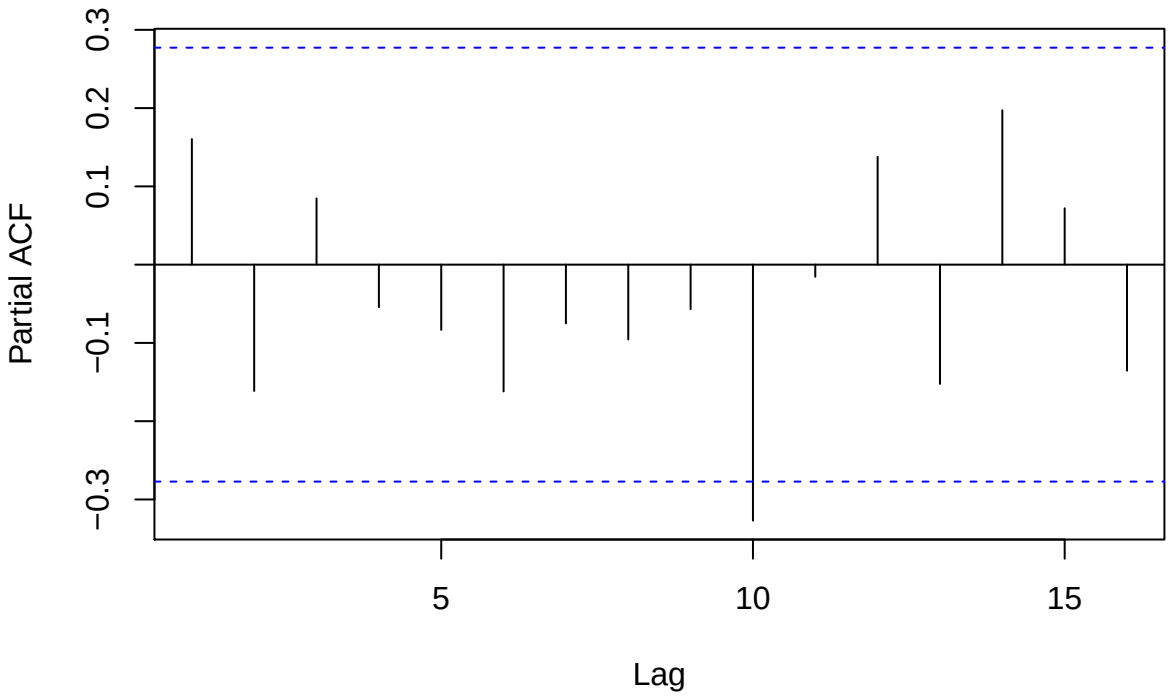
```
acf(residuos, main="Figura 24 Función de autocorrelación")
```

Figura 24 Función de autocorrelación




```
pacf(residuos, main="Figura 25 Función de autocorrelación parcial" )
```

Figura 25 Función de autocorrelación parcial



Contraste de Durbin-Watson

Si se sospecha que el término de error del modelo econométrico tiene una estructura como la siguiente:

$$\hat{e}_t = \rho \hat{e}_{t-1} + u_t$$

entonces el contraste de Durbin-Watson permite contrastar la hipótesis nula de ausencia de autocorrelación.

Dicho contraste se basa en el cálculo del estadístico d, utilizando para ello los errores mínimo-cuadráticos resultantes de la estimación:

$$d = \frac{\sum_{t=2}^{n} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{n} \hat{e}_t^2}$$

El valor del estadístico d oscila entre 0 y 4, siendo los valores cercanos a 2 los indicativos de ausencia de autocorrelación de primer orden. La interpretación exacta del test resulta compleja, ya que los valores críticos apropiados para contrastar la hipótesis nula de no autocorrelación requieren del conocimiento de la distribución de probabilidad bajo el supuesto de cumplimiento de dicha hipótesis nula, y dicha distribución depende a su vez de los valores de las variables explicativas, por lo que habría que calcularla en cada aplicación. Se ha demostrado que el valor esperado de d, cuando $\rho = 0$, está dado por la siguiente relación (Maddala,1996).

$$E(d) \approx 2 + \frac{2k}{n-k-1}$$

Para facilitar la interpretación del test, Durbin y Watson derivaron dos distribuciones: d_L y d_U, que no dependen de las variables explicativas y entre las cuales se encuentra la verdadera distribución de d, de forma que a partir de un determinado nivel de significación, se adoptan la siguientes reglas de decisión:

1. Si $d \leq d_L$ rechazamos la hipótesis nula de no autocorrelación frente a la hipótesis alternativa de autocorrelación positiva.

