

Introducción al Análisis de Regresión lineal simple y múltiple

La econometría trata de la aplicación de las técnicas estadísticas al caso de los modelos teóricos que usa el economista en situaciones reales, donde además del análisis cualitativo, el economista debe hacer análisis cuantitativo, esto es, debe especificar explícitamente una ecuación que represente el proceso económico en estudio y someterlo a prueba, contrastándolo con datos. Esto es necesario porque, al final, todo análisis teórico se debe aplicar a situaciones concretas y hay que hacer pronósticos que sean acertados, dentro de los límites que impone el hecho de trabajar con muestras y no con censos.

De todos los modelos, los más simples y frecuentes son los lineales, donde hay una variable dependiente Y cuyos valores se desea explicar a partir de los valores de una u otras variables independientes o explicativas X_1, X_2, \dots, X_p . En este contexto, un supuesto económico es que la respuesta de Y es proporcional a las v. independientes. Específicamente, que *cambios en cada v.i. X_j generan cambios en Y a una tasa constante, digamos β_j , con ciertas variaciones aleatorias ε que se suman a Y .*

El caso más sencillo, que es el modelo inicial de la econometría, es el del modelo de regresión lineal simple o bivariado, especificado por la ecuación: $E(Y|X) = \alpha + \beta X$, donde se incorpora el efecto de variación aleatoria o residuo aleatorio a la tendencia promedio de Y dado el valor de la v. independiente X , de modo que se pasa del modelo económico determinista $Y = \varphi(X; \alpha, \beta)$, $\frac{\partial \varphi}{\partial X} \neq 0$, al modelo econométrico $E(Y|X) = \varphi(X; \alpha, \beta) = \alpha + \beta X$, $\beta \neq 0$ y de ahí al modelo de datos $Y = \alpha + \beta X + \varepsilon$, $\beta \neq 0$, $\varepsilon \sim N(0, \sigma^2)$.

Una extensión natural del modelo de regresión simple es el modelo de regresión múltiple, donde el valor esperado de la variable dependiente Y es función lineal de p variables independientes o “explicativas” X_1, X_2, \dots, X_p vía el modelo econométrico $E(Y|X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$. El modelo de datos correspondiente es $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

En este documento veremos desde un punto de vista práctico, cómo se analizan estadísticamente estos dos modelos, con ayuda del paquete estadístico R, que es muy completo y gratuito. Empezaremos con el modelo de regresión simple.

1.1 Uso del modelo de regresión lineal simple.

Este modelo se aplica cuando se tiene una teoría que dice que, salvo variaciones aleatorias, una determinada variable cuantitativa X condiciona a otra variable cuantitativa Y de modo que cambios en X inducen cambios *proporcionales* en Y ; *deseamos contrastar la teoría con datos de una muestra aleatoria.*

Geométricamente la proporcionalidad equivale a que en un plano cartesiano XY , las parejas de valores (X, Y) describen o siguen una trayectoria rectilínea.

Algebraicamente la proporcionalidad equivale a que X e Y satisfacen la ecuación $Y = \alpha + \beta X + \varepsilon$, donde α y β son constantes características o sea son “parámetros” y ε es una variación aleatoria debida a que los agentes económicos no siempre tienen el mismo comportamiento.

Ejemplo 1

Un economista trabaja para una cadena de supermercados que desea abrir una nueva sucursal en un distrito de la ciudad siempre que haya posibilidad de tener suficientes ventas mensuales. El economista sugiere que las ventas deben estar asociadas directamente y proporcionalmente a la riqueza de todos los hogares del distrito y que una buena aproximación a esta riqueza es el total del valor de autoavalúo de las viviendas del distrito, que en este distrito es 28 millones. El economista consigue, de los catastros de los distritos donde la cadena tiene sucursales, el valor total de autoavalúos de las viviendas del distrito y también el valor de las ventas de los correspondientes establecimientos de la cadena en el último mes, ambos en millones de unidades monetarias.

Los datos son:

Sucursal	1	2	3	4	5	6	7	8	9	10
Total de Autoavalúos	66.0	14.0	47.9	77.9	57.8	30.6	36.0	64.2	70.0	42.2
Ventas	4.0	2.3	2.9	3.9	3.3	2.7	3.1	3.5	3.8	3.3

Si hubiera relación la proporcionalidad sugerida, entonces un simple diagrama de parejas *XY* debería seguir una tendencia lineal.

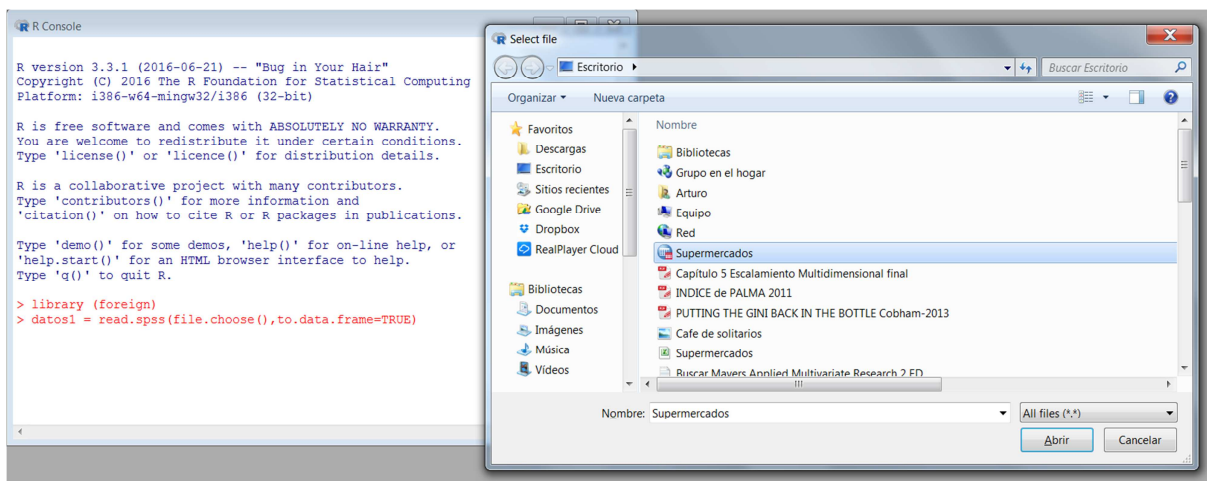
Los datos están digitados en SPSS que es un programa estadístico de uso corriente pero comercial. El archivo se llama Supermercados.sav (la extensión “sav” es la que usa SPSS para sus archivos de datos). Para leer los datos con R debemos cargar la librería “foreign” que ya viene en el paquete básico (“base”) de instalación de R. Para activar “foreign”, basta escribir en la consola de R (que es la ventana que se abre por default al cargar R en la computadora) la orden “**library (foreign)**” e ingresarla, como se muestra abajo:

library (foreign)

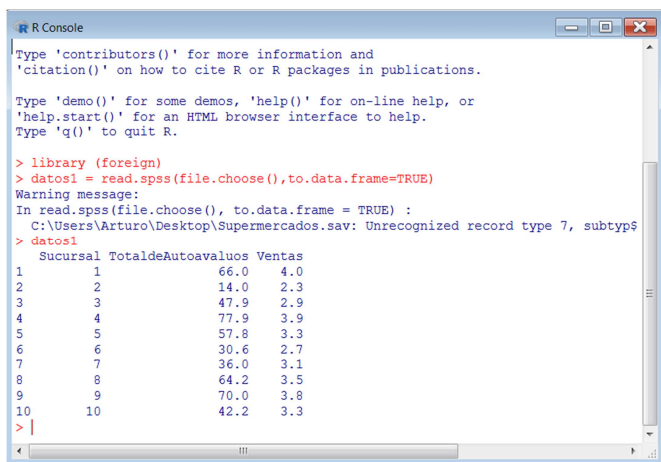
Luego para que R lea el archivo Supermercados.sav, usamos el comando “**read.spss(file.choose(),to.data.frame=TRUE)**” que indica a R que va a leer un archivo SPSS. La opción file.choose() dentro del paréntesis indica a R que nos permita ubicar el archivo dentro del directorio donde está colocado. De no escribir esta opción tendríamos que especificar la ruta completa de la ubicación del archivo.

Antes del comando read.spss debemos indicar a R el nombre que usará para identificar el archivo de datos ya convertido a formato de R. En este caso llamaremos al archivo “datos1” e indicaremos a R que ése es el nombre escribiéndolo antes del comando read.spss seguido de =, como se muestra abajo, teniendo cuidado de mantener el espacio en blanco entre datos1, = y read.spss . Al escribir el comando e ingresarlo, R abre una ventana para que podamos ubicar el archivo (que en este ejemplo está en el escritorio de Windows).

datos1 = read.spss(file.choose(),to.data.frame=TRUE)



seleccionamos el archivo Supermercados.sav con “Abrir” y R lee el archivo. Para verificarlo ingresamos la orden **datos1** y R nos muestra los datos (esto no es obligatorio, lo hacemos ahora para verificar la correcta importación de datos):

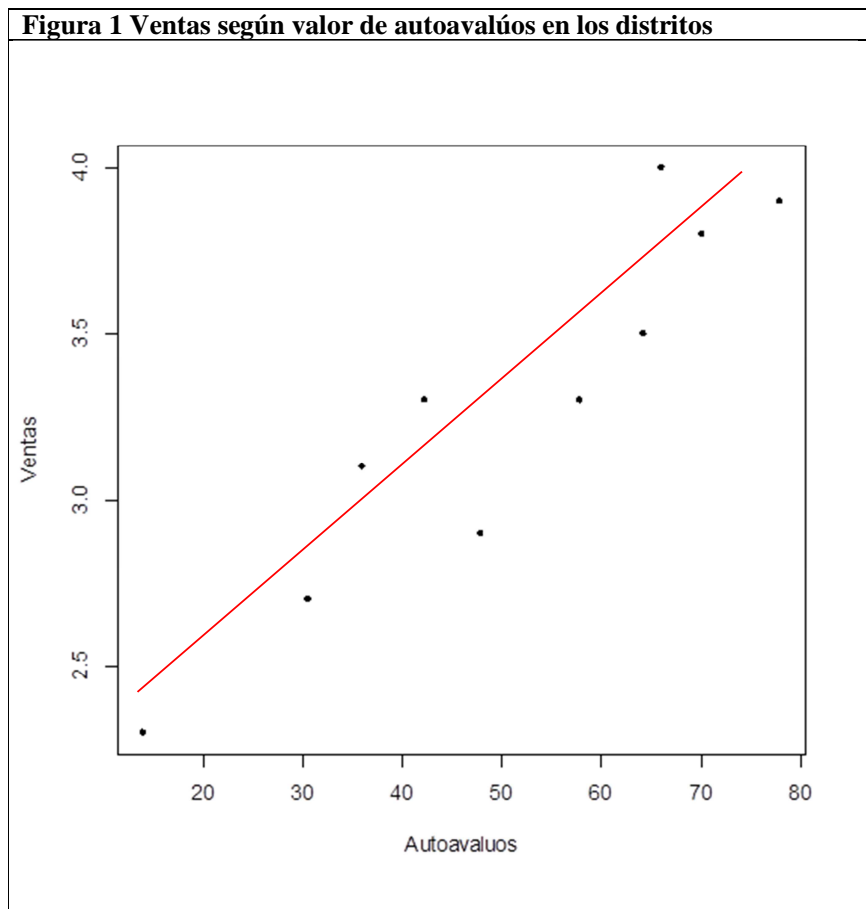


Regresando a la verificación visual de la linealidad con un diagrama XY (“Diagrama de dispersión” es su nombre técnico): Usamos el comando “plot”

```
plot(datos1[,2], datos1[,3], xlab = "Autoavaluos", ylab = "Ventas", pch=20)
```

dentro del comando plot, al escribir datos1[,2] estamos indicando a R que debe tomar del archivo de datos, llamado datos1 que está en formato de matriz, la segunda columna de datos, que corresponde a la variable TotaldeAutoavaluos y que esta variable irá en el eje X; análogamente datos1[,3] indica a R que la tercera columna de datos1 es la variable que irá en el eje Y. La etiqueta del eje X que R denota xlab (de x label) la escribimos entre comillas "Autoavaluos" y la del eje Y, ylab, la escribimos "Ventas".

