El modelo de regresión lineal clásico en R

Lino AA Notarantonio, Departamento de Economía

Tec de Monterrey, Campus Santa Fe

Contents

Inferencia estadística

Inferencia y distribución normal

Teorema Central del Límite

Sea

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

la media muestral de una muestra aleatoria X_1, X_2,..., X_n, *de tamaño* n *tomada de una población con media* μ *y desviación estándar* σ. *Entonces, la distribución de muestreo de* $\overline{X}$ *es aproximadamente normal*[1], *con media* μ *y desviación estándar* $\sigma/\sqrt{n}$, *cuando* $n \to \infty$.

La demostración del Teorema no es muy complicada, pero conviene "verla en acción" mediante unas simulaciones. Usaremos, como distribución poblacional, la distribución uniforme en el intervalo $(0, 1)$, con tamaño muestral sucesivamente igual a $n = 1$, $n = 100$ y $n = 1000$.

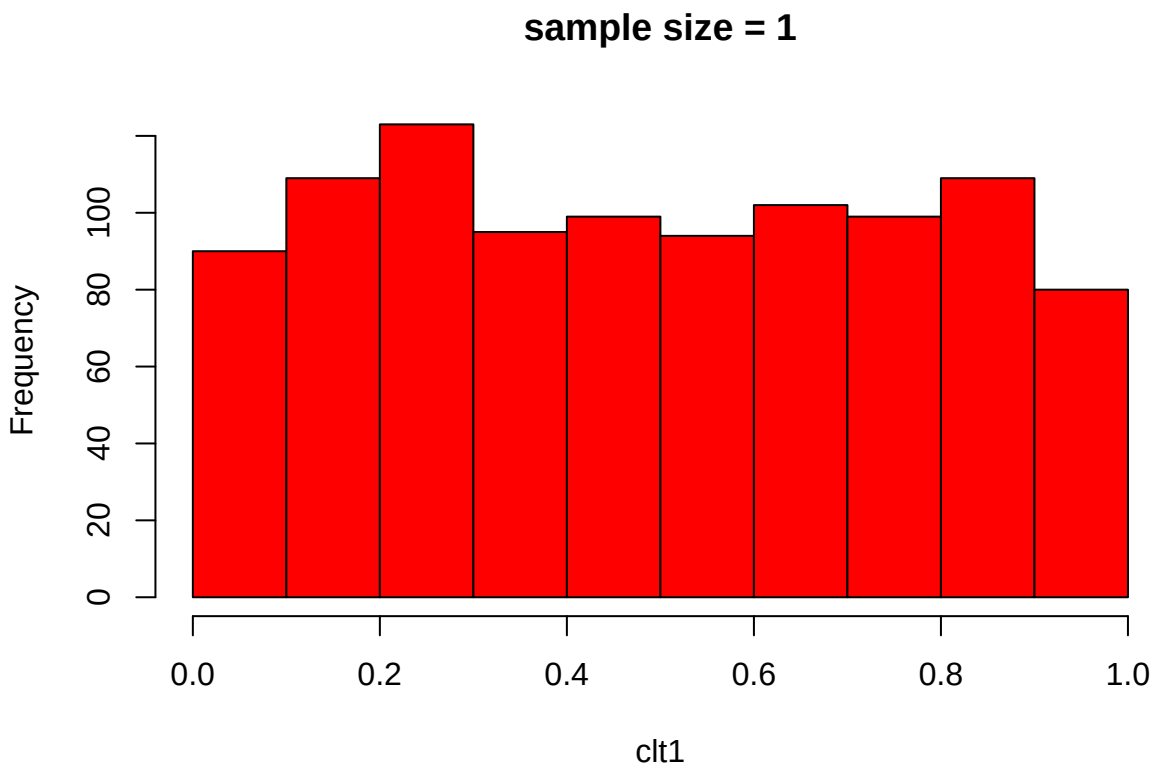
Tamaño muestral $n = 1$

```
clt1=numeric(1000)
for (i in 1:1000){x=runif(1);clt1[i]=mean(x)}
hist(clt1,col="red",main="sample size = 1")
```

[1]Más precisamente, la función de distribución de la media muestral estandardizada

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

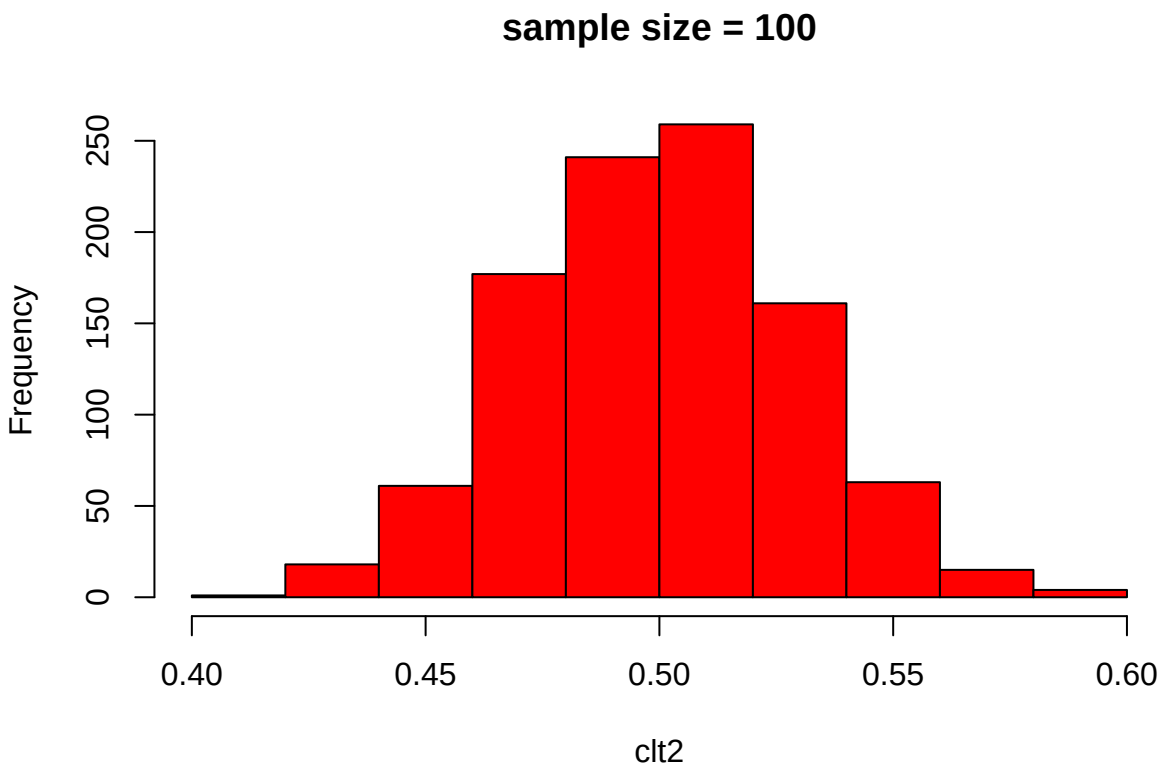
converge a la función de distribución de $N(0, 1)$.



En la simulación, observamos como el histograma es típico de una distribución uniforme en $(0, 1)$, en el sentido que los valores observados son todos igualmente probables.

Tamaño muestral $n = 100$

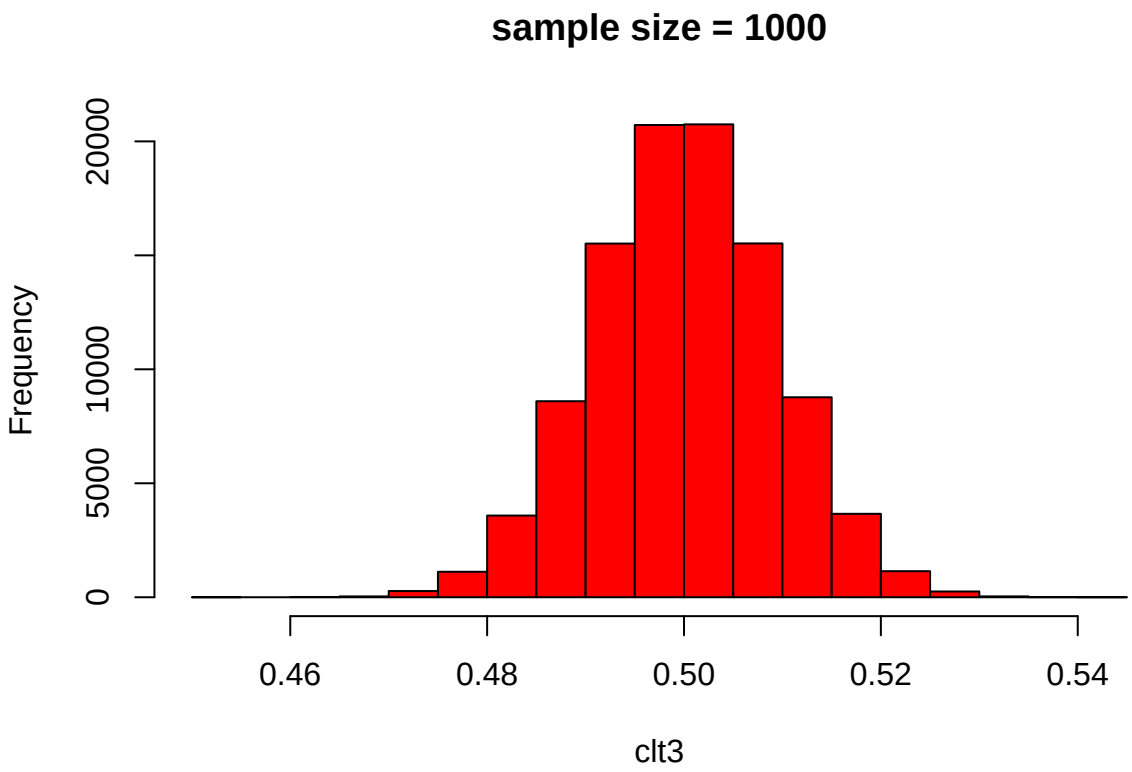
```
clt2=numeric(1000)
for (i in 1:1000){x=runif(100);clt2[i]=mean(x)}
hist(clt2,col="red",main="sample size = 100")
```



Observamos que el histograma comienza a tener forma de la "campana" típica de la distribución normal, aún si todavía se observan asimetrías alrededor de la moda (valor más frecuentemente observado). Los valores más observados se encuentran alrededor de la media poblacional $\mu = .5$.

Tamaño muestral $n = 1000$

```
clt3=numeric(100000)
for (i in 1:100000){x=runif(1000);clt3[i]=mean(x)}
hist(clt3,col="red",main="sample size = 1000")
```



El histograma tiene una forma simétrica, típica de la distribución normal. A comparación con el anterior ($n = 100$), la dispersión alrededor de $\mu = .5$ es mucho menor, efecto de la fórmula de la varianza muestral del estimador de la media muestral.

En conclusión, conforme el tamaño muestral aumente, no solamente la forma del histograma se acerca a la "campana" típica de la distribución normal; también la moda de las observaciones se acercan al valor $\mu = .5$ y la dispersión alrededor de μ es cada vez menor como función de $1/\sqrt{n}$.

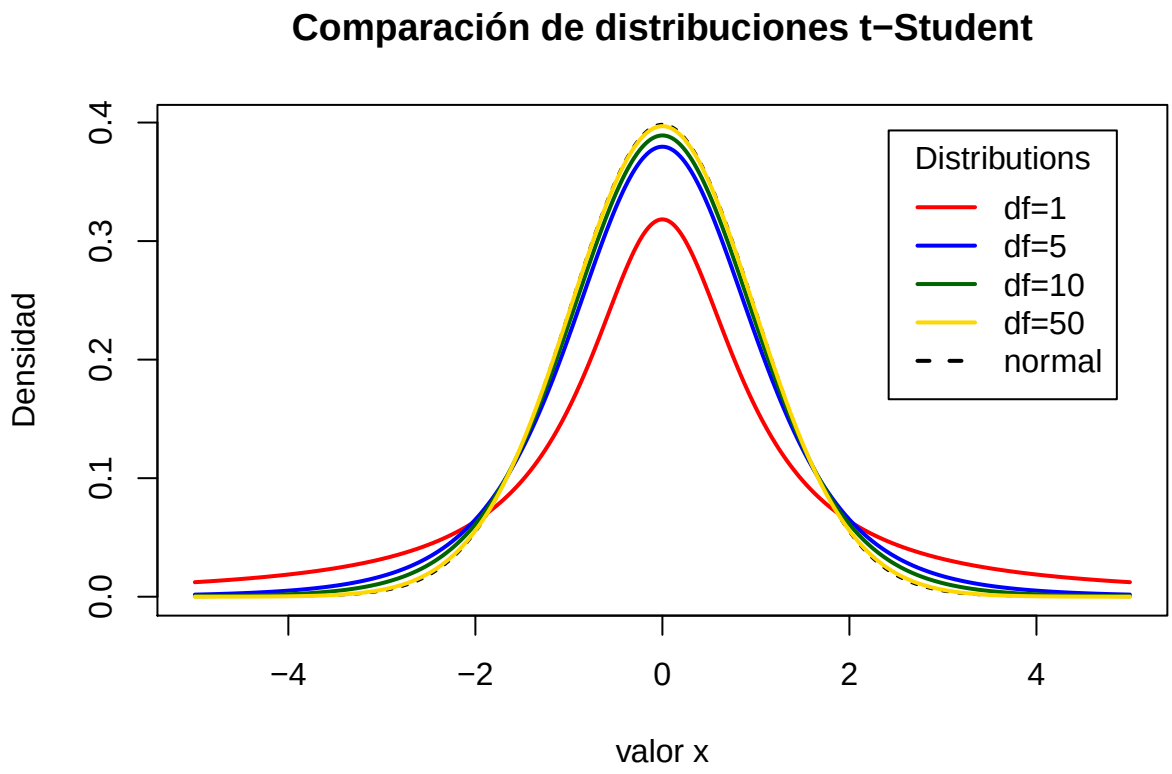
Comparación entre la distribución normal y la t-Student

En esta sección se comparan la distribución normal con la distribución t-Student, con un número de grados de libertad creciente.

El ejercicio puede ser interesante por el uso de vectorización en la definición de los colores de las diferentes curvas y de las etiquetas; además, se muestra como se puede usar un bucle `for` para añadir curvas a la gráfica final. Finalmente, el uso de una leyenda para la identificación y comparación rápida de las diferentes curvas.

```
x<-seq(-5,5,length=1000) # intervalo
hx <- dnorm(x) # densidad de la normal
degf <- c(1, 5, 10, 50) # g.l. distr t
colors <- c("red", "blue", "darkgreen", "gold", "black")
```

```
labels <- c("df=1", "df=5", "df=10", "df=50", "normal")
plot(x, hx, type="l", lty=2,
     xlab="valor x", ylab="Densidad",
     main="Comparación de distribuciones t-Student")
for (i in 1:4){ lines(x, dt(x,degf[i]), lwd=2, col=colors[i])}
legend("topright", inset=.05,
       title="Distributions", labels, lwd=2,
       lty=c(1, 1, 1, 1, 2), col=colors)
```



Prueba de hipótesis

Una prueba de hipótesis es una afirmación acerca de uno o más parámetros poblacionales.