2. Si $d \geq (4 - d_L)$ rechazamos la hipótesis nula de no autocorrelación frente a la hipótesis alternativa de autocorrelación negativa.
3. Si $d_U \leq d \leq (4 - d_U)$ aceptamos la hipótesis nula de no autocorrelación.
4. Si $d_L \leq d \leq d_U$ ó $(4 - d_U) \leq d \leq (4 - d_L)$ la prueba no es concluyente.

El estadístico de Durbin-Watson es aproximadamente igual a $2(1 - r)$ en donde r es el coeficiente de autocorrelación simple muestral del retardo 1.

En R, el test de Durbin-Watson se encuentra en el Package-R "lmtest", y su sintaxis es:

```
library(lmtest)
z <- cbind.data.frame(y=cars$dist, x=cars$speed)
dwtest(y~x, data = z)
```

```
##
##  Durbin-Watson test
##
## data:  y ~ x
## DW = 1.6762, p-value = 0.09522
## alternative hypothesis: true autocorrelation is greater than 0
```

Contraste de Breush-Godfrey

El test de correlación serial de Breusch–Godfrey es un test de autocorrelación en los errores y residuos estadísticos en un modelo de regresión. Hace uso de los errores generados en el modelo de regresión y un test de hipótesis derivado de éste. La hipótesis nula es que no exista correlación serial de cualquier orden de ρ .

El test es más general que el de Durbin–Watson, que solo es válido para regresores no-estocásticos y para testear la posibilidad de un modelo autorregresivo de primer orden para los errrores de regresión. El test Breusch–Godfrey no tiene estas restricciones, y es estadísticamente más poderoso que el estadístico d.

Los pasos para realizar el contraste son los siguientes:

1. Estimar el modelo original y obtener la serie de residuos estimados.
2. Estimar la ecuación de regresión auxiliar:

$$\hat{e}_t = \alpha + \omega_1 X_1 + \omega_2 X_2 + ... + \omega_k X_k + \delta_1 \hat{e}_{t-1} + ... + \delta_p \hat{e}_{t-p} + \epsilon_t$$

3. Al aumentar el tamaño muestral, el producto $(n - p)R^2$ (donde n es el número de observaciones, p el número de retardos del error utilizados en la regresión auxiliar y R^2 es el coeficiente de determinación de dicha regresión) sigue una distribución Chi-cuadrado con p grados de libertad. Se aceptará la hipótesis de existencia de autocorrelación cuando el valor del estadístico supere el valor crítico de la distribución Chi-cuadrado (c) al nivel de significación estadística fijado $(n - p)R^2 > c$.

El test de Breusch–Godfrey tambien se realiza con la librería-R "lmtest", y se programa para $p = 3$ del siguiente modo:

```
library(lmtest)
bgtest(y~x, order = 3, data = z)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 3
##
## data:  y ~ x
## LM test = 3.0936, df = 3, p-value = 0.3774
```

Deficiencias muestrales

El fenómeno de la multicolinealidad aparece cuando las variables exógenas de un modelo econométrico están correlacionadas entre sí, lo que tiene consecuencias negativas para la estimación por MCO, ya que la existencia de una relación lineal entre las variables exógenas implica que la matriz $(X'X)$ va a tener determinante cero, es decir será una matriz singular y, por tanto, no será invertible. Dado que $\hat{\beta} = (X'X)^{-1}X'y$, no será posible calcular la estimación mínimo cuadrática de los parámetros del modelo ni, lógicamente, la varianza de los mismos. Esto es lo que se conoce por el nombre de **multicolinealidad exacta**.

Errores de especificación

Los errores de especificación hacen referencia a un conjunto de errores asociados a la especificación de un modelo econométrico. En concreto cabe referirse a:

- Omisión de variables relevantes.
- Inclusión de variables innecesarias.
- Adopción de formas funcionales equivocadas.