R abre una ventana de gráficos y nos muestra el diagrama de dispersión XY de abajo (el gráfico se puede copiar y pegar a un documento, como hemos hecho en este ejemplo, donde además hemos añadido la recta de tendencia lineal, que R no muestra)



Se observa que hay relación, pero los puntos no caen sobre una recta, aunque la siguen, pero con algo de variabilidad, esto es, hay puntos algo alejados de la tendencia.

La cuestión es ¿Cómo verificar la hipótesis? O sea ¿Sería válido el modelo $Y = \alpha + \beta X + \varepsilon$? Y si fuera así ¿Cómo usar los datos para resolver el problema y tomar la decisión de dar la recomendación de abrir o no la nueva sucursal

1.2 Supuestos y parámetros del modelo lineal simple

Dado el modelo $Y = \alpha + \beta X + \varepsilon$, evaluado en una muestra aleatoria de n parejas $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, en general se cumple $Y_j = \alpha + \beta X_j + \varepsilon_j$ donde ε_j representa el efecto mezclado de otras variables no controladas, para el cual se asume que se comporta como “el azar”, como variable aleatoria: $\varepsilon \sim N(0, \sigma^2)$

habiendo independencia entre observaciones

Formalmente:

(a) $E(\varepsilon_j) = 0 \quad \forall j$

(b) $V(\varepsilon_j) = \sigma^2 = \text{constante} \quad \forall j$

(c) $\rho_{\varepsilon_j \varepsilon_{j'}} = 0 \quad \forall j \neq j'$

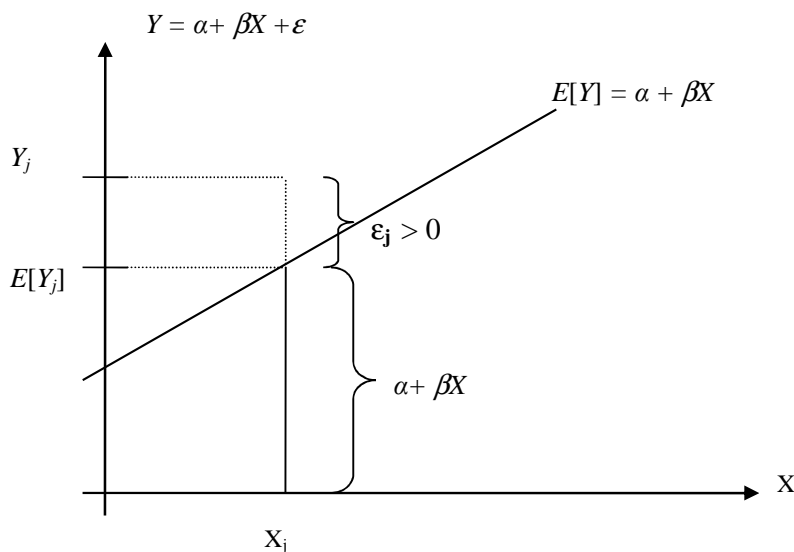
(d) X es de valores predeterminados, medidos antes de registrar los valores de Y .

Los supuestos implican que $E(Y) = \alpha + \beta X$ y $V(Y) = \sigma^2$

Parámetros del modelo:

- La constante α , llamada “constante” o “intercepto” en programas como Excel, R, SPSS o Stata, y en otros textos, “ordenada en el origen” o “intercepto” de la recta. En general es el valor esperado o promedio de Y cuando X es cero. Pero no siempre esto último tiene interpretación práctica y eso sólo un parámetro de ajuste.
- La constante β , llamada la “pendiente” de la recta. En general mide en cuántas unidades se espera que varíe Y cuando X aumenta en 1 unidad. Es la inclinación de la recta: Si $\beta > 0$ la recta se inclina a la derecha, si $\beta < 0$ se inclina a la izquierda.
- σ^2 es la varianza del azar representado por el residuo aleatorio ε ; σ es la variación promedio arriba o debajo de la recta $E(Y|X) = \alpha + \beta X$
- Se considera que $Y = \alpha + \underbrace{\beta X}_{\text{Efecto de } X} + \underbrace{\varepsilon}_{\text{Efecto de azar}}$

Figura 2 Efectos de X y del azar



Descomposición de Y_j según el modelo lineal

El gráfico muestra la "descomposición" del valor de Y_j como una parte debida a X_j (a través del efecto lineal) y una parte debida al azar, que en este caso ha dado un valor positivo (un $\varepsilon_j > 0$)

Podemos pronosticar el valor esperado de Y , $E(Y) = \alpha + \beta X$, estimando α y β y reemplazando luego en el modelo. El "error" en el pronóstico se puede aproximar en una primera instancia con la estimación de σ .

1.3 Los estimadores

Se obtienen aplicando el método de mínimos cuadrados:

- Si $\varepsilon_j = [Y_j - \alpha - \beta X_j]$ la función objetivo por minimizar es $Q(\alpha, \beta) = \sum_{j=1}^n [Y_j - \alpha - \beta X_j]^2$
- Aplicando derivadas:

$$\nabla Q(\alpha, \beta) = 0 \Leftrightarrow \frac{\partial Q(\alpha, \beta)}{\partial \alpha} = 0 \text{ y } \frac{\partial Q(\alpha, \beta)}{\partial \beta} = 0 \Leftrightarrow$$

$$\frac{\partial Q(\alpha, \beta)}{\partial \alpha} = 0 \Rightarrow \frac{\partial}{\partial \alpha} \sum_{j=1}^n [Y_j - \alpha - \beta X_j]^2 = -2 \sum_{j=1}^n [Y_j - \alpha - \beta X_j] = 0$$

$$\frac{\partial Q(\alpha, \beta)}{\partial \beta} = 0 \Rightarrow \frac{\partial}{\partial \beta} \sum_{j=1}^n [Y_j - \alpha - \beta X_j]^2 = -2 \sum_{j=1}^n [Y_j - \alpha - \beta X_j] X_j = 0$$

El sistema 2x2 resultante es:

$$\sum_{j=1}^n Y_j - \sum_{j=1}^n \alpha - \beta \sum_{j=1}^n X_j = 0 \Rightarrow n\bar{Y} - \alpha n - \beta n\bar{X} = 0 \Rightarrow \alpha n + \beta n\bar{X} = n\bar{Y}$$

$$\sum_{j=1}^n X_j Y_j - \sum_{j=1}^n \alpha X_j - \beta \sum_{j=1}^n X_j^2 = 0 \Rightarrow \alpha n\bar{X} + \beta \sum_{j=1}^n X_j^2 = \sum_{j=1}^n X_j Y_j$$

- Matricialmente el sistema anterior es: $\begin{pmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{j=1}^n X_j^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ \sum_{j=1}^n X_j Y_j \end{pmatrix}$ que se puede resolver aplicando inversa o con la Regla de Cramer o por sustitución. Resolviendo y verificando el mínimo, tenemos que

$$\beta = \frac{\sum_{j=1}^n X_j Y_j - n\bar{X}\bar{Y}}{\sum_{j=1}^n X_j^2 - n\bar{X}^2} \quad \text{y} \quad \alpha = \bar{Y} - \beta\bar{X} \quad \text{son los estimadores MCO de } \beta \text{ y } \alpha \text{ respectivamente.}$$

Estos mismos estimadores se obtienen aplicando Máxima Verosimilitud, bajo el supuesto adicional $\varepsilon \sim N(0, \sigma^2)$

Adicionalmente obtenemos:

- La predicción de $E(Y)$ es $\hat{Y} = \alpha + \beta X$ o $\hat{Y}_j = \hat{\alpha} + \hat{\beta} X_j$
- Aproximamos ε con $\hat{\varepsilon} = Y - \hat{Y}$, o sea $\hat{\varepsilon}_j = Y_j - \hat{Y}_j$
- σ^2 se estima con $\sigma^2 = S_{\hat{\varepsilon}}^2 = \frac{\sum_{j=1}^n \hat{\varepsilon}_j^2}{n-2} = \frac{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}{n-2}$, (aplicando máxima verosimilitud y corrigiendo el sesgo del estimador obtenido), que viene a ser "la distancia promedio al cuadrado entre un valor real de Y y su valor esperado estimado $E(Y)$ o predicción a partir de X .
- El Error Estándar o Error típico de estimación de $E(Y)$ es $\sigma \equiv S_{\hat{\varepsilon}} = \sqrt{\sigma^2}$. Es el "margen de error" asociado pronóstico: $Y = \hat{Y} \pm \sigma = \hat{\beta}_0 + \hat{\beta}_1 X \pm \sigma$ determina el "intervalo de estimación de Y "
- Como $\hat{\beta}$ es una estimación del verdadero β , tiene también un "error de estimación", cuyo cuadrado es

$$S_{\beta_1}^2 = \frac{\sigma^2}{\sum_{j=1}^n X_j^2 - n\bar{X}^2}, \text{ así que el Error Estándar de estimación de } \beta \text{ es}$$

$$e.e.(\beta) \equiv S_{\beta} = \sqrt{\frac{\sigma^2}{\sum_{j=1}^n X_j^2 - n\bar{X}^2}} = \sqrt{\frac{\sigma^2}{(n-1)S_X^2}}.$$

1.4 Ajuste del Modelo

“Ajuste del modelo” se refiere a medir qué tan bien representa el modelo a los datos.

Una medida natural es calcular la correlación entre el valor real de Y y su valor estimado \hat{Y} según el modelo; Esta correlación siempre es positiva (pues se espera que Y coincida con su estimación \hat{Y}) y debiera ser grande o al menos mediana. Se denota R y es definida como $R = r_{Y\hat{Y}}$. Un problema con esta medida es que aún siendo grande, no mide exactamente la coincidencia, pues la correlación de Pearson mide asociación, no necesariamente coincidencia.

Otra manera alternativa de medir el “ajuste”, siguiendo un enfoque explicativo, es *cuantificar el ajuste con la proporción de variabilidad en Y que se puede atribuir a X* :

A partir de $(Y_j - \bar{Y}) = (\hat{Y}_j - \bar{Y}) + (Y_j - \hat{Y}_j)$ se puede demostrar que

$$\underbrace{\sum_{j=1}^n (Y_j - \bar{Y})^2}_{\text{Variabilidad total en } Y} = \underbrace{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2}_{\text{Variabilidad debida a } X} + \underbrace{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}_{\text{Variabilidad residual}}$$

$$\underbrace{\sum_{j=1}^n (Y_j - \bar{Y})^2}_{\text{Variabilidad total en } Y} = SCT = \text{Suma de cuadrados total}$$

$$\underbrace{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2}_{\text{Variabilidad debida a } X} = SCR = \text{Suma de cuadrados de la regresión}$$

$$\underbrace{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}_{\text{Variabilidad residual}} = SCE = \text{Suma de cuadrados residual o del error}$$

Se escribe entonces $SCT = SCR + SCE$ y en este contexto se define:

El Coeficiente de Determinación R^2

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2} = \frac{\text{Variabilidad en } Y \text{ generada por } X}{\text{Variabilidad total en } Y}$$

R^2 mide la proporción de la variabilidad total en Y que es “explicada” o atribuible a las diferencias en la variable independiente X a través de la regresión. Es la proporción de diferencias en Y que se deben a las diferencias en X .