Se toma una muestra representativa de la población de interés y se formula una hipótesis nula, H_0 (efecto nulo), y una hipótesis alternativa, H_1 (el efecto que se quiere observar).

Desarrollaremos las ideas de inferencia estadísticas con ejemplos concretos a continuación.

Ejemplo introductorio

En un proceso de inspección, se considera que la proporción de defectuosos es del 10% ($p = .10$). Si se toma una muestra aleatoria de tamaño $T = 100$ y se encuentran 9 artículos defectuosos, entonces la proporción observada de defectuosos no permite rechazar $p = .10$, porque la probabilidad de que este evento ocurra es

```
dbinom(9,100,.1)
```

```
## [1] 0.1304163
```

relativamente elevada.

Pero, si en la muestra aleatoria, $T = 100$, se encontraran 25 elementos defectuosos, entonces esta sí es evidencia en contra de $p = .10$, porque en este último caso la probabilidad que ocurran 25 éxitos en una secuencia de Bernoulli $T = 100$ es igual a

```
dbinom(25,100,.1)
```

```
## [1] 8.972934e-06
```

Esto es, es poco probable que 25 artículos de 100 sean defectuosos, si se supone que $p = .10$.

Planteamiento de una prueba de hipótesis

En una prueba de hipótesis, con un *nivel de significancia dado*, se necesitan

1. la declaración de H_0 y H_1;
2. un valor de probabilidad (nivel de significancia) que permita decidir claramente cuándo el evento observado es poco probable, si H_0 es verdadera;
3. un estadístico de prueba para el cálculo numérico.

Nivel de significancia; potencia de una prueba

En inferencia estadística se consideran dos tipos de errores.

Error tipo I es el rechazo de H_0 cuando es verdadera: el *nivel de significancia* está asociado a un Error Tipo I mediante

$$\alpha = \Pr(\text{ Error tipo I })$$

Error tipo II es el no rechazo de H_0 cuando es falsa:

$$\beta = \Pr(\text{ Error tipo II })$$

La *potencia* de una prueba es el complemente de un error tipo II:

$$\text{Potencia } = 1 - \beta = \Pr(\text{ Rechazar } H_0 \text{ cuando es falsa}).$$

A continuación, siempre usaremos un nivel de significancia dado en una prueba de hipótesis.

Falsos positivos; falsos negativos

Considerando que H_0 es el efecto nulo, un error tipo I se considera como un *falso positivo*; en una prueba médica, por ejemplo, un falso positivo es cuando un paciente no tiene cierta enfermedad, pero la prueba señala que sí.

Un error tipo II se considera como un *falso negativo*, es decir, el paciente sí tiene la enfermedad dada, pero la prueba médica no la detecta.

Por lo general, se quiere tener un nivel de significancia (probabilidad de tener un falso positivo) pequeño y una potencia de la prueba cercana a 1.

Prueba de hipótesis con nivel de significancia dado: metodología

- Se plantea H_0, H_1, con un nivel de significancia dada.
- Bajo H_0, se establece el estadístico de prueba y la correspondiente región de *rechazo*.
- Usando el valor del estadístico de prueba, se rechaza H_0 si el valor está en el área de rechazo; de otra manera, no se rechaza H_0.

Comentarios

- Formalmente, la hipótesis nula nunca se acepta; más bien, no existe evidencia es su contra.
- Si se rechaza H_0, entonces H_1 es cierta. Una prueba de potencia permite el cálculo de la probabilidad de rechazar H_0 cuando es falsa.

Ejemplo 1

En una muestra de tamaño $n = 100$, el valor de media muestral $\overline{x} = .003$, con desviación estándar muestral$s = .01$.

Se considera la prueba de hipótesis de media

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$

con $\alpha = .05$.

Entonces, el valor del estadístico de prueba, bajo H_0,

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{.003}{.01/10} = 10\frac{.003}{.01} = 3;$$

La distribución de probabilidad del estadístico de prueba[2] $T \sim t_{n-1} = t_{99}$.

[2]En los libros de estadísticas, se considera que el tamaño muestral permite aproximar la distribución t_{99} mediante la normal estándar; usando software estadístico no se considera esta aproximación.

El *área de rechazo* asociada a la prueba de hipótesis corresponde a las colas $t_{99} > t_c$, $t_{99} < -t_c$, donde t_c es el valor crítico de t_{99}, que deja una probabilidad de .025 en cada cola. Su valor se calcula, con R, mediante

```
qt(.025,99,lower.tail = F)
```

```
## [1] 1.984217
```

El valor observado $T = 3$ se encuentra por lo tanto en el área de rechazo; se rechaza H_0 y concluimos que $\mu \neq 0$.

Valor-p y regla de rechazo

En una prueba de hipótesis el valor-p es la probabilidad que la distribución que sigue el estadístico de prueba tiene valores más extremos que el valor observado.

En el Ejemplo 1 de la sección anterior, el *valor-p* asociado al estadístico de prueba se define como

$$\text{valor-}p = \Pr(t_{99} > |T|) = \Pr(t_{99} > |3|) = 0.003415508$$

(se usa el valor absoluto porque el área de rechazo tiene dos colas).

Con R

```
2*pt(3, df = 99, lower.tail = F)
```

```
## [1] 0.003415508
```

Dado un nivel de significancia α, una comparación de áreas bajo la curva, permite enunciar la regla general de rechazo mediante el valor-p.

- Si valor-$p < \alpha$, entonces se rechaza H_0.
- Si valor-$p > \alpha$, entonces no se rechaza H_0.

Prueba de hipótesis e intervalo de confianza

Un intervalo de confianza y una prueba de hipótesis están relacionados de la siguiente manera:

- El área de rechazo de una prueba de hipótesis (a un nivel de significancia $\alpha = .05$) es el *complemento* del intervalo de confianza al 95%.
- El intervalo de confianza coincide con el área de no rechazo de la prueba de hipótesis.

Por lo tanto, si el estadístico de prueba se encuentra fuera del intervalo de confianza, entonces se rechaza H_0.

Si el estadístico de prueba está en el intervalo de confianza, entonces no se rechaza H_0.

Ejemplos con R

Ejemplo 2

Se quiere determinar si el retorno diario del precio promedio del índice FTSE de Londres es más grande, en promedio, que el correspondiente retorno del índice DAX de Frankfurt.

Usarmeos los datos `EuStockMarkets`, en la librería `datasets` cargada por defecto al arranque de R. El dataset está en forma de matriz, por lo que cada columna de la matriz corresponde a una variable

```
data("EuStockMarkets")
head(EuStockMarkets) # visualizar las primeras 6 observaciones
```

```
##          DAX    SMI    CAC   FTSE
## [1,] 1628.75 1678.1 1772.8 2443.6
## [2,] 1613.63 1688.5 1750.5 2460.2
## [3,] 1606.51 1678.6 1718.0 2448.2
## [4,] 1621.04 1684.1 1708.1 2470.4
## [5,] 1618.16 1686.6 1723.1 2484.7
## [6,] 1610.61 1671.6 1714.3 2466.8
```

```
                    # con el nombre de las variables
```

Extraemos las variables de interés, `dax`, `ftse`:

```
dax <- EuStockMarkets[,1]
ftse <- EuStockMarkets[,4]
```

Calculamos los retornos diarios asociados

```
dax.rtn <- diff(log(dax))
ftse.rtn <- diff(log(ftse))
```

Los estadísticos descriptivos

```
summary(dax.rtn)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.0962770 -0.0046854  0.0004726  0.0006520  0.0063552  0.0507601
```

```
summary(ftse.rtn)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -4.140e-02 -4.319e-03  8.021e-05  4.320e-04  5.254e-03  5.440e-02
```

El planteamiento de la prueba de hipótesis es

$$H_0 : ftse.rtn = dax.rtn$$
$$H_1 : ftse.rtn > dax.rtn$$

Usaremos el nivel de significancia $\alpha = .05$.

En R

```
t.test(ftse.rtn, dax.rtn, mu = 0, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  ftse.rtn and dax.rtn
## t = -0.72891, df = 3493.3, p-value = 0.7669
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.0007167649           Inf
## sample estimates:
##    mean of x    mean of y
## 0.0004319851 0.0006520417
```

Comentario En un prueba de hipótesis de diferencia de medias, $\mu_1 - \mu_2$ con R, se consideran por defecto varianzas diferentes, $\sigma_1^2 \neq \sigma_2^2$ con el consecuente ajuste de Welch al estadístico de prueba.

Para probar si las varianzas son diferentes,

```
var.test(ftse.rtn, dax.rtn, ratio = 1)
```

```
##
##  F test to compare two variances
##
## data:  ftse.rtn and dax.rtn
## F = 0.59681, num df = 1858, denom df = 1858, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5449119 0.6536422
## sample estimates:
## ratio of variances
##           0.596806
```

Concluimos que $\sigma_1^2 \neq \sigma_2^2$.

Ejemplo 3 (Limpieza de datos y prueba de potencia)

Usaremos los datos acerca del salario por hora (por género; grupo étnico) de "the Economic Policy Institute's State of Working America Data Library" [**EPI**]

Limpieza de datos

Los datos incluyen el símbolo "$" y esto causa que los datos no se carguen como datos numéricos.

```
library(readr)
epiwages <- read_csv("EPIwages.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Date = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

Solo la fecha se carga como variable numérica.

Para cargar los datos correctamente se debe eliminar el símbolo monetario y declarar las variables de interés como variables numéricas:

```
wage.h <- as.numeric(gsub("\\$", "", epiwages$`Men Average`))
wage.m <- as.numeric(gsub("\\$", "", epiwages$`Women Average`))
```

Comentario sobre el código El símbolo "$" es un carácter especial en R, por lo que para considerarlo como símbolo monetario se usa como prefijo el carácter "\" (carácter "escape", en inglés); a su vez, el carácter "escape" es también un símbolo reservado en R, por lo que se debe usar otro "escape" carácter anten de él. El método `gsub()` cambia el carácter del primer argumento con el carácter en el segundo argumento en la variable declarada como tercer argumento del método. En este caso, se elimina el símbolo de dólar (la función técnicamente sustituye este símbolo con nada) en las variables deseadas. Una vez eliminado el símbolo de dólar, declaramos el resultado como variable numérica usando el método `as.numeric()`.

Prueba de hipótesis

Probaremos la hipótesis que el salario promedio de los varones es mayor al salario promedio de las mujeres.

La hipótesis estadística es

$$H_0 : \mu_H - \mu_M = 0$$
$$H_1 : \mu_H - \mu_M > 0$$

con $\alpha = .05$.

```
t.test(wage.h, wage.m, mu = 0, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wage.h and wage.m
```

```
## t = 13.829, df = 79.633, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.443947      Inf
## sample estimates:
## mean of x mean of y
##  24.34674  18.15804
```

Se rechaza H_0 y concluimos que los varones, en promedio, ganan más que las mujeres.

Prueba de potencia

Para una prueba de potencia, se debe considerar una hipótesis alternativa específica. Dado el resultado, consideramos como H_1 la siguiente

$$H_1 : \mu_H - \mu_M = 4$$

es decir, la diferencia verdadera en salario promedio entre hombres y mujeres es de $4.00 por hora.

La prueba de potencia necesita el tamaño muestral, n; el valor `delta`, la diferencia verdadera en las media, la desviación estándar, el nivel de significancia y la alternativa.

```
power.t.test(n = length(wage.h),
             sd = sqrt(var(wage.h)+var(wage.m)),
             delta = 4, sig.level = .05, alternative = "one.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 46
##           delta = 4
##              sd = 3.035143
##       sig.level = 0.05
##           power = 0.9999981
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

Como desviación estándar consideramos las muestras (salario promedio por género) son independientes, por lo que la varianza de la diferencia es igual a la suma de las varianzas.

La potencia es más de .99, por lo que es muy probable rechazar H_0 cuando es falsa.

Problemas

Problema 1

Un productor afirma que la fuerza de tensión de un cable que fabrica es igual a 84 kilogramos.

Para probar la afirmación, se toman 50 muestras del cable y se prueban en condiciones típicas; el resultado de la prueba es

$$\overline{x} = 86.7, \ \ s = 6.28$$

(unidades en kilogramos).

¿Existe suficiente evidencia que respalde la afirmación del fabricante? Plantea y prueba la hipótesis estadística correspondiente; usa $\alpha = .05$.

Tip Valor del estadístico de prueba

$$T = \frac{86.7 - 84}{6.28/\sqrt{50}} = 3.0401.$$

El valor crítico de la distribución t_{49}, usando R, es

```
qt(.025, 49, lower.tail = F)
```

```
## [1] 2.009575
```

(el área de rechazo son las dos colas de la distribución).

¿Cuál es la conclusión?

Problema 2

Los salarios, por hora, de cierta industria, son $13.50. La firma ACME afirma que sus trabajadores ganan, en promedio, más que la media de la industria.

Para probar la afirmación de la empresa, se toma una muestra de 100 trabajadores de la firma ACME y se encuentra que el promedio muestral de las ganancias es igual a $\overline{x} = 13.30$, con $s = 1.20$ (unidades en dólares por hora).

¿Existe suficiente evidencia para respaldar la afirmación de ACME? Plantea y prueba la hipótesis estadística correspondiente ($\alpha = .05$).

Tip El valor crítico de la distribución t_{99}, usando R, es

```
qt(.05, 99, lower.tail = F)
```

```
## [1] 1.660391
```

(el área de rechazo es la cola derecha de la distribución).

¿Cuál es la conclusión?

Problema 3

¿Qué pasa si en el Problema 2 la desviación estándar $s = 1.51195$?

Problema 4

Pruebas las hipótesis estadísticas en los Problemas 1, 2 usando el valor-p.

Tip Supón que el valor del estadístico de prueba $T = 3.0401$, con distribución t_{49}, en un prueba de hipótesis de *dos colas*; el valor-p asociado es

```
2*pt(3.0401,49, lower.tail = F)
```

```
## [1] 0.003789543
```

El modelo de regresión lineal clásico

Se considera el modelo de regresión lineal múltiple

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Por simplicidad de escritura denotaremos en negritas, $\mathbf{x}$, el conjunto de las variables explicativas x_1, ..., x_k.

Supuestos del modelo de regresión lineal clásico

La *estimación puntual* de los parámetros es una aplicación de optimización multivariable.

Por otro lado, propiedades estadísticas deseables necesitan que la muestra $(y_i, x_{1,i}, \ldots, x_{k,i})$, $i = 1, \ldots, k$ cumpla con ciertos supuestos que listamos a continuación.

1. El término de error tiene media cero: $E(u) = 0$. Esta propiedad es una consecuencia de la estimación por Mínimos Cuadrados Ordinarios (MCO).
2. **Homocedasticidad** La varianza del término de error, condicionada sobre los valores de las variables explicativas, es constante (y finita): $var(u|x_1, \ldots, x_k) = \sigma^2$.
3. **Autocorrelación serial** (Muestreo aleatorio) Los errores son linealmente independientes entre si: $cov(u_t, u_s) = 0$.
4. Los errores son *media independiente* de las variables explicativas:

   $$E[u|x_1, \ldots, x_k] = 0.$$

   En las aplicaciones, se puede usar el supuesto más débil de no correlación: $cov(u, x) = 0$. Este supuesto sera crucial en el análisis *ceteris paribus* (véase más adelante).