En Economía, la teoría no suele concretar la forma funcional de las relaciones que estudia. Así, por ejemplo, cuando se analiza la demanda se señala que la cantidad demandada es inversamente proporcional al precio; cuando se estudia el consumo agregado se apunta que la propensión marginal a consumir (relación entre renta y/o consumo) es mayor que cero y menor que uno. Por otro lado, es frecuente utilizar la condición "ceteris paribus" para aislar la información de otras variables relevantes que influyen y/o modifican la relación estudiada. Por esta razón, la existencia de errores de especificación en la relación estimada es un factor a considerar y a resolver en el proceso de la estimación econométrica.

Con independencia de la naturaleza de los errores de especificación, dado que en el proceso de estimación MCO deben de cumplirse determinadas hipótesis básicas (que los estimadores MCO deben de ser insesgados, eficientes y consistentes, y que el estimador de la varianza del término de error ha de ser insesgado), debemos preguntarnos: ¿qué ocurriría con estas propiedades ante errores de especificación?

La omisión de variables relevantes provoca que el estimador MCO sea sesgado en media (se obtiene un valor distinto que el que se obtendría al incorporar la variable relevante omitida) y en varianza. La incorporación de variables innecesarias determina que la varianza de los estimadores MCO sea mas elevada (sesgo en la varianza), dificultando la interpretación del contraste de significación individual sobre los parámetros.

Si especificamos la forma funcional de una relación (ya sea lineal, cuadrática, cúbica, exponencial, logarítmica, etc.) y la verdadera relación presenta una forma diferente a la especificada tiene, en algunos casos, las mismas consecuencias que la omisión de variables relevantes, es decir, proporciona estimadores sesgados e inconsistentes. En general, una especificación funcional incorrecta lleva a obtener perturbaciones heteroscedásticas y/o autocorrelacionadas, o alejadas de los parámetros de la distribución del término de error del modelo correctamente especificado.

Métodos de selección de variables en el modelo lineal general

Una de las cuestiones más importantes a la hora de encontrar el modelo de ajuste más adecuado cuando se dispone de un amplio conjunto de variables explicativas, es la correcta especificación del modelo teórico, ya que como se ha visto la inclusión de una variable innecesaria o la omisión de una variable relevante, condiciona los estadísticos que resultan en la estimación MCO del modelo. Por otro lado, en un elevado número de explicativas no cabe descartar la existencia de correlaciones que originen un problema de multicolinealidad aproximada, y en estos casos hay que determinar cual de ellas cabe incluir en la especificación del modelo.

En otras palabras, ante un conjunto elevado de variables explicativas debemos seleccionar de entre todas, un subconjunto de ellas que garanticen que el modelo esté lo mejor especificado posible. Este análisis cabe hacerlo estudiando las características y propiedades de cada una de las variables independientes, a partir, por ejemplo, de los coeficientes de correlación de cada una de ellas y la dependiente, y de cada explicativa con las restantes, seleccionando modelos alternativos y observando los resultados estadísticos de la estimación MCO

de cada uno de ellos. Sin embargo, en la práctica, la selección del subconjunto de variables explicativas de los modelos de regresión se deja en manos de procedimientos más o menos automáticos.

Los procedimientos más usuales son los siguientes:

• Método backward: se comienza por considerar incluidas en el modelo teórico a todas las variables disponibles y se van eliminando del modelo de una en una según su capacidad explicativa. En concreto, la primera variable que se elimina es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente-o lo que es equivalente, un menor valor del estadístico t– y así sucesivamente hasta llegar a una situación en la que la eliminación de una variable más suponga un descenso demasiado acusado en el coeficiente de determinación.