Equivalentemente, la cantidad $100R^2\%$ es el porcentaje de variabilidad (por extensión, también se dice “% de la varianza”) de Y explicada por el modelo. En economía un % de 80% es bastante bueno.

Nota: Estimación de la correlación de Pearson ρ_{XY}

Usando el límite en probabilidad Plim, se puede probar que una estimación consistente de la correlación de

Pearson ρ_{XY} entre dos variables aleatorias X e Y es $\hat{\rho}_{XY} \equiv r_{XY} = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{(n-1)S_Y S_X} = \frac{\sum_{j=1}^n (X_j Y_j - n\bar{X}\bar{Y})}{(n-1)S_Y S_X}$. r_{XY} se

llama “correlación muestral de Pearson”.

Con este dato se tienen dos resultados importantes:

Relación entre r_{XY} , β y R^2

(1) De la fórmula de r_{XY} , tenemos $\sum_{j=1}^n (Y_j - \bar{Y})(X_j - \bar{X}) = r_{XY}(n-1)S_X S_Y$ y así resulta $\beta = \frac{r_{XY}S_Y}{S_X}$

(2) Análogamente obtenemos $SCR = \beta \sum_{j=1}^n (Y_j - \bar{Y})(X_j - \bar{X}) = \frac{r_{XY}S_Y}{S_X} \times r_{XY}(n-1)S_X S_Y = r_{XY}^2(n-1)S_Y^2$, luego

$$R^2 = \frac{SCR}{SCT} = \frac{r_{XY}^2(n-1)S_Y^2}{(n-1)S_Y^2} = r_{XY}^2, \text{ que nos proporciona una nueva interpretación del coeficiente de}$$

correlación: r_{XY}^2 es la proporción de varianza de Y compartida con X . Si además se puede asumir relación de dependencia, r_{XY}^2 es proporción de varianza de Y generada o explicada por X .

Además como $SCT = SCR + SCE \Rightarrow SCE = SCT - SCR$ tenemos $SCE = (n-1)S_Y^2(1 - r_{XY}^2)$.

Ejemplo (cálculo manual de los estimadores)

En el caso inicial tomado para introducir el tema, podemos calcular la correlación: formemos una tabla auxiliar de cálculos:

Sucursal	1	2	3	4	5	6	7	8	9	10	Media	D. Estándar
Total de Autoavalúos	66.0	14.0	47.9	77.9	57.8	30.6	36.0	64.2	70.0	42.2	50.66	20.0725
Ventas	4.0	2.3	2.9	3.9	3.3	2.7	3.1	3.5	3.8	3.3	3.28	0.5473
XY	264	32.2	138.91	303.81	190.74	82.62	111.6	224.7	266.0	139.26	SumaXY	1753.84

Aplicando la fórmula del coeficiente de correlación r_{XY} se obtiene:

$r_{XY} = \frac{\sum_{i=1}^{12} X_i Y_i - n\bar{X}\bar{Y}}{(n-1)S_X S_Y} = \frac{1,753.84 - 10 \times 50.66 \times 3.28}{(10-1) \times 20.0725 \times 0.5473} = 0.9324$ que resulta positiva y grande (mayor que 0.9 que es el estándar para datos del área de Economía). O sea, hay una relación lineal suficientemente fuerte como para animarse a aplicar un modelo lineal $E(Y|X) = \alpha + \beta X$

Pasando a estimar parámetros:

$$\hat{\beta} = r_{XY} \frac{S_Y}{S_X} = 0.9324 \times \frac{0.5473}{20.0725} = 0.02542 \text{ y } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 3.28 - 0.02542 \times 50.66 = 1.9922$$

Nuestra “fórmula de pronóstico” es

$$\hat{Y} = 1.9922 + 0.02542X \Leftrightarrow \text{Ventas} = 1.922 + 0.02542 \text{ Autoavalúos}$$

Finalmente, tratemos de obtener una respuesta al problema de si se debe abrir la sucursal:

La condición para empezar la construcción de la nueva sucursal es que las ventas Y sean 3.5 millones o más. De los datos conseguidos por el economista se sabe que el total de autoavalúos de las viviendas en ese distrito es $X=28$ millones. Necesitamos aplicar la recta de regresión $\hat{Y} = a + bX$, reemplazando X por 28 en la ecuación y ver si $\hat{Y} > 3.5$, si así ocurriera, sí procedería la construcción de la nueva sucursal. Ya vimos que “fórmula de pronóstico” es

$$\hat{Y} = 1.9922 + 0.02542X \Leftrightarrow \text{Ventas} = 1.922 + 0.02542 \text{ Autoavalúos}$$

En el caso del distrito $X = \text{Autoavalúos} = 28 \Rightarrow \text{Ventas} = 1.922 + 0.02542 \times 28 = 2.6338$ millones < 3.5

A nivel de estimación puntual, la decisión es: No abrir la sucursal, no tendría suficientes ventas.

En el desarrollo anterior hemos aplicado cálculo manual para hacer las estimaciones básicas. Continuando con R, veamos ahora cómo se hace todo lo anterior y más con el programa estadístico R. Asumiendo que ya hemos aplicado los comandos R descritos en las páginas 2 a 3, ahora aplicamos el comando `attach(datos1)` que permite acceder a las variables del archivo `datos1` sin necesidad de especificar sus posiciones dentro de la matriz `d` de datos, como se hizo con el comando `plot`.

`attach(datos1)`


Para pedir a R que haga la regresión lineal simple usamos el comando “lm” (lm viene de linear models) como se muestra abajo:

modelo1=lm(Ventas~TotaldeAutoavaluos)

Al escribir **modelo1** delante del “=” le decimos a R que los resultados del análisis de regresión se guardarán en un **objeto** llamado **modelo1**.

lm(Ventas~TotaldeAutoavaluos) indica a R que la variable dependiente Y es Ventas y la independiente X es TotaldeAutoavaluos. Ingresada la orden, R la ejecuta pero no muestra nada. Para ver los resultados básicos se usa el comando “summary(modelo1)” que pide a R que muestre un resumen básico del análisis de regresión:

summary(modelo1)



```

> library (foreign)
> datos1 = read.spss(file.choose(),to.data.frame=TRUE)
Warning message:
In read.spss(file.choose(), to.data.frame = TRUE) :
  C:\Users\Arturo\Desktop\Supermercados.sav: Unrecognized record type 7, subtype 18 encountered in system file
> attach(datos1)
The following objects are masked from datos1 (pos = 3):

    Sucursal, TotaldeAutoavaluos, Ventas

> modelo1=lm(Ventas~TotaldeAutoavaluos)
> summary(modelo1)

Call:
lm(formula = Ventas ~ TotaldeAutoavaluos)

Residuals:
    Min       1Q   Median       3Q      Max
-0.30983 -0.11132 -0.05897  0.15161  0.32999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.992007   0.188548  10.565 5.62e-06 ***
TotaldeAutoavaluos 0.025424   0.003484   7.298 8.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2098 on 8 degrees of freedom
Multiple R-squared:  0.8694,    Adjusted R-squared:  0.8531
F-statistic: 53.26 on 1 and 8 DF,  p-value: 8.407e-05

> |

```

Como se muestra en la imagen de pantalla anterior, ejecutado el comando **summary(modelo1)** R muestra el resumen del análisis de regresión lineal simple:

Call:

lm(formula = Ventas ~ TotaldeAutoavaluos)

Primero R muestra el modelo aplicado, donde especifica que Ventas es la v. dependiente (o sea la **Y**) y TotaldeAutoavaluos es la v. independiente (la variable **X** del modelo)

Residuals:

Min	1Q	Median	3Q	Max
-0.30983	-0.11132	-0.05897	0.15161	0.32999

Bajo el encabezamiento Residuals, R muestra las estadísticas básicas de los residuos estimados $\hat{\epsilon}_i$, el valor mínimo Min (-0.30983), el primer cuartil 1Q(-0.11121), la mediana (0.15161), el tercer cuartil 3Q (0.15161) y el máximo Max (0.31999).

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.992007	0.188548	10.565	5.62e-06 ***
TotaldeAutoavaluos	0.025424	0.003484	7.298	8.41e-05 ***

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

Debajo del encabezamiento Coefficients: R muestra los valores estimados de α : $\hat{\alpha} = 1.992007$ que R llama Intercept y al costado su error estándar de estimación (Std. Error): $e.e(\hat{\alpha}) = 0.188546$ y el cociente

$t = \frac{\hat{\alpha}}{e.e(\hat{\alpha})} = \frac{1.992007}{0.188546} = 10.565$; Si α fuese 0, entonces $\hat{\alpha}$ debiera ser 0 o cercano a 0. O con el margen de error e.e. debiera estar dentro del intervalo $0 \pm e.e.$ o sea entre -0.188546 y 0.188546 pero en este caso

$\hat{\alpha} = 1.992007$ no sólo no es 0 sino que cae lejos de 0, el valor t cuán lejos de 0 está cayendo: $t = 10.565$ indica que $\hat{\alpha}$ está cayendo 10.565 veces más alejado de 0 de lo que podría caer sólo por azar y es un indicador de que el parámetro α realmente no es 0, sino que es positivo.

Análogamente, R muestra la estimación del parámetro β : $\hat{\beta} = 0.025424$ con un $e.e(\hat{\beta}) = 0.003484$ y

un t igual a $t = \frac{\hat{\beta}}{e.e(\hat{\beta})} = \frac{0.025424}{0.003484} = 7.298$, esto es $\hat{\beta}$ cae a la derecha de 0 7.298 veces más alejado de 0 de lo

que permitiría el margen de error (o del azar) si fuese el caso de ser $\beta = 0$. Esto es un indicador de que el parámetro β no es 0 y que en realidad es positivo. Esto último es lo más importante para este problema: Tenemos evidencia de que el modelo propuesto realmente es bueno y que hay, como se suponía por teoría económica, que sí hay una relación lineal directa entre Valor de Autoavalúos de las casas de los distritos y el Valor de las Ventas del supermercado de la cadena sito en el distrito. Por tanto sería permisible pronosticar el valor de las Ventas de un supermercado de la cadena usando el Valor total de Autoavalúos de las casas del distrito donde esté ubicado.

Al costado de los valores t , R calcula, para cada parámetro, la probabilidad de que ocurra un valor t mayor al valor absoluto del $|t|$ obtenido en la muestra:

En el caso de $\hat{\alpha}$: $P(t > 10.565) = 5.62 \times 10^{-6} = 0.00000562 \cong 0.00$; es decir no hay probabilidad de que siendo el parámetro $\alpha = 0$, ocurra un $\hat{\alpha} = 1.992007$ y por tanto un $t \geq 10.565$; *esto es evidencia de que realmente $\alpha \neq 0$ y más bien es evidencia a favor de $\alpha > 0$.*

En el caso de $\hat{\beta}$: $P(t > 7.298) = 8.41 \times 10^{-5} = 0.0000841 \cong 0.00$; es decir no hay probabilidad de que siendo el parámetro $\beta = 0$, ocurra un $\hat{\beta} = 0.025424$ y por tanto un $t \geq 7.298$; *esto es evidencia de que realmente $\beta \neq 0$ y más bien es evidencia a favor de $\beta > 0$.*

Como veremos más abajo, realmente R está sometiendo a prueba o contrastando, respectivamente, las hipótesis nulas $H_0: \alpha = 0$ y $H_0: \beta = 0$ usando la estadística t y su distribución que es $t - Student$ $t(k = n - 2)$ para cada parámetro.