5. **Colinealidad perfecta** No existe ninguna dependencia lineal entre las variables explicativas $x_1, \ldots, x_k$. Este supuesto sólo aplica a regresiones lineales múltiples.
6. **Normalidad de los errores** $u \sim N(0, \sigma^2)$. El supuesto de normalidad se usará en la inferencia de las estimaciones por MCO (significancia estadística; cálculo de intervalos de confianza, entre otros).

Propiedades de los estimadores por MCO

Bajo los primeros cinco supuestos del modelo de regresión lineal clásico, los estimadores por MCO son los Mejores Estimadores Lineales Insesgados (MELI; en la literatura anglosajona, BLUE: Best Linear Unbiased Estimators).

- Son *estimadores* porque están determinados por una fórmula que involucra la variable aleatoria y.
- Los estimadores son *lineales* porque dependen de manera lineal de la variable aleatoria y.
- Los estimadores son *insesgados*, porque $E[\widehat{\beta}_i | x_1, \ldots, x_k] = \beta_i$: dados los valores de las variables explicativas, los valores de los estimadores son iguales, en promedio, a los correspondientes valores poblacionales.
- Son los *mejores* porque se puede probar que los estimadores son también *consistentes* y *eficientes*.
  - La *consistencia* (o límite en probabilidad) se define como sigue

    $$\lim_{T \to \infty} \Pr\left[|\widehat{\beta}_i - \beta_i| > \delta\right] = 0, \forall \delta > 0.$$

    El significado es que, si el tamaño muestral es lo suficientemente grande, entonces el valor estimado $\widehat{\beta}_i$ se acercará cada vez más al valor poblacional β_i. Para la consistencia es suficiente que $E[u] = 0$, $E[xu] = 0$.
  - La *eficiencia* significa que los estimadores por MCO tienen la menor varianza entre los estimadores insesgados de β_i.

Función de regresión poblacional

Considera el modelo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

Si los errores son media independiente de las variables explicativas,

$$E[u|\mathbf{x}] = 0,$$

entonces

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

que se conoce como la *función de regresión poblacional*.

Estimación: Modelo de regresión lineal simple

El caso de un modelo de regresión lineal simple consiste en la estimación de

$$y = \beta_0 + \beta_1 x + u.$$

Si (y_i, x_i), $i = 1 \ldots, n$ una muestra de observaciones tomada de una población. Sea

$$u_i = y_i - (\beta_0 + \beta_1 x_i)$$

la diferencia entre el valor observado de la variable y y el valor estimado por el modelo lineal; el conjunto (u_i), $i = 1, \ldots, n$ se conocen como los *residuales* (o residuos).

El método de MCO se basa en la *minimización* de los cuadrados de los residuales considerados como función de los parámetros incógnitos β_0, β_1:

$$\min SRC(\beta_0, \beta_1),$$

donde

$$SRC(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2 .$$

(SRC: suma de los residuales cuadrados)

El mínimo de la función $SRC(\beta_0, \beta_1)$ está determinado por el punto crítico de la función SRC:

$$0 = \frac{\partial SRC}{\partial \beta_0} = (-2) \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]$$
$$0 = \frac{\partial SRC}{\partial \beta_1} = (-2) \sum_{i=1}^{n} x_i \left[y_i - (\beta_0 + \beta_1 x_i) \right]$$

Ecuaciones normales

Reacomodando las variables en la expresión del punto crítico de la función $SRC()$, en el sistema lineal de dos ecuaciones en dos incógnitas (*ecuaciones normales*):

$$n\beta_0 + \left(\sum_{i=1}^{n} x_i \right) \beta_1 = \sum_{i=1}^{n} y_i$$
$$\left(\sum_{i=1}^{n} x_i \right) \beta_0 + \left(\sum_{i=1}^{n} x_i^2 \right) \beta_1 = \sum_{i=1}^{n} x_i y_i$$

El sistema tiene solución si, y sólo si, el determinante de la matriz de coeficientes es diferente de cero:

$$0 \neq \det \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2$$

El determinante de la matrix de coeficientes es un múltiplo de varianza muestral de x, por lo que es necesario que la variable $x \neq const$ para que el sistema tenga solución.

Bajo esta condición, el vector solución $\widehat{\beta}_0, \widehat{\beta}_1)$ es igual a

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

El término $\overline{x}$ (resp., $\overline{y}$) es la *media muestral* de la variable x (resp., y).

Comentarios

- El vector solución $(\widehat{\beta}_0, \widehat{\beta}_1,)$ es un *estimador* del vector poblacional (β_0, β_1).
- La expresión para el estimador $\widehat{\beta}_0$ indica que el punto con coordenadas $(\overline{x}, \overline{y})$ (media muestral de x, y, respectivamente) pertenece a la recta de regresión estimada

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

- La expresión para el estimador $\widehat{\beta}_1$ contiene al numerador el término de *covarianza muestral* de las variables x, y.
- El sistema de ecuaciones lineales usado para la estimación de β_0, β_1 son equivalentes a las correspondientes ecuaciones muestrales asociadas (método de momentos)

$$E[u] = 0$$
$$E[xu] = 0$$

  Las dos ecuaciones en conjunto dan $cov(x, y) = 0$. Esto es, usando el método de MCO, la media muestral de los errores y la covarianza muestral entre x, u son iguales a cero.
- El método de MCO se extiende sin dificultad al caso de regresión lineal múltiple

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Estimación: Modelo de regresión lineal múltiple

Consideramos, por simplicidad de exposición, el caso de dos variables (*ecuación poblacional*):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

La estimación del modelo se realiza mediante una *muestra aleatoria* dada por una base de datos

$$(x_{i,1}, x_{i,2}, y_i), \quad i = 1, \ldots, n,$$

donde n es el tamaño muestral.

Ecuaciones normales

Como en el caso de regresión lineal simple, se define la *suma de residuos cuadrados*

$$SRC(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2$$

Entonces, las ecuaciones normales se determinan *minimizando la suma de residuos cuadrados*

$$\min_{\beta_0, \beta_1, \beta_2} SRC(\beta_0, \beta_1, \beta_2).$$

Las *condiciones de primer orden*, que dan lugar a la ecuación normal, se dan igualando a cero el gradiente de $SRC(\beta_0, \beta_1, \beta_2)$:

$$\begin{aligned}
0 = \frac{\partial(SRC)}{\partial \beta_0} &= (-2) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2}) \\
0 = \frac{\partial(SRC)}{\partial \beta_1} &= (-2) \sum_{i=1}^{n} [(y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2})(x_{i,1})] \\
0 = \frac{\partial(SRC)}{\partial \beta_2} &= (-2) \sum_{i=1}^{n} [(y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2})(x_{i,2})]
\end{aligned}$$

Simplificando, se obtienen las ecuaciones normales, en forma vectorial

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,2} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,1}^2 & \sum_{i=1}^n x_{i,1} x_{i,2} \\ \sum_{i=1}^n x_{i,2} & \sum_{i=1}^n x_{i,1} x_{i,2} & \sum_{i=1}^n x_{i,1}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,1} y_i \\ \sum_{i=1}^n x_{i,2} y_i \end{bmatrix}$$

Colinealidad perfecta

La determinación del vector solución pasa por el cálculo del determinante de la matriz de coeficientes.

Se dice que el modelo tiene *colinealidad perfecta* si el determinante de la matriz de coeficientes es igual a cero. Esto es equivalente a afirmar que las variables explicativas son una combinación lineal.

Si las variables no tienen colinealidad perfecta, entonces los estimadores $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ de los respectivos parámetros poblaciones β_0, β_1, β_2, se determinan usando las herramientas usuales de Álgebra Lineal (Método de eliminación de Gauss-Jordan es particularmente efectivo numéricamente).

Los estimadores $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ son los *estimadores del modelo de regresión por Mínimos Cuadrados Ordinarios (MCO).*

Valores ajustados y residuales

Una vez estimados el intercepto y las pendientes $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, respectivamente, podemos determinar

- el *valor ajustado* (valor predicho, o valor ajustado por MCO) de la variable dependiente

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \widehat{\beta}_2 x_{i,2};$$

- el *residual* (o residual de MCO) de la observación i

$$\widehat{u}_i = y_i - \widehat{y}_i.$$

  Hay un residual para cada observación.

- El *signo* de cada residual es de interés:
  - si $\widehat{u}_i > 0$, entonces significa que $y_i > \widehat{y}_i$, es decir, el modelo de regresión *subestima* el valor observado de la variable dependiente;
  - si $\widehat{u}_i < 0$, entonces significa que $y_i < \widehat{y}_i$, es decir, el modelo de regresión *sobreeestima* el valor observado de la variable dependiente;
- el promedio muestral de los residuales es cero:

$$\sum_{i=1}^{n} \widehat{u}_i = 0.$$

- la covarianza muestral entre cada una de las variables independientes y los residuales es cero. Por lo tanto, la covarianza muestral entre los valores ajustados por MCO y los residuales de MCO es también cero.
- el punto $(\overline{x}_1, \overline{x}_2, \overline{y})$ siempre se encuentra sobre la línea de regresión de MCO:

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}_1 + \widehat{\beta}_2 \overline{x}_2.$$

Análisis *ceteris paribus*

Cuando se considera una función, por ejemplo, de dos variables

$$z = f(x_1, x_2)$$

el análisis *ceteris paribus* consiste en determinar el efecto del *cambio* de una variable independiente sobre la variable dependiente, *manteniendo los valores de la otra variable independiente constantes*; matemáticamente, esto corresponde al cálculo de la *función marginal*

$$\frac{\partial z}{\partial x_i} = \frac{\partial f}{\partial x_i}, \ i = 1, 2.$$

Cuando se estima un modelo de regresión lineal (simple o múltiple), tenemos también que considerar el *efecto de los factores no observables*, modelados por el error u.

Si consideramos el modelo poblacional

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Para determinar como el cambio, *ceteris paribus*, de x_1, por ejemplo, afecta la variable dependiente, calculamos

$$\Delta y = \beta_1 \Delta x_1 \;\Rightarrow\; \boxed{\beta_1 = \frac{\Delta y}{\Delta x_1}}$$

siempre y cuando $\Delta u = 0$, es decir, siempre y cuando los factores no observados no cambian cuando cambia x_1. Matemáticamente,

$$E[u|x_1, x_2] = 0.$$

Analizaremos más en detalle el impacto de los errores en la estimación de las pendientes β_j en la siguiente sección *Efectos parciales y estimación por MCO*.

Variables exógenas

Si los errores son *media independiente* de las variable explicativas, $E[u|x_1, x_2] = 0$,entonces se dice que las variables (independientes) son *exógenas*.

Comentario

El análisis *ceteris paribus* es correcto cuando todas las variables independientes del modelo son exógenas.

Independencia en media y no correlación

En las aplicaciones, se puede reemplazar la propiedad de independencia en media con la más débil de *no correlación* entre los errores y las variables independientes:

$$cov(u, x_j) = 0,\; j = 1, \ldots, k$$

Efectos parciales y estimación por MCO

En esta parte, ampliaremos la discusión acerca del efecto *ceteris paribus* en el modelo de regresión múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Para mantener el análisis más concreto, a continuación nos enfocaremos en β_1; claro está, un análisis semejante si hace para las demás pendientes.

La estimación de la pendiente β_1 se puede expresar mediante

$$\boxed{\widehat{\beta}_1 = \frac{\sum_{i=1}^n y_i \widehat{r}_{i,1}}{\sum_{i=1}^n \widehat{r}_{i,1}^2}},$$

donde $\widehat{r}_{i,1}$ son los residuales de la regresión

$$x_1 = \gamma_0 + \gamma_1 x_2 + v,$$

usando la muestra dada.

El significado de la fórmula de arriba es que la estimación de la pendiente β_1 sólo se considera el efecto de la variable x_1 sobre y porque el residual $r_{i,1}$ tiene correlación cero con la variable x_2. Es decir,

$$\widehat{\beta}_1 = \frac{\Delta y}{\Delta x_1},$$

descontando los efectos de las demás variables (análisis *ceteris paribus*). Es importante que en el análisis ceteris paribus los errores sean media independiente de las variables explicativas.

La prueba de la fórmula enmarcada arriba no es complicada.

Se escribe

$$x_1 = \widehat{x}_1 + \widehat{r}_{i,1}$$

donde $\widehat{x}_1$ son los valores ajustados de x_1 de la regresión sobre x_2 y $\widehat{r}_{1,1}$ son los correspondientes residuos. Entonces,

$$\widehat{x}_1 = \widehat{\gamma}_0 + \widehat{\gamma}_1 x_2, \quad \widehat{r}_{i,1} = x_1 - \widehat{x}_1$$

Las siguientes expresiones se usarán a menudo en las estimaciones de las pendientes:

$$\sum_{i=1}^n x_2 \widehat{r}_{i,1} = 0, \quad \sum_{i=1}^n \widehat{x}_1 \widehat{r}_{i,1} = 0, \quad \sum_{i=1}^n x_1 \widehat{r}_{i,1} = \sum_{i=1}^n \widehat{r}_{i,1}^2, \quad \sum_{i=1}^n \widehat{r}_{i,1} = 0$$

Ahora, usando valores ajustados en la primera ecuación de la condición de primer orden

$$\begin{aligned}
0 &= \sum_{i=1}^n x_1(\widehat{y} - \widehat{\beta}_0 - \widehat{\beta}_1 x_1 - \widehat{\beta}_2 x_2) \\
&= \sum_{i=1}^n (\widehat{x}_1 + \widehat{r}_{i,1}) (\widehat{y} - \widehat{\beta}_0 - \widehat{\beta}_1 x_1 - \widehat{\beta}_2 x_2) \\
&= \sum_{i=1}^n \widehat{x}_1(\widehat{y} - \widehat{\beta}_0 - \widehat{\beta}_1 x_1 - \widehat{\beta}_2 x_2) + \sum_{i=1}^n \widehat{r}_{i,1}(\widehat{y} - \widehat{\beta}_0 - \widehat{\beta}_1 x_1 - \widehat{\beta}_2 x_2) \\
&= \sum_{i=1}^n (\widehat{\gamma}_0 + \widehat{\gamma}_1 x_2) u_i + \sum_{i=1}^n \widehat{r}_{i,1}(\widehat{y} - \widehat{\beta}_0 - \widehat{\beta}_1 x_1) \\
&= \sum_{i=1}^n \widehat{r}_{i,1} \left[\widehat{y} - \widehat{\beta}_0 - \widehat{\beta}_1(\widehat{x}_1 + \widehat{r}_{i,1})\right] \\
&= \sum_{i=1}^n \widehat{r}_{i,1}(\widehat{y} - \widehat{\beta}_1 \widehat{r}_{i,1}).
\end{aligned}$$

Simplificando se obtiene

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n y_i \widehat{r}_{i,1}}{\sum_{i=1}^n \widehat{r}_{i,1}^2},$$

(Véase también la Apéndice 3A, p. 114, del libro de Wooldridge [**W**].)