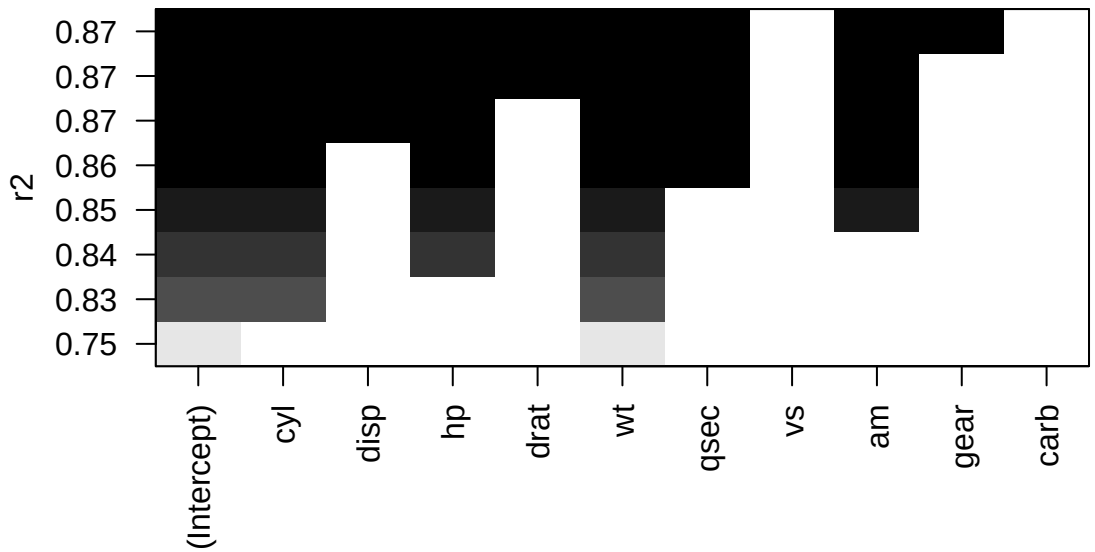
• Método forward: se comienza por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación -en valor absoluto- con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquella variable que presenta un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando el incremento en el coeficiente de determinación debido a la inclusión de una nueva variable explicativa en el modelo ya no es importante.

• Método stepwise: es uno de los más empleados y consiste en una combinación de los dos anteriores. En el primer paso se procede como en el método forward pero a diferencia de éste, en el que cuando una variable entra en el modelo ya no vuelve a salir, en el procedimiento stepwise es posible que la inclusión de una nueva variable haga que otra que ya estaba en el modelo resulte redundante.

El modelo de ajuste al que se llega partiendo del mismo conjunto de variables explicativas es distinto según cuál sea el método de selección de variables elegido, por lo que la utilización de un procedimiento automático de selección de variables no significa que con él se llegue a obtener el mejor de los modelos a que da lugar el conjunto de datos con el que se trabaja.

Para realizar la selección de un modelo por cualquira de los métodos descritos, necesitamos instalar la librería-R: "leaps". Una vez instalada ejecutamos el siguiente Chunk, para seleccionar según el método "forward":

```
library(leaps)
regfit.fwd = regsubsets(mpg~.,data=mtcars,method="forward")
plot(regfit.fwd,scale="r2")
```




```
summary(regfit.fwd)
```

```
## Subset selection object
```

```
## Call: regsubsets.formula(mpg ~ ., data = mtcars, method = "forward")
## 10 Variables  (and intercept)
##      Forced in Forced out
## cyl      FALSE      FALSE
## disp     FALSE      FALSE
## hp       FALSE      FALSE
## drat     FALSE      FALSE
## wt       FALSE      FALSE
## qsec     FALSE      FALSE
## vs       FALSE      FALSE
## am       FALSE      FALSE
## gear     FALSE      FALSE
## carb     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          cyl disp hp  drat wt  qsec vs  am  gear carb
## 1  ( 1 ) " " " "  " " " "  "*" " "  " " " " " "  " "
## 2  ( 1 ) "*" " "  " " " "  "*" " "  " " " " " "  " "
## 3  ( 1 ) "*" " "  "*" " "  "*" " "  " " " " " "  " "
## 4  ( 1 ) "*" " "  "*" " "  "*" " "  " " "*" " "  " "
## 5  ( 1 ) "*" " "  "*" " "  "*" "*"  " " "*" " "  " "
## 6  ( 1 ) "*" "*"  "*" " "  "*" "*"  " " "*" " "  " "
## 7  ( 1 ) "*" "*"  "*" "*"  "*" "*"  " " "*" " "  " "
## 8  ( 1 ) "*" "*"  "*" "*"  "*" "*"  " " "*" "*"  " "
```

Si queremos ver las estimaciones MCO de los parámetros del modelo 8:

```
coef(regfit.fwd, 8)
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.56350226 -0.23126963  0.01611609 -0.02339020  0.70893592 -4.08155351
##        qsec          am        gear
##  0.91812006  2.47759723  0.50403957
```