Residual standard error: 0.2098 on 8 degrees of freedom
Multiple R-squared: 0.8694, Adjusted R-squared: 0.8531
F-statistic: 53.26 on 1 and 8 DF, p-value: 8.407e-05

Finalmente, R muestra estadísticas complementarias importantes:

Primero la estimación de la desviación estándar del residuo $\epsilon \sim N(0, \sigma^2)$ que es $\hat{\sigma} = 0.2098$ y que llama Residual estándar error que está asociado a una distribución χ^2 , pues se demuestra que, bajo el supuesto de homocedasticidad (varianzas homogéneas de residuos), se cumple que $W = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(k = n - 2)$

Luego R muestra el $R^2 = 0.8694$ del modelo que es bastante alto y en la siguiente línea el estadístico F de

Fisher $F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = (n - 2) \frac{R^2}{(1-R^2)}$. Se demuestra que si en la población $R^2 = 0$, entonces $F \sim F(1, n - 2)$ y

debiera ser cero aunque su valor esperado sería 1. En este ejemplo $F = 53.26$ que es 53.26 veces más grande que su valor esperado 1 y por tanto indica que realmente $R^2 > 0$. La probabilidad de que, siendo realmente $R^2 = 0$, hayamos obtenido un R^2 muestral igual a $R^2 = 0.8694$ y no 0 como hubiera debido ser, es $P(F \geq 53.26) = 8.407 \times 10^{-5} = 0.00008407 \cong 0.00$; *esto es evidencia de que en la población $R^2 > 0$, es decir el modelo aplicado es correcto.*

Para entender mejor las probabilidades anteriores necesitamos desarrollar algo más de teoría.

1.5 Contrastes en el Análisis de Regresión.

Si asumimos $\varepsilon_j \sim N(0, \sigma^2)$ entonces se cumplen las siguientes proposiciones que permiten realizar contrastes de hipótesis:

Proposición 1

En el modelo de Regresión Lineal Simple $Y_j = \alpha + \beta X_j + \varepsilon_j$ si además $\varepsilon_j \sim N(0, \sigma^2)$, entonces:

$$(a) \quad \beta \sim N(\beta_1, \sigma_{\beta_1}^2) \text{ donde } \sigma_{\beta_1}^2 = \frac{\sigma^2}{\sum_{j=1}^n X_j^2 - n\bar{X}^2}$$

$$(b) \quad W = (n-2)\sigma^2 / \sigma^2 \sim \chi^2(n-2). \text{ Es decir } SCE / \sigma^2 \sim \chi^2(n-2)$$

$$(c) \quad \text{Si } \beta = 0 \Rightarrow F = \frac{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2 / 1}{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2 / (n-2)} = \frac{SCR/1}{SCE/(n-2)} \sim F(k_1 = 1, k_2 = n-2)$$

Proposición 2

En el contexto anterior, se cumple que $t = \frac{(\beta_1 - \beta_1)}{S_{\beta_1}} \sim t(n-2)$ donde $S_{\beta_1} = \sqrt{\frac{\sigma^2}{\sum_{j=1}^n X_j^2 - n\bar{X}^2}}$

Contraste sobre R^2

De acuerdo con (c) si $\beta = 0$, entonces $R^2 = 0$ (pues el modelo se reduce a $Y_j = \alpha + \varepsilon_j$ y en este contexto $\hat{Y}_j = \bar{Y} \forall j$ y por tanto en la población $\rho_{XY} = 0$ de modo que $R^2 = (\rho_{XY})^2 = 0$), de modo que para contrastar $H_0: R^2 = 0$ vs $H_0: R^2 > 0$, si $F = \frac{SCR/1}{SCE/(n-2)} > F_{0.95}(1, n-2)$ se debe rechazar $H_0: R^2 = 0$, pues de ser cierta H_0 esperamos un $F = 0$ y si ocurre un F muestral “muy grande”, tenemos evidencia contra H_0 .

En este caso estamos usando un nivel de significación (o P(Error I)) de 0.05; Si usáramos un nivel de significación α (**no** confundir el nivel de significación α con el parámetro α del modelo!), se rechazará H_0 si $F = \frac{SCR/1}{SCE/(n-2)} > F_{1-\alpha}(1, n-2)$

Contraste general sobre β

Queremos contrastar $H_0: \beta = b$ donde b es un valor hipotético predeterminado, contra las distintas alternativas H_1 uni o bilaterales. Aplicando variantes del Teorema de Neyman-Pearson se llega a:

Como en general $t = \frac{(\beta - \beta_1)}{S_{\beta_1}} \sim t(n-2)$ entonces si $H_0: \beta_1 = b$ es cierta, el estadístico $t = \frac{(\beta_1 - b)}{S_{\beta_1}}$ debe

seguir la distribución t -Student y se espera que su valor sea cero o cercano a cero.

Entonces un valor de t muy alejado de cero es buena razón para rechazar $H_0: \beta = b$.

Considerando “muy alejados” de cero a los valores de t que tienen probabilidad menor que un nivel de significación α (**no** confundir con el intercepto del modelo!) predeterminado, llegamos a la siguiente tabla de decisiones:

Hipótesis Nula	Hipótesis Alterna	Rechazar H_0 si	Tipo de contraste
$H_0: \beta = b$	$H_1: \beta > b$	$t > t_{1-\alpha}$	Unilateral derecho
	$H_1: \beta < b$	$t < -t_{1-\alpha}$	Unilateral izquierdo
	$H_1: \beta \neq b$	$ t > t_{1-\alpha/2}$	Bilateral
$t_{1-\alpha}$ y $t_{1-\alpha/2}$ percentiles $1-\alpha$ y $1-\alpha/2$ de la tabla t de Student: $t(k=n-2)$			

Contraste de coeficiente nulo $H_0 : \beta = 0$

Contrastar $H_0 : \beta = 0$ equivale a decir que Y no depende de X . El estadístico de contraste es $t = \beta / S_{\beta_1}$; el procedimiento para rechazar H_0 es el mismo descrito en la tabla anterior haciendo $b=0$ y con $t = \beta / S_{\beta} \sim t(n-2)$. Este es el contraste que por default suelen hacer los programas estadísticos como R.

Hipótesis Nula	Hipótesis Alterna	Rechazar H_0 si	Tipo de contraste
$H_0 : \beta = 0$	$H_1 : \beta > 0$	$t > t_{1-\alpha}$	Unilateral derecho
	$H_1 : \beta < 0$	$t < -t_{1-\alpha}$	Unilateral izquierdo
	$H_1 : \beta \neq 0$	$ t > t_{1-\alpha/2}$	Bilateral
$t_{1-\alpha}$ y $t_{1-\alpha/2}$ percentiles $1-\alpha$ y $1-\alpha/2$ de la tabla t de Student: $t(k=n-2)$			

En nuestro curso, como usamos siempre un nivel de significación $\alpha=0.05$, la tabla anterior se convierte en:

Hipótesis Nula	Hipótesis Alterna	Rechazar H_0 si	Tipo de contraste
$H_0 : \beta = 0$	$H_1 : \beta > 0$	$t > t_{0.95}$	Unilateral derecho
	$H_1 : \beta < 0$	$t < -t_{0.95}$	Unilateral izquierdo
	$H_1 : \beta \neq 0$	$ t > t_{0.975}$	Bilateral
$t_{0.95}$ y $t_{0.975}$ percentiles 0.95 y 0.975 de la tabla t de Student: $t(k=n-2)$			

Notas:

- (1) Para el parámetro α hay un contraste de hipótesis similar, pero no es muy usado.
- (2) Programas estadísticos como R, realizan automáticamente el contraste bilateral $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$, pero no muestran el percentil $t_{1-\alpha/2}$ de la distribución t -Student, sino la “significación” (o “valor p ” según el programa estadístico) que es la probabilidad de obtener un valor $|t|$ mayor o igual que el valor absoluto del calculado en la muestra. Si esta probabilidad es menor que 0.05, entonces se rechaza $H_0 : \beta = 0$ y se concluye que la variable X sí tiene efecto sobre la v . dependiente Y .
- (3) Como $\beta = \frac{r_{XY}S_Y}{S_X}$ y es posible pasar de un coeficiente a otro, entonces *el contraste de $H_0 : \beta = 0$ es totalmente equivalente al de $H_0 : \rho_{XY} = 0$ y por tanto al de $H_0 : R^2 = 0$. La consecuencia es que las hipótesis $H_0 : \beta = 0$, $H_0 : \rho_{XY} = 0$ y $H_0 : R^2 = 0$ son todas equivalentes en el modelo simple.*

Regresando a los resultados obtenidos con R:

Como ya se dijo, primero R muestra las estimaciones de los parámetros con los respectivos errores estándar de estimación (*e.e*) y al costado las estadísticas *t-Student*. Como R no sabe cuál es el nivel de significación o $P(\text{Error I})$ que deseamos usar, realiza un contraste **bilateral** pero en lugar de mostrar el percentil $1 - \frac{\alpha}{2}$ de la tabla *t-Student*, esto es $t_{1-\frac{\alpha}{2}}(k = n - 2)$, calcula la probabilidad de un **t** mayor que el **valor absoluto del t** obtenido en la muestra, si esta probabilidad es menor que 0.05 entonces podemos rechazar la hipótesis nula H_0 de que el correspondiente parámetro es 0 y no hay necesidad de usar la tabla *t-Student*.

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.992007    0.188548  10.565 5.62e-06 ***
TotaldeAutoa    0.025424    0.003484   7.298 8.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los tres *** que muestra R quieren decir que la probabilidad de Error I es menor que 0.001

Y con respecto al contraste de $H_0: R^2 = 0$, R muestra el estadístico F y con el nombre de **p-value** la probabilidad de Error I que se tendría si se rechaza H_0 . Si esta probabilidad es menor que 0.05, se puede rechazar H_0 con toda confianza.

Residual standard error: 0.2098 on 8 degrees of freedom
Multiple R-squared: 0.8694, Adjusted R-squared: 0.8531
F-statistic: 53.26 on 1 and 8 DF, p-value: 8.407e-05

En el caso del modelo de regresión simple, el contraste F es innecesario, basta con el contraste para β .

2. Modelo de regresión lineal múltiple.

2.1 Uso del modelo de regresión lineal múltiple

Como ya se mencionó, el modelo de regresión múltiple es una extensión del modelo lineal simple, donde el valor esperado de la variable dependiente Y es función lineal de p variables independientes o “explicativas” X_1, X_2, \dots, X_p vía el modelo econométrico $E(Y|X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$. El modelo de datos correspondiente es $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

Este modelo asume una relación de proporcionalidad entre los cambios en cada variable independiente X_j y la variable dependiente Y , esto es $\frac{\partial E(Y|X_1, X_2, \dots, X_p)}{\partial X_j} = \beta_j$ y que los efectos de los cambios son aditivos, de ahí el modelo lineal:

$\beta_j =$ Cambio en $E(Y|X_1, X_2, \dots, X_p)$ (o cambio **esperado** en Y) cuando X_j **aumenta** en 1

2.2 Supuestos y parámetros del modelo lineal múltiple

Dado el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$, evaluado en una muestra aleatoria de n casos, en general se cumple $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$ donde ε_i representa el efecto mezclado de otras variables no controladas, para el cual se asume que se comporta como “el azar”, como variable aleatoria: $\varepsilon_i \sim N(0, \sigma^2)$ habiendo independencia entre observaciones.

Los supuestos son los supuestos clásicos:

- (a) $E(\varepsilon_i) = 0 \quad \forall i$
- (b) $V(\varepsilon_i) = \sigma^2 = \text{constante} \quad \forall i$
- (c) $\rho_{\varepsilon_i \varepsilon_j} = 0 \quad \forall i \neq j$
- (d) Las X_1, X_2, \dots, X_p son de valores predeterminados, o sea $\rho_{X_i X_k} = 0$ para todo par de variables independientes X_j y X_k
- (e) El modelo está “bien especificado”.