Insesgamiento de los estimadores por MCO

Considerando que

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

se puede reescribir la estimación de pendiente

$$\boxed{\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n \widehat{r}_{i,1} u_i}{\sum_{i=1}^n \widehat{r}_{i,1}^2}}.$$

Entonces,

$$E[\widehat{r}_{i,1}|x_1, x_2] = r_{i,1}$$

porque $\widehat{r}_{i,1}$ es función solo de las variables independientes.

Por otro lado, los errores $E[u|x_1, x_2] = 0$, por lo que

$$\begin{aligned}
E[\widehat{\beta}_1|x_1, x_2] &= \beta_1 + E[(\sum_{i=1}^n \widehat{r}_{i,1} u_i)/(\sum_{i=1}^n \widehat{r}_{i,1}^2)] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n \widehat{r}_{i,1}^2} \sum_{i=1}^n E[\widehat{r}_{i,1} u_i|x_1, x_2] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n \widehat{r}_{i,1}^2} \sum_{i=1}^n \widehat{r}_{i,1} E[u_i|x_1, x_2] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n r_{i,1}^2} \sum_{i=1}^n r_{i,1} \cdot 0 \\
&= \beta_1.
\end{aligned}$$

Es decir, $\widehat{\beta}_1$ es un estimador insesgado del parámetro poblacional β_1.

Homocedasticidad: Varianza de los estimadores

Seguimos considerando, por simplicidad de exposición, el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

con dos variables explicativas.

Homocedasticidad

La varianza de los errores es constante:

$$var(u|x_1, x_2) = \sigma^2$$

Comentario

Considerando que $E[u|x_1, x_2] = 0$, entonces[3]

$$var(u|x_1, x_2) = E[u^2|x_1, x_2]$$

Usando la expresión del estimador de pendiente

$$\widehat{\beta}_j = \beta_j + \frac{\sum_{i=1}^n \widehat{r}_{i,j} u_i}{\sum_{i=1}^n \widehat{r}_{i,j}},\ j = 1, 2.$$

y considerando que $\widehat{r}_{i,j}$, condicionada sobre x_1, x_2, no son aleatorias, $E[\widehat{r}_{i,j}|x_1, x_2] = \widehat{r}_{i,j}$, entonces

$$var(\widehat{\beta}_{i,j}|x_1, x_2) = \left(\frac{1}{\sum_{i=1}^n \widehat{r}_{i,j}^2}\right)^2 \sum_{i=1}^n \widehat{r}_{i,j}^2 var(u|x_1, x_2) = \left(\frac{1}{\sum_{i=1}^n \widehat{r}_{i,j}^2}\right)^2 \sum_{i=1}^n \widehat{r}_{i,j}^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n \widehat{r}_{i,j}^2}.$$

Factor de inflación de la varianza

Por simplicidad se discute el caso $j = 1$.

Considerando que $SRC_1 = \sum_{i=1}^n \widehat{r}_{i,1}^2$, entonces $SRC_1 = (1 - R_1^2)STC_1$, donde

$$STC_1 = \sum_{i=1}^n (x_{i,1} - \overline{x}_1)^2;$$

y R_1^2 es el valor de ajuste de la regresión de la variable x_1 sobre las demás variables.

Por lo tanto,

$$var(\widehat{\beta}_1|x_1, x_2) = \frac{\sigma^2}{STC_1(1 - R_1^2)}.$$

Se define

$$VIF = \frac{1}{1 - R_1^2}$$

el *factor de inflación de varianza* (Variance Inflation Factor, en inglés).

Estimación con R del modelo de regresión

Usaremos el dataset `hprice.xls` que contiene el precio de casas y otras variables.

[3]En general,

$$var(u|x_1, x_2) = E[u^2|x_1, x_2] - \left(E[u|x_1, x_2]\right)^2.$$

Cargar archivos Excel; nombrar variables

El dataset no tiene rotuladas las variables, por lo que se mostrará como se pueden asignar los nombres a las variables en un dataset.

Para cargar archivos de Excel, se necesita la librería `readxl`

```
library(readxl)
house <- read_excel("hprice.xls")
names(house) <-
  c("price", "assess",
                  "bedrooms", "lotsize", "sqft", "colonial",
                  "lprice", "lassess", "llotsize", "lsqrft")
```

Se asigna el nombre a las variables (columnas) del dataset usando el método `names()` que toma sus valores del vector a la derecha.

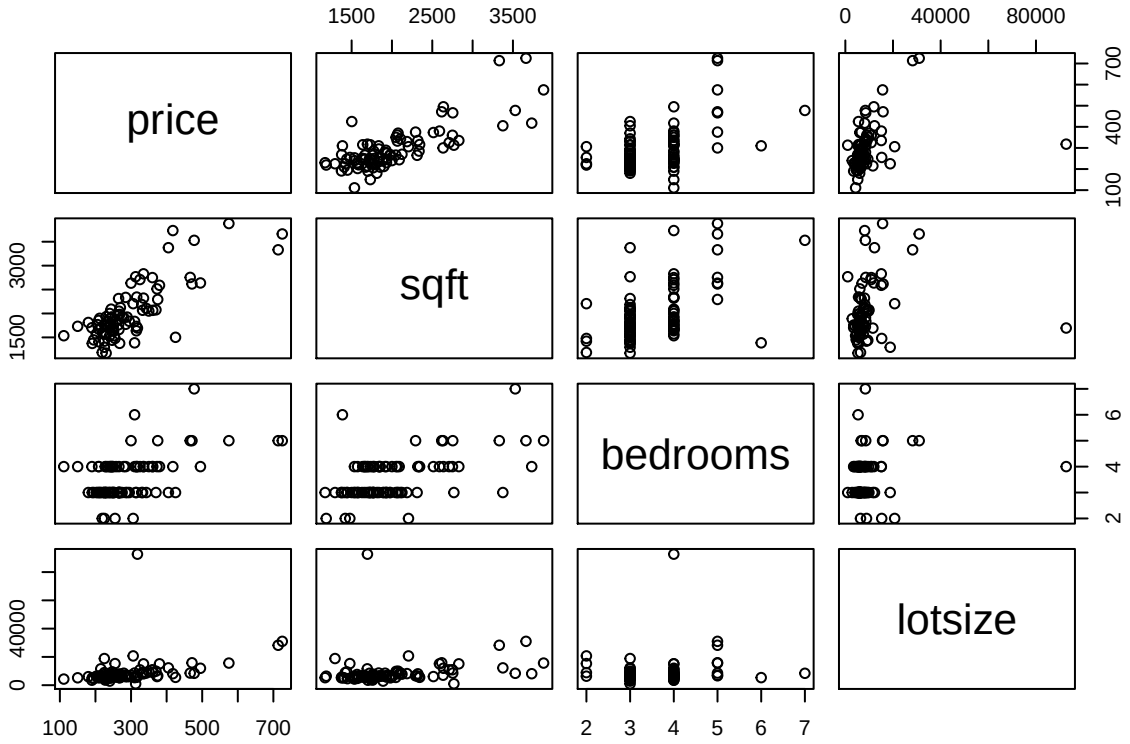
El precio de las casas es en miles de dólares; *assess* es el valor estimado de la casa por un profesional; *bedrooms* es el número de cuartos en la casa; *lotsize* es el tamaño del lote; $sqft$ es el tamaño de la casa (construcción); *colonial* es una variable "dummy" (cualitativa) que vale 1 si la casa tiene ciertas características hedónicas; las últimas cuatro variables son los logaritmos de *price*, *assess*, *lotsize* y $sqft$.

Visualización de pares de variables

Puede ser conveniente visualizar gráficamente la relación (directa; inversa) entre pares de variables antes de estimar un modelo de regresión lineal. Usamos la función `pairs()`. La función produce un scatterplot de los pares de variables declaradas.

Por ejemplo, nos interesa conocer la relación que existe entre el precio de una casa (*price*), su tamaño ($sqft$), el número de cuartos (*bedrms*) y el tamaño del lote (*lotsize*):

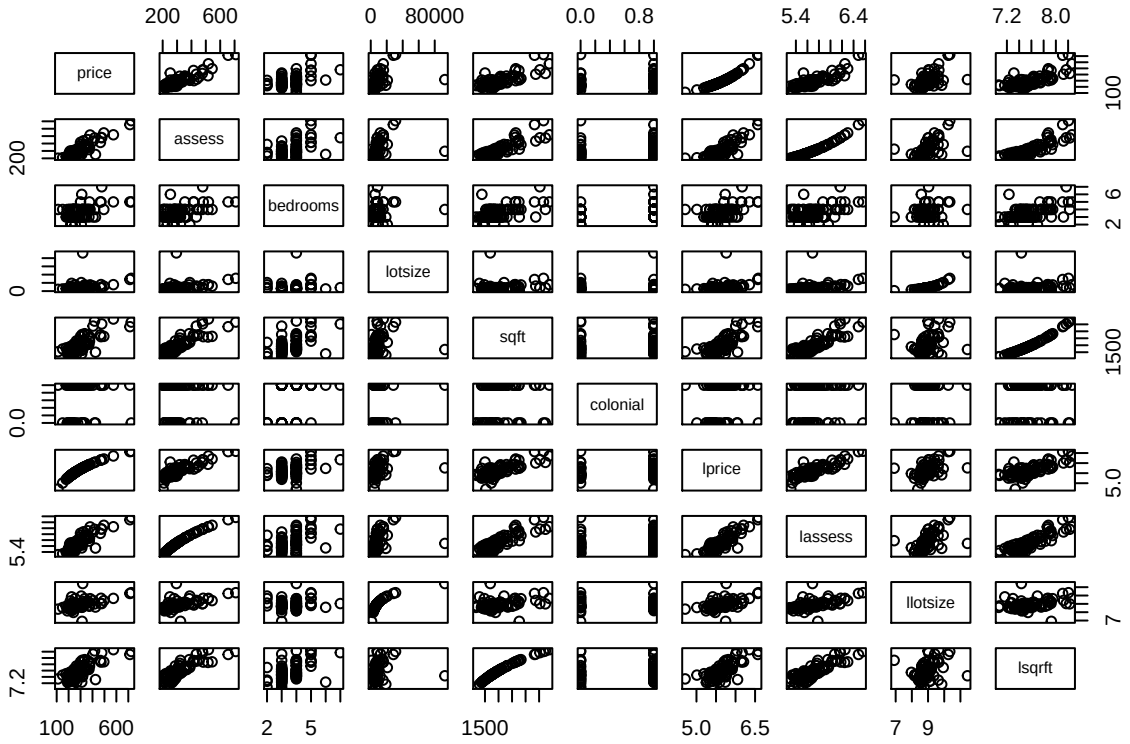
```
pairs(~ price+sqft+bedrooms+lotsize, data = house)
```



Se observa que existe una relación lineal entre *price* y *sqft*; *price* y *lotsize*.

También se puede usar la función `plot()` sobre todo el dataset.

```
plot(house)
```



Estimación del modelo

La estimación del modelo se realiza usando la función `lm()`.

Se estimará el modelo econométrico

$$price = \beta_0 + \beta_1 sqft + \beta_2 bedrooms + \beta_3 lotsize + u.$$

```
price.fit <- lm(price ~ sqft +
                  bedrooms + lotsize, data = house)
```

El resultado se la función se asigna al objeto `price.fit`. Se puede desplegar el resumen usando la función `summary()`

```
summary(price.fit)
```

```
##
## Call:
## lm(formula = price ~ sqft + bedrooms + lotsize, data = house)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -120.402  -37.762   -6.738   31.810  209.174
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.280e+01  2.958e+01  -0.771  0.44290
## sqft         1.234e-01  1.330e-02   9.283 1.77e-14 ***
## bedrooms     1.398e+01  9.034e+00   1.547  0.12563
## lotsize      2.044e-03  6.444e-04   3.173  0.00212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.98 on 83 degrees of freedom
## Multiple R-squared:  0.6747, Adjusted R-squared:  0.6629
## F-statistic: 57.37 on 3 and 83 DF,  p-value: < 2.2e-16
```

Estimación sin intercepto

Si se necesita un modelo sin intercepto, entonces se le añade "-1" a la declaración del modelo:

```
price.fit0 <- lm(price ~ sqft +
                   bedrooms + lotsize -1 , data = house)
```

Regresión sobre todas las variables

A veces se necesita regresar la variable y sobre todas las demás variables. En lugar de escribir todas las variables explicitamente, se puede usar el siguiente atajo:

```
price.fit1 <- lm(price ~ ., data = house)
```

Estimación puntual

Se quiere estimar el precio de una casa cuando $sqft = 3000$, $bedrooms = 4$, $lotsize = 4000$:

```
c <- c(1,3000,4,4000)
sum(price.fit$coefficients*c)
```

```
## [1] 411.5967
```

El precio estimado es de 411,596 dólares.

Comentarios sobre el código Se define primeramente un vector con los valores deseados; se incluye el valor 1 para añadir el valor estimado del intercepto. La segunda instrucción calcula el producto punto (valor estimado del precio) del vector de los parámetros estimados del modelo por los valores declarados de las variables.

Intervalo de confianza para un valor estimado

Se quiere determinar un intervalo de confianza al 99% para el valor estimado de *price* cuando $sqft = 3000$, $bedrooms = 4$, $lotsize = 4000$:

```
predict(price.fit, newdata =
          data.frame(sqft = 3000,
              lotsize=4000, bedrooms=4),
        interval = "confidence", level = .99)
```

```
##        fit      lwr      upr
## 1 411.5967 374.2917 448.9017
```

(Por defecto, el intervalo de confianza es al 95%.)

Intervalo de predicción para un valor estimado

En este caso, se estima para un individuo no necesariamente parte de la muestra. El intervalo incluye también la estimación de la varianza de la regresión:

```
predict(price.fit, newdata =
          data.frame(sqft = 3000,
              lotsize=4000, bedrooms=4),
        interval = "prediction", level = .99)
```

```
##        fit      lwr      upr
## 1 411.5967 249.1272 574.0663
```

Puntos influyentes y leverage

Como referencia se puede usar la Sección "Influential Observations", Cap. 14, del libro [**IS**].

Punto influyente

La influencia de una observación se puede pensar en cómo los valores predicho de las otras obsercaciones cambiarían si esta observación no se incluyera en la estimación.

La *distancia de Cook* D es una medida para determinar la influencia de una observación y es proporcional a la suma de los cuadrados de las predicciones con todas las observaciones y las predicciones obtenidas eliminando esta observación.

Una regla empírica indica que una distancia de Cook $D > 1$ es evidencia que la observación en cuestión tiene mucha influencia.

Leverage

El leverage de una observación es la capacidad de esta observación de afectar el valor de tendencia de la variable dependiente.

Una observación con un valor igual al promedio de la variable dependiente no tiene leverage porque esta observación se encuentra sobre la recta de regresión.

Por otro lado, una observación que es un outlier, puede tener mucha influencia sobre la pendiente de la recta de regresión.

El leverage h_{ii} es el i-ésimo elemento de la "hat matrix" $X(X^TX)^{-1}X^T$, donde X es la *matriz de diseño* (matriz $T \times k$; T tamaño muestral, k número de variables independientes).

El leverage cumple

$$0 \leq h_{ii} \leq 1.$$

Si $e_i = y_i - \widehat{y}_i$, residual de la i-ésima observación entonces, bajo homocedasticidad,

$$var(e_i) = (1 - h_{ii})\sigma^2.$$

Validación "in sample", "out of sample"

Para determinar cuánto es buena la estimación de la regresión se puede realizar un proceso de *validación* que consiste en dividir la muestra en dos[4]: una parte (se puede usar el 80% de los datos) se usa para la estimación del modelo de regresión (dentro de muestra; *in sample*, en inglés) y el 20% remanente se compara con la predicción del modelo estimado (fuera de muestra; *out of sample*).

Se dividirá la muestra de manera aleatoria (sin remplazo). El uso de la función `set.seed()` es para obtener un resultado que se pueda replicar por el lector.

```
set.seed(123)
myvariables <- c("price", "bedrooms",
                 "lotsize", "sqft")
house.red <- house[myvariables]
EntrenRows <- sample(1:nrow(house.red),
                     .8*nrow(house.red),
                     replace = FALSE)
Entren <- house.red[EntrenRows,]
Validac <- house.red[-EntrenRows, ]
modelo.lineal.Entren <-
  lm(formula = price ~ ., data = Entren)
pricePredic <-
```

[4]Los valores que se usan aquí, 80% para la estimación "in sample" y 20% para el pronóstico "out of sample" son medidas empíricas comunemente usadas; se pueden encontrar otros valores en la literatura.

```
  predict(modelo.lineal.Entren, Validac)
actual.predic <-
  data.frame(cbind(actual = Validac$price,
                   predicted = pricePredic))
```

Exactitud: matriz de correlación

Después, se puede calcular la correlación entre valores predicho (in sample) con los valores out of sample:

```
corr.accuracy <- cor(actual.predic)
corr.accuracy
```

```
##               actual predicted
## actual    1.0000000 0.8602343
## predicted 0.8602343 1.0000000
```

Los valores in-sample y out of sample tienen una correlación bastante buena.

Exactitud: Min-Max

El valor *min-max* considera el promedio entre el valor mínimo y el máximo de la predicción.

El valor $min\text{-}max = 1$ indica una completa exactitud. Valores empíricos $min\text{-}max > .9$ indican una excelente exactitud.

```
min.max.accuracy <-
  mean(apply(actual.predic,
             1, min)/apply(actual.predic, 1, max))
min.max.accuracy
```

```
## [1] 0.872623
```

En este caso, la exactitud es razonablemente buena.

Exactitud: Mean Absolute Error (MAE)

El error absoluto promedio (MAE) se define

$$MAE = \frac{1}{n}\sum_{i=1}^{T}|A_i - F_i|$$

donde A_i es el valor actual (observado) de *price* en la muestra; F_i es el valor pronosticado usando la estimación de la regresión, T es el tamaño muestral.

```
mae <- mean(abs((actual.predic$predicted -
                   actual.predic$actual)))
mae
```

```
## [1] 41.41314
```

Exactitud: Mean Absolute Percentage Error (MAPE)

El error porcentual absoluto promedio (MAPE) se define

$$MAPE = \frac{1}{n}\sum_{i=1}^{T}\left|\frac{A_i - F_i}{A_i}\right|.$$

- Se puede interpretar el valor de $MAPE$ como un porcentaje promedio de error.
- En ciertas situaciones el valor de $MAPE$ puede tener inestabilidad numérica. Por ejemplo, si $A_i \approx 10^{-3}$ y $|F_i - A_i| \approx 10^{-1}$, entonces

$$\left|\frac{A_i - F_i}{A_i}\right| \approx 10^2$$

- En un pronóstico exacto, se obtiene $MAPE = 0$.

```
mape <- mean(
  abs((actual.predic$predicted -
         actual.predic$actual))/actual.predic$actual)
mape
```

```
## [1] 0.1386397
```

Pruebas de diagnóstico

Un diagnóstico de los supuestos del Modelo de Regresión Lineal Clásico incluye

- Homocedasticidad
- Autocorrelación serial
- Normalidad de los errores
- prueba RESET de especificación del modelo

Además, se considerará también la prueba de multicolinealidad, aún si el efecto multicolineal es más un problema de datos que de estimación del modelo.

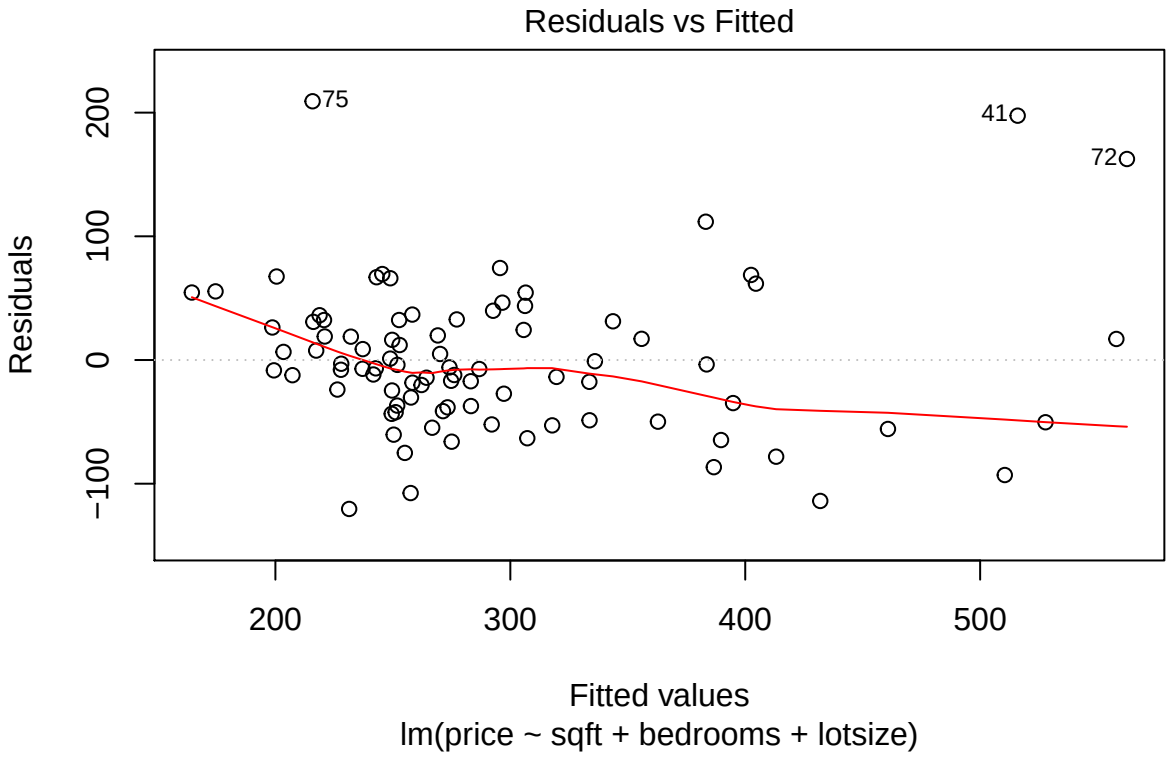
Diagnóstico gráfico

Se puede usar la función `plot()` de un objeto tipo `lm` para el diagnóstico gráfico de un modelo de regresión lineal estimado. Se le pasa un parámetro numérico, de 1 a 6, para el diagnóstico.

Usaremos el modelo de precio de inmuebles estimado arriba.

Prueba de nolinealidad en el modelo Si la línea de los residuales es horizontal, no hay evidencia de nolinealidades en el modelo.

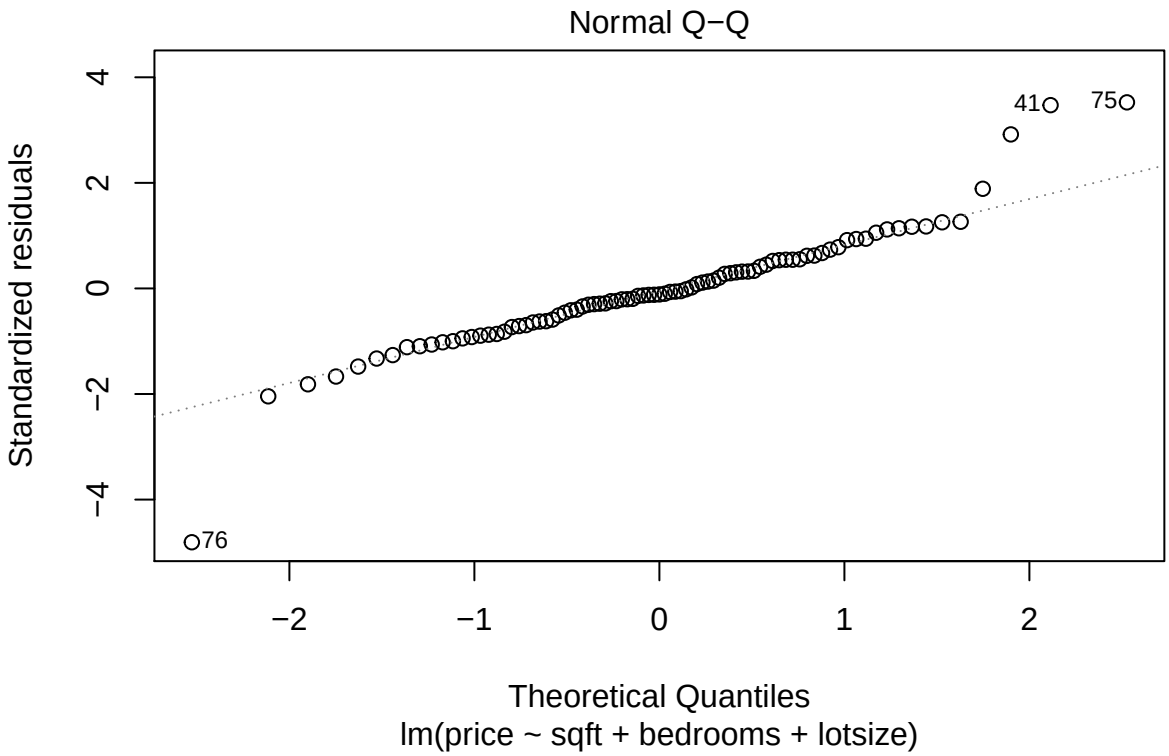
```
plot(price.fit,1)
```



En este caso, la evidencia de no linealidades no parece ser fuerte. Se realizará una prueba `RESET` más adelante para probar la adecuación funcional del modelo.

Prueba de normalidad de los errores La prueba gráfica es un *QQ plot*. Si los errores son normales, los residuales se deben alinear a la recta que pasa por el origen.

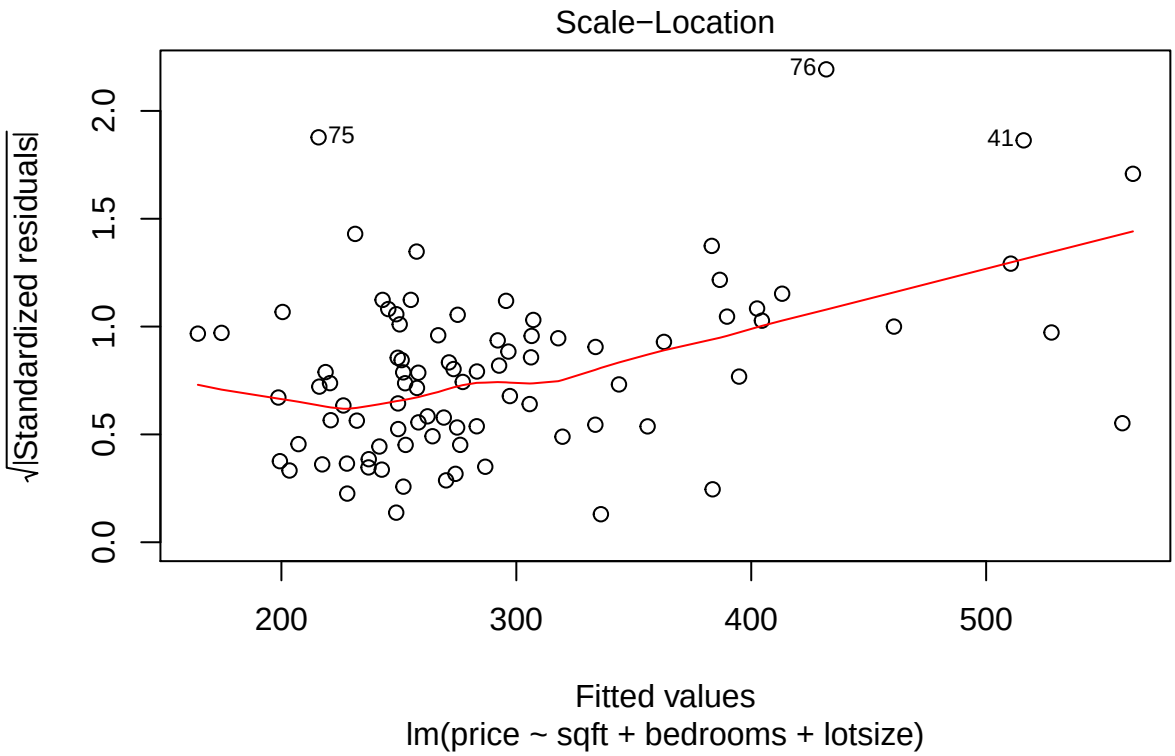
```
plot(price.fit,2)
```



Por lo general, el ajuste es bueno, aún si hay cuatro observaciones, de las cuales las observaciones 41 y 75 son las más extremas. Estas observaciones podrían tener mucha influencia/leverage. Lo veremos más adelante. Más adelante se llevará a cabo una prueba de hipótesis de normalidad de los residuales.

Prueba de heterocedasticidad Idealmente, la linea de los residuales debería ser horizontal, con una dispersión de los valores observados uniforme.