Los supuestos implican que $E(Y|X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ y $V(\varepsilon_i) = \sigma^2$

Parámetros del modelo.

Son $(p + 2)$ parámetros, a saber:

- Las constantes $\beta_0, \beta_1, \beta_2, \dots, \beta_p$. β_0 es llamada “*constante*” o “*intercepto*” en programas como Excel, SPSS o Stata, y en otros textos, “*ordenada en el origen*” o “*intercepto*” de la recta. En general es el valor esperado o promedio de Y cuando todas las X_i son cero. Pero no siempre esto último tiene interpretación práctica y eso sólo un parámetro de ajuste. Hasta aquí son $(p + 1)$ parámetros. Cada constante β_j mide en cuántas unidades se espera que varíe Y cuando X_j aumenta en 1 unidad, manteniendo las otras v. independientes constantes. Si $\beta_i > 0$ la relación de Y con X_i es “directa”, si $\beta_i < 0$ la relación es “inversa”. Estos suelen ser parámetros económicos (derivados de la “especificación de un modelo económico”) y convertidos en parámetros estadísticos con fines básicamente instrumentales.
- σ^2 es la varianza del azar representado por el residuo aleatorio ε ; σ es la variación promedio arriba o debajo de $E(Y|X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$. σ^2 es el parámetro “de ruido”, es parámetro estadístico derivado del modelo de datos, no suele tener interpretación económica.

2.3 Estimación de parámetros.

El método para hallar las mejores estimaciones de los parámetros es el de Mínimos Cuadrados (o máxima verosimilitud que bajo el supuesto de normalidad de residuos coincide con los MCO). Las estimaciones serán denotadas $\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_p$

- La estimación de Y_i es $\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_p X_{ip}$
- La estimación del error o residuo ε_i es $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$
- La estimación de σ^2 es $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 / (n - p - 1) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n - p - 1)$
- El Error Estándar o típico de estimación es $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ y mide el “margen de error” asociado al uso del modelo como base para el pronóstico de valores de Y .

2.4 Ajuste del Modelo.

Como en el caso lineal simple, debemos medir el efecto conjunto de las variables $\{X_j\}$ en Y . Lo anterior se hace notando que las diferencias en los valores de Y_i se deben en parte a las diferencias en las variables independientes y en parte al azar, pues $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$

- La “variabilidad total en Y ” es medida por $(n-1)S_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$.

Esta cantidad, denotada **SCT** se llama Suma de Cuadrados Total y mide cuán diferentes son los casos en la v.d. Y : cuanto mayores sean las diferencias en Y entre los n casos, mayor será **SCT**. Ya sabemos que SCT es el numerador de la varianza muestral S_Y^2 .

- La “variabilidad en \hat{Y} ” es $\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2$.

Esta cantidad llamada Suma de Cuadrados de la Regresión y denotada **SCR**, mide en total qué tan diferentes son los casos, debido a los distintos valores que tienen en las variables independientes $\{X_j\}$.

O sea es la variabilidad originada en las variables independientes $X_1, X_2, X_3, \dots, X_p$

- La “variabilidad residual originada por ε ” es $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$.

Esta cantidad se llama Suma de Cuadrados Residual o del Error, se denota **SCE** y mide las diferencias entre los n casos que se deberían al azar, o sea a razones fortuitas y no originadas por las variables independientes.

Partiendo de $(Y_i - \bar{Y}) = (Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$ se puede probar formalmente que se

cumple $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ o equivalentemente **SCT = SCR + SCE**.

A partir de lo anterior:

Para medir el ajuste global de los datos al modelo

Usamos el Coeficiente de Determinación definido por:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2} = \frac{\text{Variabilidad originada por } X_1, X_2, \dots, X_p}{\text{Variabilidad total en } Y}$$

que siempre está entre 0 y 1 y mide la proporción de la variabilidad total en Y que es “explicada” o atribuible a las diferencias en las variables independientes X_1, X_2, \dots, X_p a través de la regresión. Es la proporción de diferencias en Y que se deben a las diferencias en X_1, X_2, \dots, X_p .

El número $100R^2\%$ es “el porcentaje de variabilidad (por extensión, también se dice “porcentaje de la varianza”) de Y explicada por el modelo”. En economía un % de 80% es bastante bueno.

El Coeficiente de Correlación Múltiple $R = \sqrt{R^2}$. Es la correlación de Pearson entre Y_i y su “valor pronosticado” según el modelo: $\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_p X_{ip}$, o sea $R = r_{Y\hat{Y}}$, y se espera que sea un valor alto. Si $R = 1$, las predicciones son exactas y si $R = 0$, las predicciones son malas. Este coeficiente debe tener un valor alto, cercano a 1. En nuestro curso pediremos que esté alrededor de 0.9 o más para considerarlo alto y como mínimo que sea mayor o igual que 0.5.

2.5 Contrastes en el Análisis de Regresión Múltiple

De manera análoga al modelo lineal simple, en el caso múltiple hay una serie de contrastes que podemos hacer. Nos centraremos en los más básicos, dejando otros para los cursos de econometría. Estos contrastes se basan en las siguientes proposiciones, enunciadas sin demostrar pero que usaremos libremente.

Proposición 1

En el modelo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$, si $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$ es cierta (o equivalentemente, si es $H_0 : R^2 = 0$ cierta), entonces se cumple que

$$F = \frac{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2 / p}{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2 / (n-p-1)} = \frac{SCR/p}{SCE/(n-p-1)} \sim F(k_1 = p, k_2 = n - p - 1). \text{ Si fuera cierta } H_0, F \text{ sería cero y si}$$

en la muestra ocurre un F mucho mayor que cero, eso es motivo para rechazar $H_0 : R^2 = 0$ y aceptar $H_1 : R^2 > 0$, o equivalentemente aceptar que $\beta_i \neq 0$ para al menos una de las v. independientes X_i .

Proposición 2

En el modelo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$, si además $\varepsilon_i \sim N(0, \sigma^2)$, entonces se cumplen:

- $\hat{\beta}_i \sim N(\beta, \sigma_{\hat{\beta}_i}^2)$ donde $\sigma_{\hat{\beta}_i}^2$ es la varianza del estimador MCO (o de Máxima verosimilitud pues ambos coinciden bajo los supuestos clásicos).
- $W = \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$, es decir $\frac{SCE}{\sigma^2} \sim \chi^2(n-p-1)$

Proposición 3

En el contexto anterior, se cumple que $t = \frac{(\hat{\beta}_i - \beta)}{S_{\hat{\beta}_i}} \sim t(k = n - p - 1)$, donde $S_{\hat{\beta}_i}$ es el error estándar de estimación (muestral) de $\hat{\beta}_i$. Si $H_0 : \beta_i = 0$ es cierta tenemos que $t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} \sim t(k = n - p - 1)$.

Contraste de significación conjunta

Es el contraste del modelo como un todo. Se somete a prueba la hipótesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$ que equivale a decir que no hay efecto de las variables independientes sobre Y . Este contraste equivale a contrastar la hipótesis $H_0 : R^2 = 0$ donde R^2 es el coeficiente de correlación múltiple poblacional.

De acuerdo con la proposición 1, si $\beta_1 = \beta_2 = \dots = \beta_p = 0$ para contrastar $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ hay que ver si el F muestral, $F = \frac{SCR/p}{SCE/(n-p-1)}$, es “muy grande”. Si trabajamos con un nivel de significación (o probabilidad de Error I) $\alpha = 0.05$, se debe rechazar H_0 si $F = \frac{SCR/p}{SCE/(n-p-1)} > F_{0.95}(1, n-p-1)$ pues se considera “grande” un F muestral que esté arriba del 5% de valores más grandes de la distribución F de Fisher con los correspondientes grados de libertad. Si usáramos un nivel de significación α , se rechazará H_0 si $F = \frac{SCR/p}{SCE/(n-p-1)} > F_{1-\alpha}(1, n-p-1)$, donde es el percentil $(1 - \alpha)$ de la distribución F de Fisher.

Si se rechaza esta primera hipótesis, tiene sentido dar el siguiente paso, que es evaluar cuáles de las v. independientes tienen efecto. Esto se hace para cada v.i. X_i mediante la correspondiente estadística t -Student.

Contraste de significación para cada v. independiente del modelo (para cada β_i)

Se apoya en la proposición 3. Recordando que $\hat{\beta}_i$ es una estimación de β_i con un error estándar de estimación $e.e. = S_{\hat{\beta}_i}$, si tenemos la hipótesis nula $H_0: \beta_i = 0$, de la proposición 3 se espera un valor de $t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$ cero o cercano a cero y si t cae muy alejado de cero tenemos razón para rechazar $H_0: \beta_i = 0$. Considerando “muy alejados” de cero a los valores de t que tienen probabilidad “casi cero” o sea, menor que un nivel α predeterminado, llegamos a la metodología siguiente:

Dada la muestra y calculadas estimaciones y estadísticas, para contrastar $H_0: \beta_i = 0$

Hipótesis Nula	Hipótesis Alterna	Rechazar H_0 si	Tipo de contraste
$H_0: \beta_i = 0$	$H_1: \beta_i > 0$	$t > t_{1-\alpha}$	Unilateral derecho
	$H_1: \beta_i < 0$	$t < -t_{1-\alpha}$	Unilateral izquierdo
	$H_1: \beta_i \neq 0$	$ t > t_{1-\alpha/2}$	Bilateral
$t_{1-\alpha}$ y $t_{1-\alpha/2}$ son los percentiles $1-\alpha$ y $1-\alpha/2$ de la distribución <i>t-Student</i> $t(n-p-1)$			

Nota:

Programas estadísticos como R, realizan automáticamente el contraste bilateral $H_0: \beta_i = 0$ vs $H_1: \beta_i \neq 0$, pero no muestran el percentil $t_{1-\alpha/2}$ de la distribución t-Student, sino la “significación” (o “valor p” según el programa estadístico) que es la probabilidad de obtener un valor t mayor o igual que el valor absoluto del calculado en la muestra. Si esta probabilidad es menor que 0.05, entonces se rechaza $H_0: \beta_i = 0$ y se concluye que la correspondiente variable X_i sí tiene efecto sobre la v. dependiente Y .

Ejemplo 2 (Regresión lineal múltiple con R)

Una empresa comercializadora de ropa tiene un equipo de vendedoras comisionistas y desea hacer un estudio que le permita determinar si el ingreso semanal por comisión (en decenas de unidades monetarias) obtenido por la venta de ropa de invierno (Y = variable IngSem) depende del número mensual de horas trabajadas (X_1 = HorasW), la edad (en años) de sus vendedoras (X_2 = Edad) y los meses de experiencia en algún trabajo similar (X_3 = ExpMeses).

Para llevar a cabo este estudio se seleccionó aleatoriamente una muestra de $n = 45$ vendedoras y se registró para cada una las citadas variables. Los datos están en el archivo Vendedoras.sav en formato del paquete estadístico SPSS.

Bajo el supuesto de proporcionalidad entre la variable dependiente Y y las tres variables independientes, se desea ajustar un modelo de regresión lineal múltiple para determinar cuáles variables independientes de las tres propuestas tienen efecto real sobre el ingreso.

El modelo es $IngSem = \beta_0 + \beta_1 HorasW + \beta_2 Edad + \beta_3 ExpMeses + \varepsilon$ y se asumen los supuestos clásicos.

Además de estimar el modelo y realizar las respectivas pruebas de hipótesis, se quiere verificar el cumplimiento de los supuestos estadísticos subyacentes al modelo de regresión. Aplique el paquete estadístico R para hacer este análisis.