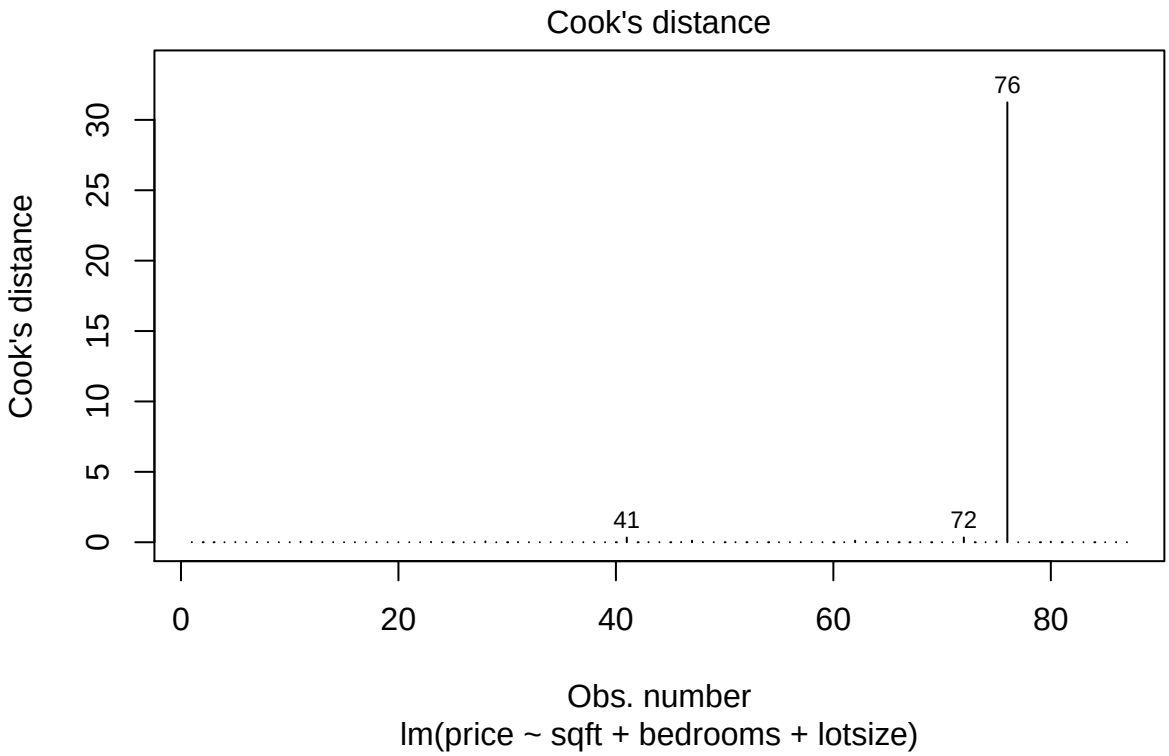
```
plot(price.fit,3)
```



Se observa una acumulación de observaciones cerca de valores ajustados más pequeños; puede ser una indicación de heterocedasticidad. Obtendremos un resultado más preciso con una prueba de Breusch-Pagan de heterocedasticidad más adelante.

Observaciones influyentes En este caso, se plotean las observaciones en función de su distancia de Cook. Las observaciones influyentes se indican explicitamente.

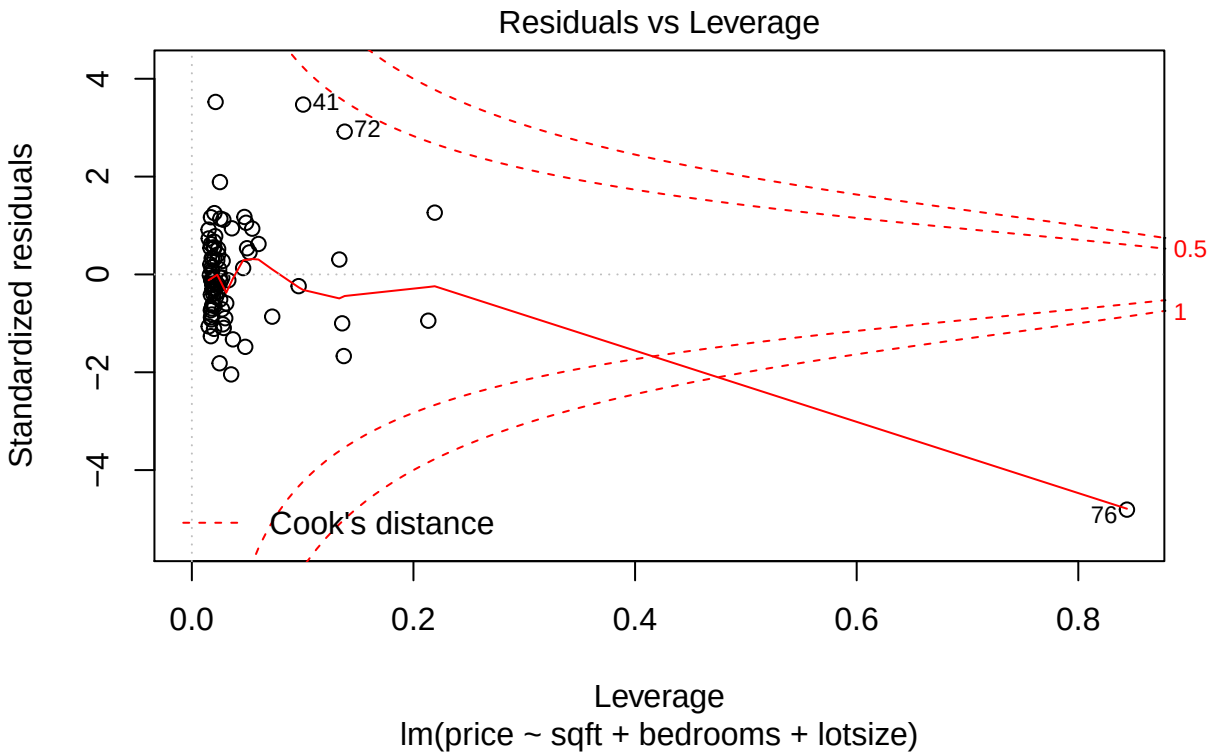
```
plot(price.fit, 4)
```



La observación 76 es influyente. Las observaciones 41, 72 son casos límites.

Leverage Observaciones influyentes caen más allá de las lineas punteadas en rojo (distancia de Cook mayor a .5; mayor a 1).

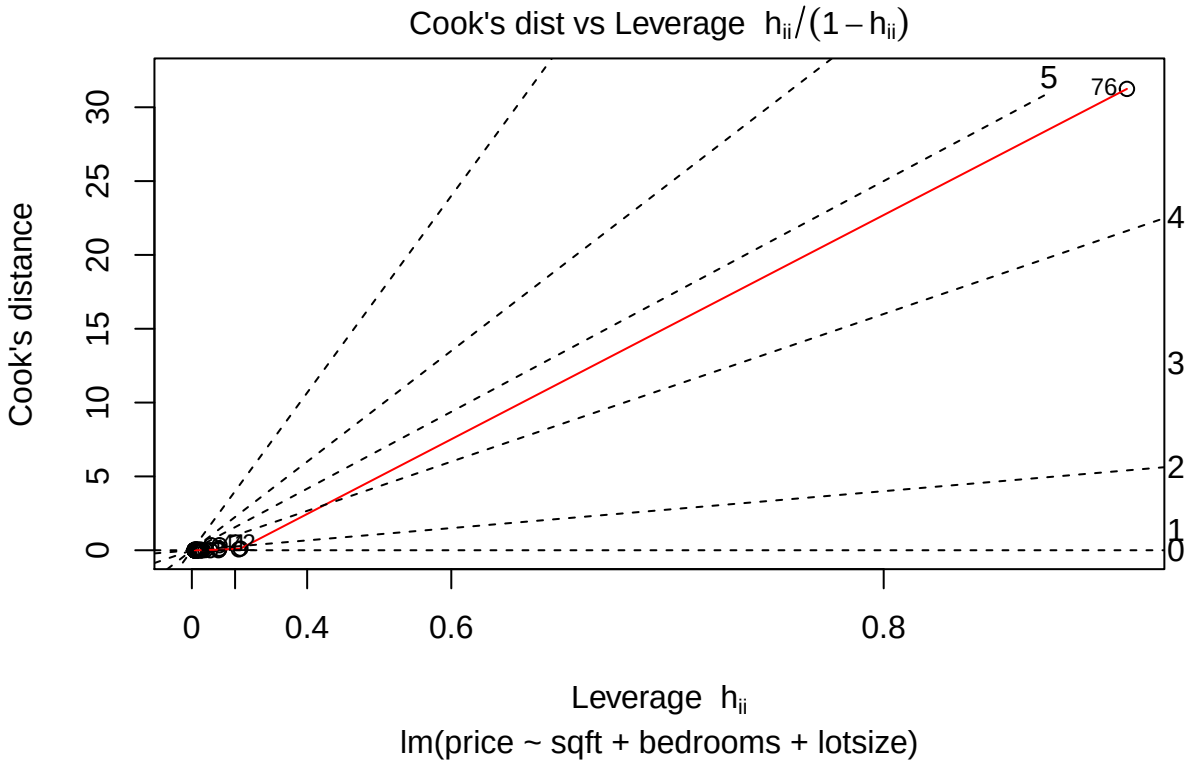
```
plot(price.fit, 5)
```



La observación 76 tiene leverage; las observaciones 41, 72, como en la gráfica anterior, son casos límites.

El último ploteo compara el leverage con la distancia de Cook.

```
plot(price.fit, 6)
```



La única observación problemática es la número 76.

Diagnóstico con pruebas de hipótesis

Se usará la librería `lmtest` para realizar el diagnóstico de un modelo de regresión lineal.

Heterocedasticidad

La prueba de hipótesis para heterocedsticidad es

$$H_0: \text{ El modelo es homocedástico}$$
$$H_1: \text{ El modelo no es homocedástico}$$

La implementación de Breusch-Pagan de la prueba de heterocedasticidad consiste en la estimación de los residuales al cuadrado mediante las variables explicativas del modelo, como se explica a continuación.

Considerando que

$$E[u|\mathbf{x}] = 0$$

la varianza es entonces igual a

$$Var[u|\mathbf{x}] = E[u^2|\mathbf{x}] - (E[u|\mathbf{x}])^2 = E[u^2|\mathbf{x}].$$

El supuesto de varianza constante (homocedasticidad), $Var(u|\mathbf{x}) = \sigma^2$, se puede escribir como

$$E[u^2|\mathbf{x}] = \sigma^2$$

que se puede interpretar como la *función de regresión poblacional* del modelo

$$u^2 = \sigma^2 + v,$$

donde v es un término de error con media cero y varianza constante.

Bajo la hipótesis nula de homocedasticidad el modelo

$$u^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_k x_k + v$$

no es significativo, es decir,

$$\gamma_1 = 0, \gamma_2 = 0, \cdots, \gamma_k = 0.$$

Por la tanto, la prueba de hipótesis de heterocedasticidad de Breusch-Pagan se plantea como

$$H_0 : \gamma_1 = 0, \gamma_2 = 0, \cdots, \gamma_k = 0$$
$$H_0 : \gamma_1 \neq 0, \gamma_2 \neq 0, \cdots, \gamma_k \neq 0$$

Para probar heterocedasticidad se estima el modelo

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

y se guardan los residuales, $\widehat{u}$. Después se estima la regresión auxiliar

$$\widehat{u}^2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + u$$

Si el modelo auxiliar es significativo, entonces se rechaza H_0 y se concluye que el modelo es heterocedástico; si el modelo no es significativo, entonces no hay evidencia de heterocedasticidad.

Consecuencias

Si un modelo es heterocedástico, entonces los estimadores son todavía insesgados y consistentes, pero no son los mejores (con menor varianza) entre los estimadores insesgados.

Una consecuencia importante es que las fórmulas de los intervalos de confianzas y las inferencias estadísticas no son válidas. Afortunadamente, es posible usar errores robustos a heterocedasticidad.

Véanse también los comentarios generales en la sección de Autocorrelación para lidiar con heterocedasticidad.

Prueba de heterocedasticidad

La prueba de heterocedasticidad de Breusch-Pagan se encuentra en la librería `lmtest`. Probaremos la hipótesis de heterocedasticidad en el modelo de precios de casas estimado en el capítulo anterior

```
library(lmtest)
```

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

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

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

```
bptest(price.fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  price.fit
## BP = 13.715, df = 3, p-value = 0.00332
```

Existe evidencia de heterocedasticidad; es necesario usar errores robustos a la heterocedasticidad.

Autocorrelación

El diagnóstico del Supuesto 3 determina si en el modelo estimado se detecta **autocorrelación serial**:

$$cov(u_i, u_j) \neq 0, \quad i \neq j.$$

Consecuencias

A continuación se presentan algunas consecuencias de la presencia de autocorrelación serial en la estimación de un modelo.

- Las estimaciones por mínimos cuadrados del modelo son todavía insesgadas, pero no son eficientes, aún con tamaños muestrales grandes.
- Por lo tanto, las estimaciones de los errores estándares pueden ser equivocadas.
- Lo que significa es que se pueden hacer inferencias estadísticas erróneas.
- En caso de correlación serial positiva se tendrá una subestimación de los errores estándares.

  - Aumento de la probabilidad de un error tipo I (rechazar la hipótesis nula cuando es verdadera).

- El valor de R^2 será más grande (comparado con su valor "correcto") bajo autocorrelación, porque la autocorrelación residua podrá subestimar el valor correcto de la varianza.

El término de error no es observable, por lo que las pruebas se efectuan sobre los residuales del modelo estimado.

Prueba de Durbin-Watson

Una de las primeras pruebas para la detección de autocorrelación serial es la prueba de Durbin-Watson.

El alcance de la prueba es determinar si u_t, u_{t-1} tiene correlación igual a cero.

La prueba de hipótesis asociada es

$$H_0 : cov(u_t, u_{t-1}) = 0$$
$$H_1 : cov(u_t, u_{t-1}) \neq 0$$

El resultado de interés es el rechazo de H_0.

El estadístico de prueba es

$$DW = \frac{\sum_{t=2}^{T}(\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=2}^{T}\widehat{u}_t^2},$$

donde T es el tamaño muestral.

Resulta que

$$DW \approx 2(1 - \widehat{\rho}),$$

donde $\widehat{\rho}$ es el estimador de la regresión

$$u_t = \rho u_{t-1} + v_t, \qquad v_t \sim N(0, \sigma_v^2)$$

y $\widehat{\rho} = corr(\widehat{u}_t, \widehat{u}_{t-1})$. Se encuentran más detalles en el libro de Brooks, Sección 5.5.3, pp. 193-196.

La correlación siempre está acotada entre -1, 1, por lo que

$$0 \leq DW \leq 4.$$

La hipótesis nula de no correlación serial, $cov(u_t, u_{t-1}) = 0$, es equivalente a $DW = 2$.

No se rechazará la hipótesis nula cuando DW está cerca de 2; el problema es que el estadístico de prueba DW no tiene una distribución típica.

El estadístico DW tiene dos valores críticos, d_L, d_U, determinados por el número de variables explicativas y del tamaño muestral.

En el libro de Brooks, p. 674, se encuentran valores tabulados asociado a DW.

La prueba de Durbin-Watson tiene como resultado tres posibles resultados. Cada resultado está determinado por el valor numérico del estadístico de prueba DW:

- No se rechaza H_0, si $d_U < DW < 4 - d_U$.
- Se rechaza H_0, si $0 < DW < d_L$ ó $4 - d_L < DW < 4$.
- La prueba no es concluyente, si $d_L < DW < d_U$ ó $4 - d_U < DW < 4 - d_L$.

Problema 1

Considera el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

con $T = 100$. Determina el área de rechaza; de no rechazo y los valores de DW para los cuales la prueba de Durbin-Watson no es concluyente. Usa los valores tabulados $d_L = 1.50$, $d_U = 1.58$.

Solución Usando los valores tabulados

- No se rechaza H_0, si $1.58 < DW < 2.42$.
- Se rechaza H_0, si $0 < DW < 1.50$, ó $2.50 < DW < 4$.
- La prueba no es concluyente, si $1.50 < DW < 1.58$, ó $2.42 < DW < 2.50$.

Prueba de Breusch-Godfrey

La prueba de Breusch-Godfrey es una prueba de autocorrelación conjunta, es decir, considera rezagos en los errores que pueden ser mayores a uno.

El modelo para los errores bajo esta prueba (hasta rezago r) es

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_r u_{t-r} + v_t, \qquad v_t \sim N(0, \sigma_v^2).$$

La prueba de hipótesis asociada es

$$\begin{aligned} H_0 &: \rho_1 = 0, \ldots, \rho_r = 0 \quad \Leftrightarrow \quad \text{no existe autocorrelación serial} \\ H_1 &: \rho_1 \neq 0, \ldots, \rho_r \neq 0 \quad \Leftrightarrow \quad \text{existe autocorrelación serial} \end{aligned}$$

El estadistico de prueba es χ^2_r, pero es también posible usar un estadístico F de exclusión.

Comentarios generales

1. En la prueba de Durbin-Watson tres supuestos son cruciales:
   (a) La regresión debe incluir un intercepto.

(b) los controles no deben ser estocásticos (dependientes de t).

(c) la regresión no puede incluir variables rezagadas.

2. En las pruebas de Durbin-Watson, Breusch-Godfrey no se considera el intercepto en las regresiones auxiliares.

3. Una desventaja de la prueba de Breusch-Godfrey es la determinación del rezago apropiado. No hay una respuesta correcta para todos los casos. Además de probar una variedad de rezagos, es útil considerar la frecuencia de los datos para decidir, la idea siendo que los errores pueden estar correlacionados sólo de manera periódica; por ejemplo, si la frecuencia de los datos son mensuales o trimestrales, valores apropiados de r pueden ser 12 o 4, respectivamente.

4. Las pruebas de Durbin-Watson, Breusch-Godfrey quieren determinar si se observa en los errores un efecto *autoregresivo* (de orden 1, en el caso de la prueba de Durbin-Watson; de orden r, en la prueba de Breusch-Godfrey).

5. La autocorrelación en los errores es un caso de heterocedasticidad en el modelo. Si se detectara autocorrelación en la estimación de un modelo hay dos caminos que se pueden seguir. El primero consiste en analizar la estructura del error, examinando el modelo autoregresivo que resulta. En el segundo caso, se trata la presencia de autocorrelación serial como un efecto que manejar mediante el uso de errores robustos a heterocedasticidad y consistentes a autocorrelación.[5]

6. De manera general, en un análisis econométrico puede interesar más la habilidad predictiva del modelo que la significancia estadística de algunas variables. En este caso, el análisis se enfocará más en el valor de parámetros como R^2 o criterios de información (AIC; BIC, etc.).

7. Si por otro lado interesa más la significancia estadística de algunas de las variables del modelo, entonces vale la pena analizar a fondo los efectos heterocedásticos, usando las herramientas apropiadas (modelos de series de tiempo AR, por ejemplo).[6]

Pruebas de autocorrelación serial con R

Probaremos la hipótesis de autocorrelación serial sobre los residuos del modelo estimado de precios de casas.

```
dwtest(price.fit)
```

```
##
##  Durbin-Watson test
```

[5] en la literatura anglosajona se usa el término HAC: *Heteroscedasticity Autocorrelation Consistent errors*

[6] La significancia estadística también debe considerar la potencia de una prueba de hipótesis (la probabilidad del rechazo de H_0 cuando es falsa). El ajuste heterocedástico aumenta, por lo general, la varianza de los estimadores, lo que disminuye la potencia de la prueba. Esto sugiere que, en este caso, podría ser mejor dejar algunos aspectos del modelo sin explicar.

```
##
## data:  price.fit
## DW = 2.0722, p-value = 0.6233
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bgtest(price.fit)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  price.fit
## LM test = 0.23651, df = 1, p-value = 0.6267
```

No hay evidencia de autocorrelación serial en los residuales.

Normalidad de los errores

El supuesto de normalidad de los errores en fundamental en la inferencia estadística de los estimadores por mínimos cuadrados.

A veces el supuesto de normalidad no se cumple porque la variable dependiente es positiva o tal vez por la presencia de *outliers* en los datos. En otros casos, la presencia de estacionalidad en los datos puede dar lugar a distribución no normal de los errores.

En muchas situaciones es posible usar considerar $\log(y)$, en lugar de y, como variable dependiente para cumplir con el supuesto de normalidad y realizar la estimación del modelo

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

En otras situaciones si el tamaño muestral es suficientemente grande, se puede usar el Teorema Central del Límite y usar *normalidad asintótica* para suplir a la falta de normalidad de los errores. A continuación se presenta el Teorema Central del Límite, con una visualización de sus consecuencias.

Pruebas de normalidad

La prueba de hipótesis asociada se plantea como

$$\begin{aligned} H_0: &\text{ Los errores son normales} \\ H_1: &\text{ Los errores no son normales} \end{aligned}$$

Las pruebas más usadas son la prueba de Shapiro-Wilk (librería `stats`)

```
shapiro.test(price.fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  price.fit$residuals
## W = 0.94288, p-value = 0.00079
```

y la prueba de Jarque-Bera (librería `tseries`; existen implementaciones de la prueba de Jarque-Bera en otras librerías también).

```
library(tseries)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

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

```
jarque.bera.test(price.fit$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  price.fit$residuals
## X-squared = 30.499, df = 2, p-value = 2.383e-07
```

Existe fuerte evidencia de errores no normales.

Nótese que la prueba gráfica de normalidad no fue muy indicativa de la falta de normalidad de los errores.

Prueba RESET

El supuesto de linealidad de los parámetros (pendietes e intercepto) y que la relación entre la variable dependiente e independientes es también lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

Se puede probar si la relación lineal entre las variables dependiente e independientes tiene que ser lineal usando la prueba RESET de Ramsey. La idea de la prueba es que si el modelo necesita términos cuadráticos o cúbicos de una(s) variable(s) independiente(s), entonces los regresores $\widehat{y}^2$, $\widehat{y}^3$ de los valores ajustado de y serán significativos en la regresión auxiliaria

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma_1 \widehat{y}^2 + \gamma_2 \widehat{y}^3 + u.$$

Entonces la prueba de hipótesis asociada es

$$\begin{aligned} H_0: &\text{ el modelo es adecuado funcionalmente } \Leftrightarrow \ \gamma_1 = 0,\ \gamma_2 = 0 \\ H_1: &\text{ el modelo no es adecuado funcionalmente } \Leftrightarrow \ \gamma_1 \neq 0,\ \gamma_2 \neq 0 \end{aligned}$$

El estadístico de prueba es ya sea $\chi^2 = TR^2_{aux} \sim \chi^2_2$, donde R^2_{aux} es el valor de R^2 de la regresión auxiliaria o se puede usar el estadístico F de exclusión de variables

$$F = \frac{SRC_r - SRC_{nr}}{SRC_{nr}} \cdot \frac{T-k-1}{2}$$

donde $T-k-1$ es el número de grados de libertad del modelo auxiliario.

Si se rechaza la hipótesis nula, es decir, se encuentra que el modelo lineal no es apropiado, La prueba RESET no indica la forma funcional correcta.

Consecuencias Usar una forma funcional no apropiada puede causar sesgo en la estimación de las pendientes e intercepto, seemejante al sesgo de variable omitida.

Una manera para obviar a una forma funcional incorrecta es añadir variables cuadráticas al modelo y eliminar sucesivamente aquellas que no son significativas.

La teoría económico-financiera puede ayudar en determinar qué variable(s) necesita(n) *retornos marginales variables*.

Prueba RESET con R

Una implementación de la prueba RESET se encuentra en la librería `lmtest`:

```
library(lmtest)
resettest(price.fit)
```

```
##
##  RESET test
##
## data:  price.fit
## RESET = 4.386, df1 = 2, df2 = 81, p-value = 0.01554
```

hay evidencia que el modelo no es adecuado funcionalmente.

Multicolinealidad

El fenómeno de *multicolinealidad* es un problema de los datos usados en una regresión lineal, no es una violación de los supuestos del Modelo de Regresión Lineal Clásico; por esto, los estimadores son MELI (mejores estimadores lineales insesgados).

Multicolinealidad afecta principalmente los errores estándares de los estimadores; más adelante, en la sección "Varianza de los estimadores" se desarrolla más en detalle este punto.

Se observa multicolinealidad entre par de variables cuando la correlación entre estas variables se encuentra, en valor absoluto, por arriba de .80[7]

[7] Vale la pena subrayar que no hay un valor matemáticamente determinado; el valor proporcionado se puede considerar una medida empírica.

Una consecuencia de multicolinealidad en una regresión múltiple es obtener un valor R^2 relativamente alto, con significancia estadística del modelo, pero con pocas variables significativas. Por ejemplo, en la siguiente simulación se observan todos estos puntos.

Se define la semilla `set.seed(123)` para poder replicar los resultados en este documento.

Se define la variable x por medio de 1000 muestra de una distribución normal estándar. La variable y se define como $3 + x$ más un pequeño error, también normal estándar.

Después, la variable $z = 2x$, más una variable normal, con media cero y varianza 10^{-4} (nótese que en R la distribución normal se define por medio de la desviación estándar).

Finalmente, se define una base de datos con las tres variables así definidas.

Se cargan las librerías `car`, `lmtest`, `sandwich` para las pruebas sucesivas.

```
library(car)
```

```
## Loading required package: carData
```

```
library(lmtest)
library(sandwich)
set.seed(123)
x <- rnorm(1000)
y <-  3 + x + rnorm(1000, 0, .5)
z <- 2*x + rnorm(1000, mean = 0, sd = .01)
datos <- data.frame(y, x, z)
```

Se estima el modelo

$$y = \beta_0 + \beta_1 x + \beta_2 z + u :$$

```
m1 <- lm(y ~ x + z)
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x + z)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49005 -0.34168  0.00378  0.34357  1.63306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.02081    0.01592 189.756   <2e-16 ***
## x           -1.86447    3.25520  -0.573    0.567
## z            1.45438    1.62773   0.894    0.372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5032 on 997 degrees of freedom
## Multiple R-squared:  0.8092, Adjusted R-squared:  0.8088
## F-statistic:  2115 on 2 and 997 DF,  p-value: < 2.2e-16
```

Observamos que el valor de $R^2 = .8092$, el modelo es altamente significativo estadísticamente, pero ninguna de las variables explicativas es significativa.

Por eso, es buena costumbre estimar la matriz de correlación del dataframe (o quizás de algunas de sus variables):

```
cor(datos)
```

```
##           y         x         z
## y 1.0000000 0.8994877 0.8995377
## x 0.8994877 1.0000000 0.9999878
## z 0.8995377 0.9999878 1.0000000
```

Se observa que la correlación entre las variables x, z es alta (por arriba del 89%), lo que implica evidencia de multicolinealidad.

Varianza de los estimadores

Se puede probar que la varianza del estimador $\widehat{\beta}_j$, $j = 1, \ldots, k$, asociado al modelo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u,$$

es

$$Var(\widehat{\beta}_j) = \frac{\sigma^2}{\sum_i (x_{j,i} - \overline{x}_j)^2} \frac{1}{1 - R_j^2}$$

El término $1/(1 - R_j^2)$ es una medida de dependencia lineal entre la variable x_j y las demás variables explicativas[8]. Este término se conoce como *factor de inflación de varianza* (VIF: Variance Inflation Factor, en la literatura anglosajona):

$$VIF = \frac{1}{1 - R_j^2}$$

Por ejemplo, si $VIF > 4$ implica un $R_j^2 > .75$. Una implementación de VIF en R se encuentra en la librería `car`:

```
vif(m1) > 4
```

```
##    x    z
## TRUE TRUE
```

[8] No coincide necesariamente con la correlación muestral entre los pares de variables explicativas

El resultado es una stringa lógica que es igual a verdadero, si el $VIF > 4$; falso, de otra manera. Rematando el resultado de correlación de arriba, concluimos que existe multicolinealidad entre las variables x, z.

Eliminando una de las variabeles,

```
m2 <-  lm(y ~x)
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51394 -0.34572  0.00213  0.35436  1.64557
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.02052    0.01591  189.80   <2e-16 ***
## x            1.04402    0.01605   65.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5032 on 998 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.8089
## F-statistic:  4229 on 1 and 998 DF,  p-value: < 2.2e-16
```

el resultado es mucho mejor, en término de significancia estadística. La diferencia en el valor de R^2-ajustado es mínima (afecta el cuarto dígito decimal).

Ejemplos generales

En esta sección presentaremos diferentes ejemplos, incluyendo el diagnóstico.

Se desgloasarán los pasos del diagnóstico sólo para el Ejemplo 4.

Ejemplo 4

En esta sección se estimará un modelo de regresión lineal múltiple y después se realizará el diagnóstico del modelo estimado.

Se necesitarán las librerías `lmtest`, `sandwich`. La segunda librería se usará para determinar errores robustos a heterocedasticidad.

El dataset es `Investment` incluido en la librería `sandwich`.

Se estimará la inversión (en términos reales) como función del GNP real y de la tasa de interés real

$$RealInv = \beta_0 + \beta_1 RealGNP + \beta_2 RealInt + u.$$

Después

- se probará la normalidad de los errores; la normalidad es una hipótesis importante en la prueba de Durbin-Watson.
- se determinará la existencia de efectos de autocorrelación serial;

Finalmente, se calcularán los errores robustos a heterocedasticidad y consistente a autocorrelación serial.

Estimación

Primeramente, cargaremos las librerías necesarias y el dataset:

```
library(lmtest)
library(sandwich)
data(Investment)
```

Después, se estimará el modelo de regresión múltiple y se desplegará el modelo estimado:

```
m1 <- lm(RealInv ~ RealGNP + RealInt, data = Investment)
summary(m1)
```

```
##
## Call:
## lm(formula = RealInv ~ RealGNP + RealInt, data = Investment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.987  -6.638   0.180  10.408  26.288
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.53360   24.91527  -0.503    0.622
## RealGNP       0.16914    0.02057   8.224 3.87e-07 ***
## RealInt      -1.00144    2.36875  -0.423    0.678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.21 on 16 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8141, Adjusted R-squared:  0.7908
```

```
## F-statistic: 35.03 on 2 and 16 DF,  p-value: 1.429e-06
```

Normalidad de los errores

Se probará la normalidad de los errores:

$$H_0 : \text{Los errores son normales}$$
$$H_1 : \text{Los errores no son normales}$$

```
res = m1$res
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.94852, p-value = 0.3728
```

No tenemos evidencia en contra del supuesto de normalidad de los errores.

Autocorrelación serial

Prueba de Durbin-Watson

Detección de efectos de autocorrelación serial. Considerando que las variables son series de tiempo, es muy probable que dependan del tiempo (no estocásticas). Por lo tanto, el resultado de la prueba de Durbin-Watson no es recomendable; se incluye de todas maneras como un ejemplo de su aplicación. Por defecto, la función `dwtest()` considera como hipótesis alternativa `la correlación verdadera es positiva`.