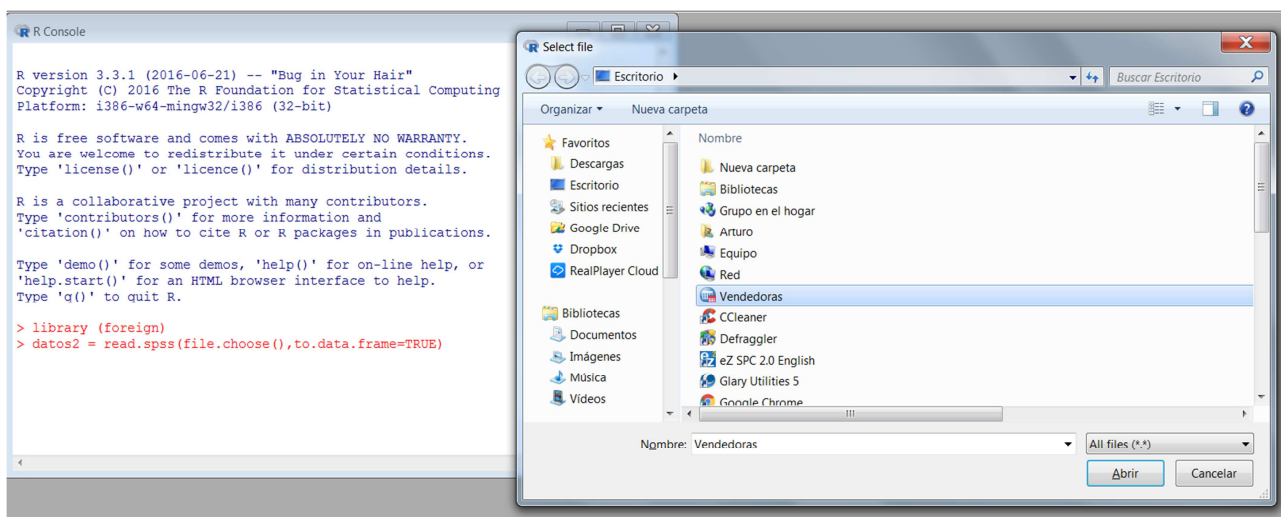
Solución:

Como en el ejemplo 1, en este caso los datos están digitados en SPSS y para leer los datos con R cargamos la librería “foreign”:

library(foreign)

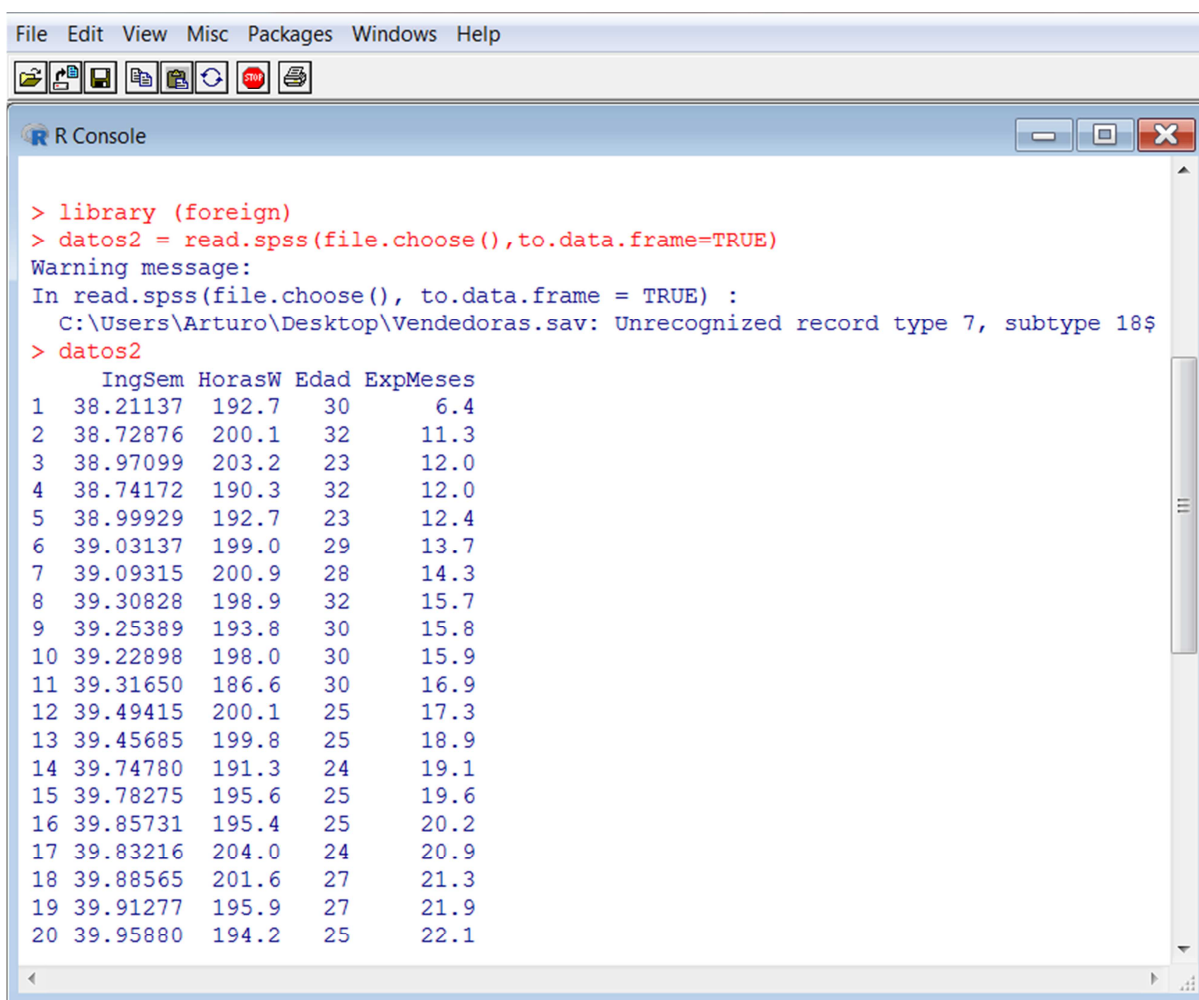
Luego para que R lea el archivo Vendedoras.sav, y lo llame **datos2** usamos el comando “**read.spss(file.choose(),to.data.frame=TRUE)**” que indica a R que va a leer un archivo SPSS y además nos permite buscar y seleccionar el archivo donde esté ubicado en la computadora. Antes del comando read.spss decimos a R el nombre que usará para identificar el archivo de datos ya convertido a formato de R, escribiéndolo antes del comando read.spss seguido de =, como se muestra abajo, teniendo cuidado de mantener el espacio en blanco entre datos2, = y read.spss. Al escribir el comando e ingresarlo, R abre una ventana para que podamos ubicar el archivo (que en este ejemplo 2 está en el escritorio de Windows).

```
datos2 = read.spss(file.choose(),to.data.frame=TRUE)
```



seleccionamos el archivo Vendedoras.sav con “Abrir” y R lee el archivo. Como en el ejemplo 1, en este caso y solo para ver los nombres de las variables ingresamos la orden “datos2” y R nos muestra los datos (en la figura de abajo sólo mostramos los 20 primeros casos, pero son $n = 45$ casos):

datos2



Ahora aplicamos el comando “attach(datos2)” que permite acceder a las variables del archivo datos2 sin necesidad de especificar sus posiciones dentro de la matriz de datos:

attach(datos2)

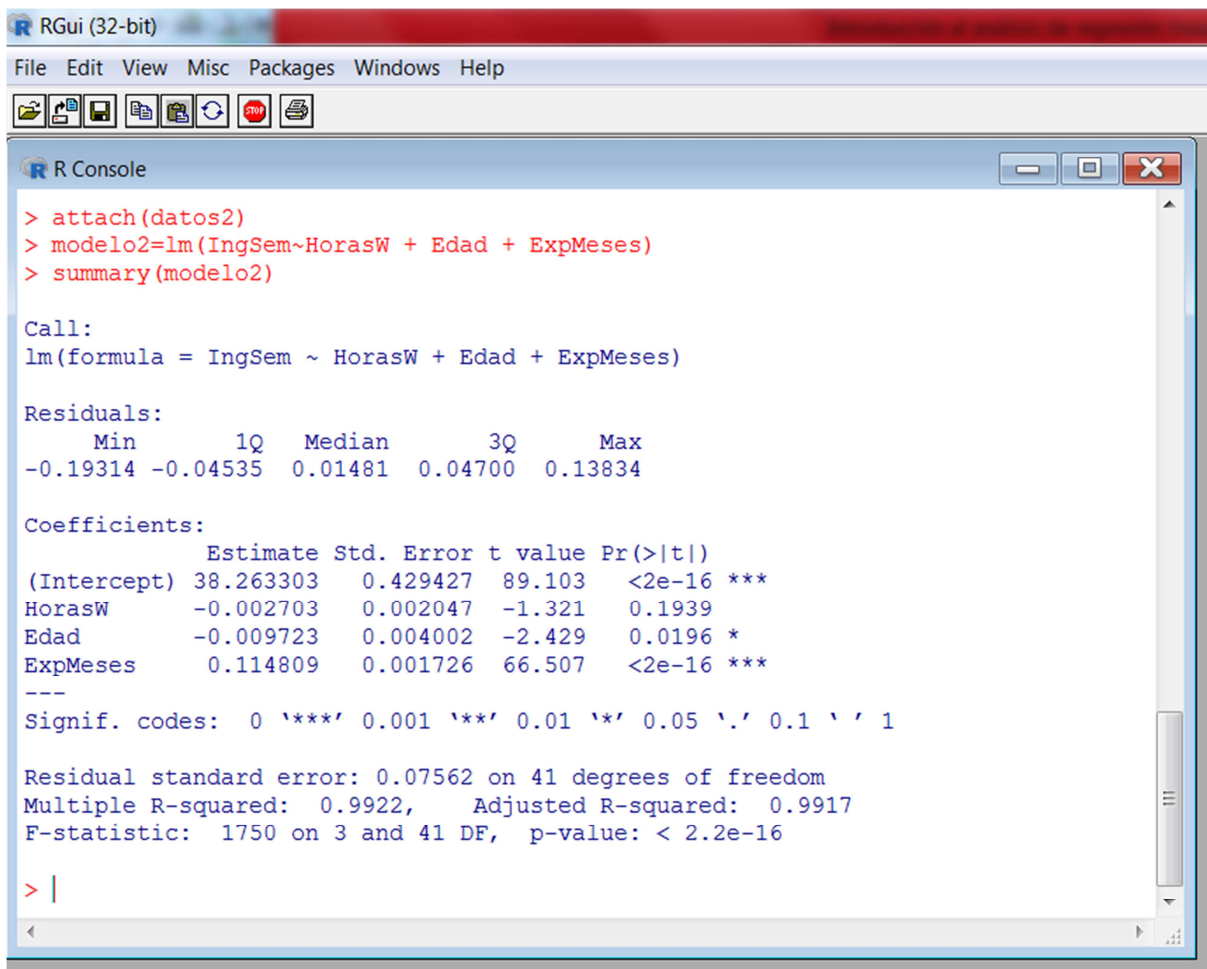
Para pedir a R que haga la regresión lineal múltiple con usamos el comando “lm” (lm viene de linear models) como se muestra abajo:

modelo2=lm(IngSem~HorasW + Edad + ExpMeses)

Al escribir **modelo2** delante del “=” le decimos a R que los resultados del análisis de regresión se guardarán en un **objeto** llamado **modelo2**.

lm(IngSem~HorasW + Edad + ExpMeses) dice a R que la variable dependiente Y es IngSem y las variables independientes son $p = 3$: $X_1 = \text{HorasW}$, $X_2 = \text{Edad}$ y $X_3 = \text{ExpMeses}$. Ingresada la orden, R la ejecuta y para ver los resultados básicos se usa el comando “summary(modelo2)” que pide a R que muestre un resumen básico del análisis de regresión:

summary(modelo2)



```
> attach(datos2)
> modelo2=lm(IngSem~HorasW + Edad + ExpMeses)
> summary(modelo2)

Call:
lm(formula = IngSem ~ HorasW + Edad + ExpMeses)

Residuals:
    Min       1Q   Median       3Q      Max
-0.19314 -0.04535  0.01481  0.04700  0.13834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.263303   0.429427  89.103  <2e-16 ***
HorasW      -0.002703   0.002047  -1.321   0.1939
Edad        -0.009723   0.004002  -2.429   0.0196 *
ExpMeses     0.114809   0.001726  66.507  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07562 on 41 degrees of freedom
Multiple R-squared:  0.9922,    Adjusted R-squared:  0.9917
F-statistic: 1750 on 3 and 41 DF,  p-value: < 2.2e-16

> |
```

Como se ve en la imagen anterior, R muestra las estimaciones de los parámetros del modelo junto con los contrastes de hipótesis en un formato similar al del ejemplo 1:

```

Call:
lm(formula = IngSem ~ HorasW + Edad + ExpMeses)

Residuals:
    Min       1Q   Median       3Q      Max
-0.19314 -0.04535  0.01481  0.04700  0.13834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.263303   0.429427  89.103  <2e-16 ***
HorasW      -0.002703   0.002047  -1.321   0.1939
Edad        -0.009723   0.004002  -2.429   0.0196 *
ExpMeses     0.114809   0.001726  66.507  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07562 on 41 degrees of freedom
Multiple R-squared:  0.9922,    Adjusted R-squared:  0.9917
F-statistic: 1750 on 3 and 41 DF,  p-value: < 2.2e-16

```

Dejando de lado por el momento las estadísticas de los residuos, vemos que R muestra los valores estimados de los coeficientes Intercepto (β_0) y de las v. independientes del modelo (los β_i) junto con sus errores estándar estimados (e. e. = S_{β_i}) y las correspondientes estadísticas *t-Student* y las probabilidades necesarias para hacer los contrastes de las respectivas hipótesis nulas $H_0: \beta_i = 0$ vs $H_0: \beta_i \neq 0$. Luego R presenta el $R^2 = 0.9922$ (que resulta bastante grande) y en la siguiente línea las estadísticas relativas al contraste del R^2 poblacional

Analizaremos los resultados empezando con el contraste global de significación, que R muestra al final:

Contraste de significación conjunta $H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$ o $H_0: R^2 = 0$

Este contraste se ubica en la línea final

```

Residual standard error: 0.07562 on 41 degrees of freedom
Multiple R-squared:  0.9922,    Adjusted R-squared:  0.9917
F-statistic: 1750 on 3 and 41 DF,  p-value: < 2.2e-16

```

R presenta la estadística *F* muestral **F-statistic: 1750** e indica sus grados de libertad: $p = 3$ y $(n - p - 1) = 45 - 3 - 1 = 41$. Recordemos que si H_0 fuera cierta *F* tendría que ser cero o en todo caso cercano a su valor esperado que es 1, pero en este caso es 1750, bastante alejado de cero y muy a la derecha de su valor esperado 1. Al final de la línea R muestra la **probabilidad de que, siendo realmente $R^2 = 0$, hayamos obtenido un R^2 muestral igual a $R^2 = 0.9922$ y no 0 como hubiera debido ser. Esta probabilidad es $P(F > 1750) = 2.2 \times 10^{-16} \cong 0.00$; o sea no hay probabilidad de que $R^2 = 0$** *Lo anterior es evidencia de que en la población $R^2 > 0$, es decir el modelo aplicado es correcto y se puede rechazar $H_0: R^2 = 0$, o equivalentemente, se puede rechazar la hipótesis $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. Concluimos que al menos una de las $p = 3$ v. independientes tiene coeficiente no nulo y sí tiene relación con la v. dependiente Ingreso semanal (IngSem)*

Contraste de significación para cada v. independiente del modelo $H_0: \beta_i = 0$ vs $H_0: \beta_i \neq 0$

R presenta estos contrastes bajo el encabezamiento **Coefficients** como se muestra abajo:

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.263303   0.429427  89.103  <2e-16 ***
HorasW      -0.002703   0.002047  -1.321   0.1939
Edad        -0.009723   0.004002  -2.429   0.0196 *
ExpMeses     0.114809   0.001726  66.507  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

R muestra las estimaciones de los parámetros con los respectivos errores estándar de estimación (*e.e*) y al costado las estadísticas *t-Student*. Como R no sabe cuál es el nivel de significación o $P(\text{Error I})$ que deseamos usar, realiza un contraste **bilateral**, pero en lugar de mostrar el percentil $1 - \frac{\alpha}{2}$ de la tabla *t-Student*, esto es $t_{1-\frac{\alpha}{2}}(k = n - p - 1)$, calcula la probabilidad de un *t* mayor que el **valor absoluto del t** obtenido en la muestra, si esta probabilidad es menor que 0.05 entonces podemos rechazar la hipótesis nula H_0 de que el correspondiente parámetro es 0 y no hay necesidad de usar la tabla *t-Student*.

En el caso de β_0 , R muestra el valor estimado de $\beta_0, \widehat{\beta}_0 = 38.263303$ que R llama Intercept y al costado su error estándar de estimación (Std. Error) $e.e(\widehat{\beta}_0) = 0.429427$ y luego el cociente $t = \frac{\widehat{\beta}_0}{e.e(\widehat{\beta}_0)} = \frac{38.263303}{0.429427} = 89.103$: Si β_0 fuese 0, entonces $\widehat{\beta}_0$ debiera ser 0 o cercano a 0 y por tanto *t* debiera caer cerca de 0. Sin embargo en la muestra ocurrió $t = 89.103$, que está bastante lejos de 0. Al costado del valor *t*, para contrastar $H_0: \beta_0 = 0$ vs $H_0: \beta_0 \neq 0$, R calcula la probabilidad de que ocurra un valor *t* mayor al valor absoluto del *t* obtenido en la muestra, $P(t > 89.103) = 2 \times 10^{-16} \cong 0.00$; es decir no hay probabilidad de que siendo el parámetro $\beta_0 = 0$, ocurra un $\widehat{\beta}_0 = 38.263303$ y por tanto un $t = 89.103$; *esto es evidencia de que realmente $\beta_0 \neq 0$. Por tanto rechazamos $H_0: \beta_0 = 0$.*

En el caso de X_1 =Horas de trabajo (HorasW en el archivo de datos), R muestra el valor estimado de β_1 que es $\widehat{\beta}_1 = -0.002703$ y su error estándar de estimación (Std. Error) $e.e(\widehat{\beta}_1) = 0.002047$; el cociente o estadística *t-Student* es $t = \frac{\widehat{\beta}_1}{e.e(\widehat{\beta}_1)} = \frac{-0.002703}{0.002047} = -1.321$ Si β_1 fuese 0, entonces $\widehat{\beta}_1$ debiera ser 0 o cercano a 0 y por tanto *t* debiera caer cerca de 0. En este caso *t* no es cero, pero para saber si ha caído “lejos” de cero es mejor hacer el contraste sobre β_1 : $H_0: \beta_1 = 0$ vs $H_0: \beta_1 \neq 0$. R calcula la probabilidad de que ocurra un valor *t* mayor al valor absoluto del *t* obtenido en la muestra, $P(t > 1.321) = 0.1939 > 0.05$; esta probabilidad no es cercana a cero (en particular no es menor que 0.05 que es el punto de corte, o nivel de significación, que usamos en nuestro curso, sino al contrario, es mayor) y por tanto no hay evidencia como para rechazar la hipótesis nula $H_0: \beta_1 = 0$. Se concluye que la variable independiente Horas de trabajo no tiene relación con el Ingreso semanal.

Pasando a X_2 =Edad, la estimación de su coeficiente β_2 es $\widehat{\beta}_2 = -0.009723$ con un error estándar de estimación $e.e(\widehat{\beta}_2) = 0.004002$ y una estadística *t-Student* $t = \frac{\widehat{\beta}_2}{e.e(\widehat{\beta}_2)} = \frac{-0.009723}{0.004002} = -2.429$; luego R hace el contraste de $H_0: \beta_2 = 0$ vs $H_0: \beta_2 \neq 0$ y para ello calcula la probabilidad de que ocurra un valor *t* mayor al valor absoluto del *t* obtenido en la muestra, $P(t > 2.429) = 0.0196$ que es lo bastante pequeña (menor que 0.05) como para considerarla cero, es decir, es decir no hay probabilidad de que siendo el parámetro $\beta_2 = 0$, ocurra un $t = -2.429$; *esto es evidencia de que realmente $\beta_2 \neq 0$. Por tanto rechazamos $H_0: \beta_2 = 0$.* Se concluye que la variable independiente Edad si tiene relación con el Ingreso semanal y como el coeficiente es negativo esta relación es inversa: a mayor edad menor ingreso.

Finalmente para la variable X_3 = Experiencia en meses (ExpMeses), la estimación de su coeficiente β_3 es $\widehat{\beta}_3 = 0.114809$ con un error estándar de estimación $e.e(\widehat{\beta}_3) = 0.001726$ y una estadística *t-Student* $t = \frac{\widehat{\beta}_3}{e.e(\widehat{\beta}_3)} = \frac{0.114809}{0.001726} = 66.507$ que cae bastante lejos de cero. El contraste de $H_0: \beta_3 = 0$ vs $H_0: \beta_3 \neq 0$ lo hace R calculando la probabilidad de que ocurra un valor *t* mayor al valor absoluto del *t* obtenido en la muestra, $P(t > 66.507) = 2 \times 10^{-16} \cong 0.00 < 0.05$: no hay probabilidad de que siendo el parámetro $\beta_3 = 0$, ocurra en la muestra un $t = 66.507$ y por tanto hay evidencia como para rechazar la hipótesis nula $H_0: \beta_3 = 0$. Se concluye que la variable independiente Experiencia si tiene relación con el Ingreso semanal y como el coeficiente es positivo, esta relación es directa: a mayor experiencia mayor ingreso.

En resumen, de las tres variables independientes postuladas por el modelo como variables explicativas del Ingreso semanal, sólo dos: Edad y Experiencia muestran relación estadísticamente significativa con el Ingreso. El $R^2=0.9922$ indica que el 99.22% de diferencias en el ingreso de las vendedoras se deberían a diferencias en Edad y/o la Experiencia.

2.6 Verificación de supuestos.

El modelo explicativo propuesto asume los supuestos clásicos, a saber:

- (a) $E(\varepsilon_i) = 0 \quad \forall i$
- (b) $V(\varepsilon_i) = \sigma^2 = \text{constante} \quad \forall i$
- (c) $\rho_{\varepsilon_i \varepsilon_{i'}} = 0 \quad \forall i \forall i'$
- (d) Las X_1, X_2, \dots, X_p son de valores predeterminados, o sea $\rho_{X_i X_k} = 0$ para todo par de variables independientes X_j y X_k
- (e) El modelo está “bien especificado”.

De estos supuestos, se supone que el primero se cumple por default, dada la construcción del modelo como la expresión formal de un valor esperado condicional. Los supuestos (b) a (d) necesitan ser verificados y también el supuesto implícito de normalidad de residuos $\varepsilon_i \sim N(0, \sigma^2)$ que es la base de los contrastes F y t-Student.