```
dwtest(m1)
```

```
##
##  Durbin-Watson test
##
## data:  m1
## DW = 1.3215, p-value = 0.01696
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(m1, alternative = "two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  m1
## DW = 1.3215, p-value = 0.03393
## alternative hypothesis: true autocorrelation is not 0
```

Prueba de Breusch-Godfrey

Tomando en cuenta que los datos son anuales, con posibles efectos estacionales, usaremos la prueba de Breusch-Godfrey, con rezagos $r = 1$, $r = 3$, $r = 6$, $r = 12$.

```
bgtest(m1, order = 1, type = "Chisq")
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  m1
## LM test = 1.2816, df = 1, p-value = 0.2576
```

```
bgtest(m1, order = 3, type = "Chisq")
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 3
##
## data:  m1
## LM test = 10.362, df = 3, p-value = 0.01573
```

```
bgtest(m1, order = 6, type = "Chisq")
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 6
##
## data:  m1
## LM test = 15.042, df = 6, p-value = 0.01993
```

```
bgtest(m1, order = 12, type = "Chisq")
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 12
##
## data:  m1
## LM test = 17.696, df = 12, p-value = 0.1252
```

Se observan posibles efectos de autocorrelación serial en los residuos.

Errores robustos HAC

Pasamos al cálculo de los errores robustos HAC. La base teórica es el artículo de Newey-West [**NW**]; los detalles se encuentran en la Bibliografía más abajo.

Matriz de covarianza HAC

Una implementación de este estimador se encuentra en la función `NeweyWest()` de la librería `sandwich`. Cuando se usa esta función, se debe incluir el parámetro adicional `prewhite = F`.

El resultado es la matriz de covarianza HAC.

```
NW <- NeweyWest(m1, lag=6, prewhite = F)
```

Errores robustos HAC

Los errores robustos HAC se pueden determinar calculando la raíz cuadrada de la diagonal de la matriz de covarianza:

```
sqrt(diag(NW))
```

```
## (Intercept)     RealGNP     RealInt
## 15.68076244  0.01339627  3.37357317
```

o usando la función `coeftest()` de la librería `sandwich`:

```
coeftest(m1, vcov = NW)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -12.533601  15.680762 -0.7993    0.4358
## RealGNP       0.169136   0.013396 12.6256 9.819e-10 ***
## RealInt      -1.001438   3.373573 -0.2968    0.7704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prueba RESET

Determinar si el modelo es adecuado funcionalmente:

```
resettest(m1)
```

```
##
##  RESET test
##
## data:  m1
## RESET = 0.63134, df1 = 2, df2 = 14, p-value = 0.5464
```

No hay evidencia que el modelo sea funcionalmente no adecuado.

Ejemplo 5: Pruebas de cambio estructural

En la estimación de una regresión lineal, una de las variables está sujeta a cambios drásticos, de manera que surge la pregunta de si, en la muestra bajo consideración, se pueden identificar

dos sub-muestras independientes. Por ejemplo, considerando variables que dependen del tiempo, se puede analizar el efecto antes y después de un "choque" en la población.

En otras instancias, se puede considerar el agrupamiento de dos muestras independientes, para aumentar el tamaño muestral o quizás para determinar si existe una diferencia significativa entre la estimación en las dos muestras.

Consideraremos primeramente el caso de un "choque" en una variable explicativa. Se estimará el efecto de los precios en el índice `FTSE100` sobre los precios de casas en UK, en el periodo 30 de abril 1994 a 31 de mayo de 2013.

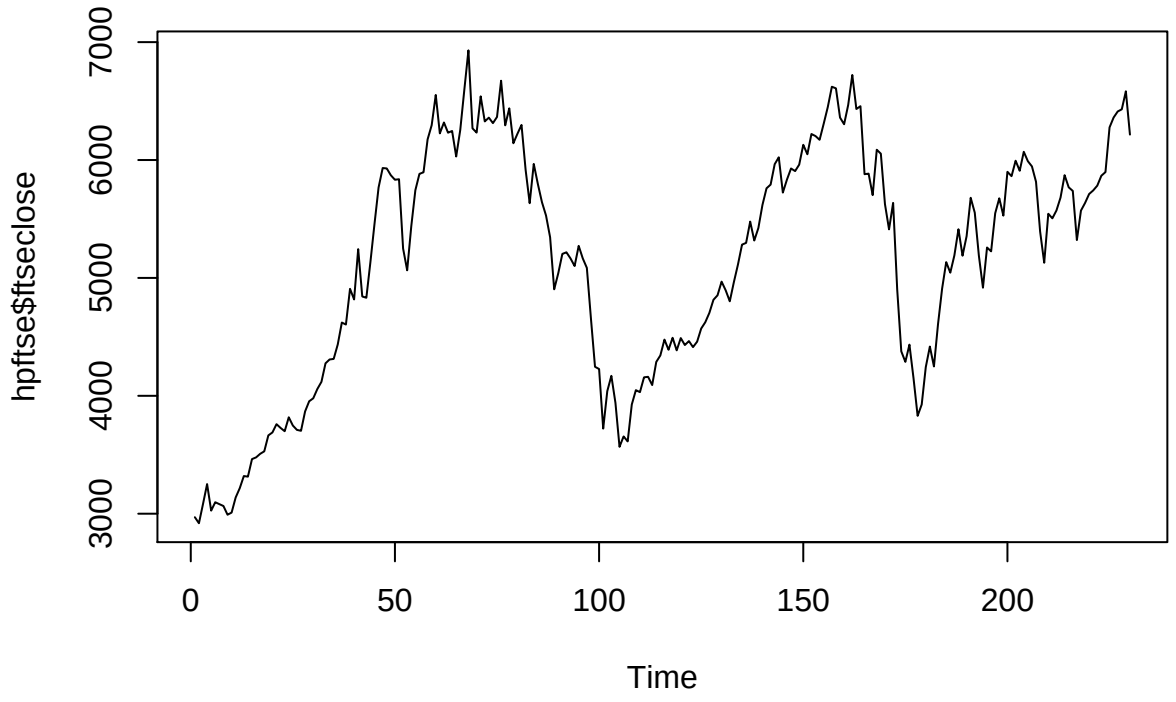
```
hpftse <- read.csv("HPFTSE100.csv")
```

Las variables en el dataset son obs, ftseclose, hp, D, donde *obs* es la fecha de la observación, *hp* es el precio de las casas británicas (en libras esterlinas); *ftseclose* es el precio mensual, al cierre, del índice `FTSE100` y D es una variable "dummy",

$$D = \begin{cases} 0, & \text{para observaciones con fecha anterior a 31 de mayo, 2007,} \\ 1, & \text{para observaciones con fecha posterior a 31 de mayo, 2007,} \end{cases}$$

Se puede observar que el índice `FTSE100` sufrió una caída alrededor del 2006-2007 (y empeoró a causa de la crisis financiera global).

```
plot.ts(hpftse$ftseclose)
```



Se considera la hipótesis: *el efecto de la crisis financiera global afectó de manera significativa los precios de las casas británicas.*

Para estimar el efecto de "antes" y "después", se considera el modelo

$$hp = \beta_0 + \beta_1 ftseclose + \delta_0 D + \delta_1 D \cdot ftseclose + u,$$

en donde la hipótesis de estudio es equivalente, entonces, a una prueba de hipótesis de exclusión de variables,

$$H_0 : \delta_0 = 0,\ \delta_1 = 0$$
$$H_1 : \delta_0 \neq 0,\ \delta_1 \neq 0$$

La estimación da

```
m_r <- lm(hp ~ ftseclose, data = hpftse)
m_nr <- lm(hp ~ ftseclose + D + I(D*ftseclose), data = hpftse)
```

El estadístico de prueba asociado a la prueba de exclusión de variables es

$$F = \frac{SRC_r - SRC_{nr}}{SRC_{nr}} \frac{n-k-1}{q},$$

es decir,

```
src_r = sum((m_r$residuals)^2)
src_nr = sum((m_nr$residuals)^2)
f <- ((src_r - src_nr)/src_nr)*(226/2)
pf(f,2,226, lower.tail = F) < .05 # prueba hip con valor-p
```

```
## [1] TRUE
```

```
qf(.05, 2, 226, lower.tail = F) < f
```

```
## [1] TRUE
```

```
          # prueba hip con valor crit distribución F(2,226)
```

Rechazamos la hipótesis nula y concluimos que sí, la crisis financiera afectó de manera significativa los precios de las casas británicas.

Recordatorio Se rechaza la hipótesis nula en cualquiera de las dos situaciones *equivalentes* a continuación:

- si el valor-p del estadístico de prueba es **menor** al nivel de significancia α;
- si el valor del estadístico de prueba es **mayor** al valor crítico de la distribución del estadístico de prueba.

Pruebas de diagnóstico

Vale la pena verificar si se cumplen los supuestos del Modelo de Regresión Lineal Clásico:

- Homocedasticidad
- Autocorrelación serial
- Normalidad de los errores
- prueba RESET de especificación del modelo

Homocedasticidad

Si el modelo no restringido es heterocedástico, entonces se debe usar el estadístico de Wald (estadístico de prueba robusto a heterocedasticidad).

```
library(lmtest)
library(sandwich)
bptest(m_nr)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m_nr
## BP = 117, df = 3, p-value < 2.2e-16
```

El modelo es heterocedástico, por lo que se debe usar el estadístico de Wald para determinar si rechazar, o no, la hipótesis nula:

```
waldtest(m_r, m_nr, vcov = vcovHC(m_nr, type = "HC0"))
```

```
## Wald test
##
## Model 1: hp ~ ftseclose
## Model 2: hp ~ ftseclose + D + I(D * ftseclose)
##   Res.Df Df      F    Pr(>F)
## 1    228
## 2    226  2 206.51 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La conclusión es la misma: la crisis financiera sí causó un cambio en los precios de las casas británicas.

Autocorrelación serial

Usaremos las pruebas de Durbin-Watson y de Breusch-Godfrey.

Prueba de Durbin-Watson

```
dwtest(m_nr, alternative = "two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  m_nr
## DW = 0.016806, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

Prueba de Breusch-Godfrey

```
bgtest(m_nr, order = 5, type = "Chisq")
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 5
##
## data:  m_nr
## LM test = 225.89, df = 5, p-value < 2.2e-16
```

En este caso se debe usar lal corrección de Newey-West de autocorrelación serial para el cálculo de valor-p asociado al estadístico de exclusión de variables:

```
nw <- NeweyWest(m_nr, lag = 5, prewhite = F)
waldtest(m_r, m_nr, vcov = nw)
```

```
## Wald test
##
## Model 1: hp ~ ftseclose
## Model 2: hp ~ ftseclose + D + I(D * ftseclose)
##   Res.Df Df      F    Pr(>F)
## 1    228
## 2    226  2 36.042 2.593e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se sigue rechazando la hipótesis nula: la conclusión es la misma.

Normalidad de los errores

La hipótesis nula es que los errores son normales; la alternativa es que no siguen la distribución normal.

```
shapiro.test(m_nr$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m_nr$residuals
```

```
## W = 0.90725, p-value = 9.412e-11
```

Los errores no son normales. En este caso, el tamaño muestral nos permite el uso de normalidad asintótica que permite la justificación de las inferencias estadísticas que se realizaron.

Prueba RESET

Ya que se determinó que el modelo (no restringido) es heterocedástico, se usará la corrección heterocedástica de esta prueba:

```
resettest(m_nr, power = 2:3, vcov = vcovHC)
```

```
##
##  RESET test
##
## data:  m_nr
## RESET = 13.517, df1 = 2, df2 = 226, p-value = 2.869e-06
```

El modelo no resulta funcionalmente adecuado; quizás, necesita términos cuadráticos en $ftseclose$.

Ejemplo 6: Comparación de dos modelos de regresión

Se consideran dos firmas con giro de negocio muy parecido (GE, WH)

La hipótesis económica es:

> La inversión (inv) para firmas diferentes, pero con giro de negocio parecido, se puede explicar por medio del valor de la firma (V) y capital (K).

Modelo para GE

$$inv = a_0 + a_1 V + a_2 K + u$$

Modelo para WH

$$inv = b_0 + b_1 V + b_2 K + u$$

Si el mismo modelo puede explicar los datos de las dos firmas, entonces

$$a_i = b_i, \quad i = 1, 2.$$

Usando variables "dummy", se puede verificar si las dos regresiones son idénticas (a un nivel de significancia dado), por medio de una prueba F (**Prueba de Chow**).

Para probar la hipótesis económica, se introduce una variable binaria

$$\delta = \begin{cases} \delta = 1, & \text{datos de WH} \\ \delta = 0, & \text{datos de GE} \end{cases}$$

y estimamos el modelo

$$inv = \beta_0 + \beta_1 V + \beta_2 K + \delta_0 \delta + \delta_1 (\delta \cdot V) + \delta_2 (\delta \cdot K) + u$$

La prueba de hipótesis apropiada ($\alpha = 0.05$) es

$$H_0 : \delta_0 = 0,\ \delta_1 = 0,\ \delta_2 = 0$$
$$H_a : \delta_0 \neq 0,\ \delta_1 \neq 0,\ \delta_2 \neq 0$$

Para la estimación se usarán los datos en el archivo `gewh.csv`.

```
gewh <- read.csv("gewh.csv")
```

Comentario Nótese como la agregación de los datos de las dos empresas aumenta el tamaño muestral. En este caso es positivo, porque hay sólo 20 observaciones por cada firma.

La estimación del modelo da

```
gewh.fit <- lm(INV ~ V + K + D + I(D*V) + I(D*K), data = gewh)
summary(gewh.fit)
```

```
##
## Call:
## lm(formula = INV ~ V + K + D + I(D * V) + I(D * K), data = gewh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.513 -10.600  -2.890   6.484  58.738
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.94704   23.62626  -0.421    0.676
## V            0.02655    0.01172   2.265    0.030 *
## K            0.15170    0.01936   7.837 4.02e-09 ***
## D            9.43765   28.80548   0.328    0.745
## I(D * V)     0.02635    0.03435   0.767    0.448
## I(D * K)    -0.05929    0.11695  -0.507    0.615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 34 degrees of freedom
## Multiple R-squared:  0.8278, Adjusted R-squared:  0.8025
## F-statistic:  32.7 on 5 and 34 DF,  p-value: 4.611e-12
```

Cálculo del estadístico de prueba

```r
gewh.fit.r <- lm(INV ~ V + K, data = gewh)
src.r <- sum((gewh.fit.r$residuals)^2)
src.nr <- sum((gewh.fit$residuals)^2)
f <- ((src.r - src.nr)/src.nr)*(34/3)
pf(f,3,34, lower.tail = F) #valor-p
```

```
## [1] 0.3284676
```

No se rechaza la hipótesis nula. La conclusión es que

> no ha evidencia **en contra** de la hipótesis económica que, para firmas con el mismo giro de negocio, la inversión se explica mediante el valor de la firma y capital.

Bibliografía

EPI Economic Policy Institute, State of Working America Data Library, "Median/average hourly wages," 2019, con datos de CPS ORG.

IS Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

NW Newey WK & West KD (1987), A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55, 703–708.

W Wooldridge Jeffrey M, *Introducción a la Econometría* Quinta edición, Cengage Learning, 2015.