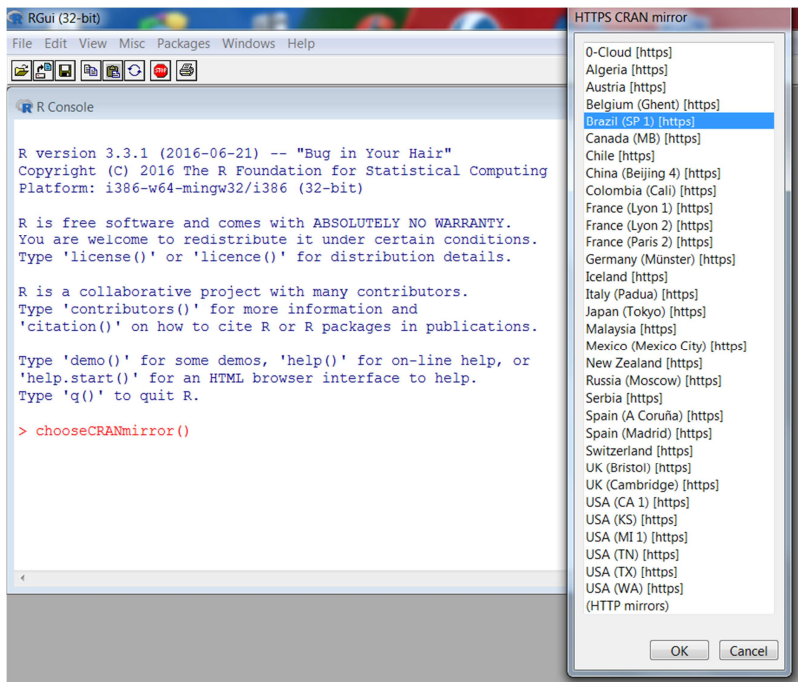
Para hacer esta verificación con R hay varios procedimientos y paquetes, de deben ser instalados pues no todos figuran dentro del paquete básico(base) que viene con la instalación de R. En estas notas usaremos los paquetes “faraway” y “lmtest”. De aquí en adelante suponemos que todas los procedimientos aplicados antes permanecen en memoria, o sea que aún seguimos en la sesión de trabajo con R con el archivo datos2.

Para verificar el supuesto (b) $V(\varepsilon_i) = \sigma^2 = \text{constante} \quad \forall i$ o supuesto de “Homocedasticidad” y el supuesto (c) $\rho_{\varepsilon_i \varepsilon_{i'}} = 0 \quad \forall i \forall i'$ (supuesto de “no Autocorrelación”) usamos el paquete “lmtest”.

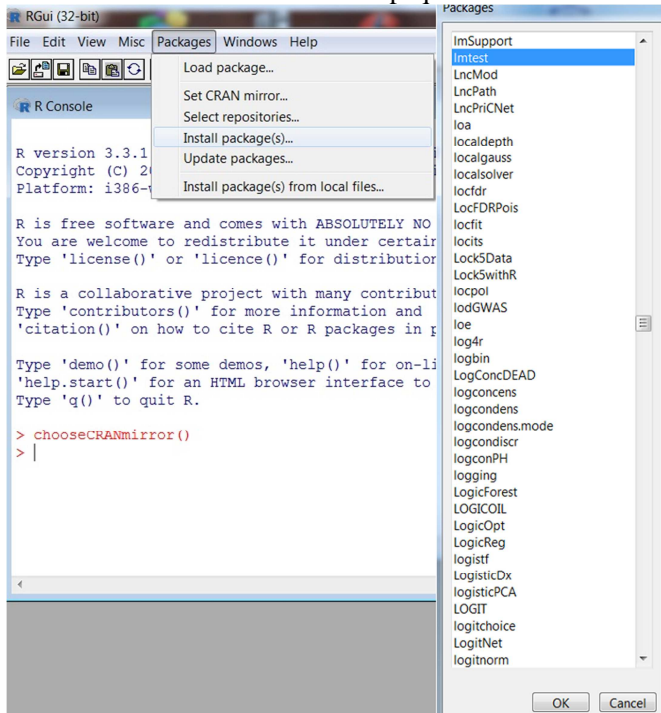
Para verificar el supuesto (d) Las X_1, X_2, \dots, X_p son de valores predeterminados, o sea $\rho_{X_i X_k} = 0$ para todo par de variables independientes X_j y X_k (supuesto de “no Multicolinealidad”) y el supuesto de normalidad de residuos $\varepsilon_i \sim N(0, \sigma^2)$, usaremos el paquete “faraway”.

Instalación del paquete lmtest para verificar los supuestos (b) y (c).

Estando en R, usamos la secuencia Packages→Set CRAN Mirror→escoger un lugar (en este caso seleccionamos Brazil (SP 1), que corresponde a la Universidad de Sao Paulo en Brasil)→OK. R busca comunicarse con el servidor seleccionado y hecho esto queda a la espera de nueva orden



A continuación para instalar el paquete **lmtest**, usamos la secuencia Packages→Install package(s) → Buscamos **lmtest** en la relación de paquetes →OK. R instala el paquete con sus librerías asociadas.



Para empezar a usar el paquete **lmtest** usamos el comando “library” que lo carga en la memoria, como se muestra abajo.

library(lmtest)

Verificación del supuesto (b) $V(\varepsilon_i) = \sigma^2 = \text{constante} \quad \forall i$. Se usa el comando “bptest” de la librería **lmtest** que aplica el Test o contraste de Breusch-Pagan que somete a prueba la hipótesis nula

H_0 : Existe homocedasticidad (o sea $V(\varepsilon_i) = \sigma^2 = \text{constante} \quad \forall i$). Si H_0 es cierta, el estadístico de Breusch-Pagan (BP) tiene una distribución $\chi^2(p)$ y su valor esperado es p (p es el número de variables independientes) de modo que si BP cae muy a la derecha de su valor esperado, o sea toma un valor “muy grande” en relación a lo esperado si fuese cierta H_0 , entonces se debe rechazar H_0 y esto implicaría que el supuesto de homocedasticidad no se está cumpliendo, en caso contrario se acepta H_0 y se concluye que si habría homocedasticidad u homogeneidad de varianzas. Se considera que BP es “grande” si su valor cae arriba del percentil 95 de la distribución $\chi^2(p)$. Como R no sabe que usamos un nivel de significación de 5%, nos proporciona la probabilidad de obtener un estadístico BP mayor o igual que el encontrado en la muestra, que llama “p-value”. **Si p-value < 0.05 el BP muestral está dentro de lo esperado y se acepta H_0 : Existe**

homocedasticidad, o sea que sí se cumple el supuesto $V(\varepsilon_i) = \sigma^2 = \text{constante} \quad \forall i$. En nuestro ejemplo 2 tenemos

bptest(modelo2)

studentized Breusch-Pagan test

data: modelo2

BP = 2.3525, df = 3, p-value = 0.5025

El BP=2.3525 cae cerca de su valor esperado que es 3 y el valor p (p-value=0.5025) es mayor que 0.05, por tanto se acepta H_0 : Existe homocedasticidad y se verifica que el supuesto de homocedasticidad sí se cumple.

Verificación del supuesto (c) $\rho_{\varepsilon_i \varepsilon_{i'}} = 0 \quad \forall i \neq i'$ (supuesto de “no Autocorrelación”). Se usa el comando

“dwtest” de la librería **lmtest** que aplica el Test de Durbin-Watson que usa el estadístico $dw = \frac{\sum_{i=2}^n (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^n (\hat{\varepsilon}_i)^2}$.

Si $H_0: \rho_{\varepsilon_i \varepsilon_{i'}} = 0$ es cierta, el estadístico dw debe caer alrededor de 2 (pues $dw \approx 2(1 - \rho)$). R proporciona la

probabilidad (que llama p-value) de obtener un dw como el de la muestra, si esta probabilidad es pequeña (menor que 0.05) se rechaza $H_0: \rho_{\varepsilon_i \varepsilon_{i'}} = 0$, en caso contrario (si $p\text{-value} \geq 0.05$) se acepta H_0 y se concluye que no hay problema de autocorrelación, esto es, que el supuesto (c) sí se cumple. En nuestro ejemplo tenemos:

dwtest(modelo2)

Durbin-Watson test

data: modelo2

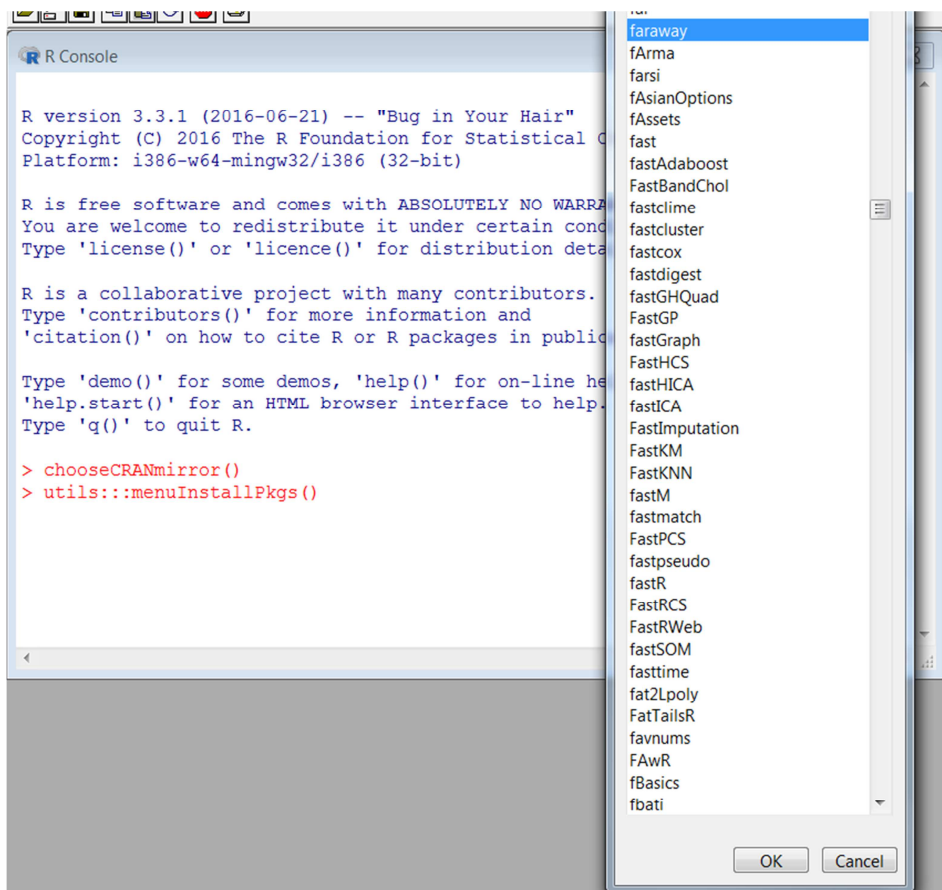
DW = 2.0461, p-value = 0.4875

alternative hypothesis: true autocorrelation is greater than 0

El valor p (p-value = 0.4875) es mayor que 0.05, así que concluimos que el supuesto de no autocorrelación sí se cumple.

Instalación del paquete faraway para verificar el supuesto (d) y el de normalidad de residuos.

Se procede como en el caso de la instalación del paquete lmtest: estando en R, usamos la secuencia Packages→Set CRAN Mirror→escoger un lugar (en este caso seleccionamos Brazil (SP 1), que corresponde a la Universidad de Sao Paulo en Brasil)→OK. Hecho lo anterior, para instalar el paquete faraway, usamos la secuencia Packages→Install package(s) → Buscamos faraway en la relación de paquetes →OK. R instala el paquete con sus librerías asociadas.



Para empezar a usar el paquete **faraway** lo cargamos en memoria con el comando “library”:

library(faraway)

Instalado el paquete faraway procedemos a usarlo para verificar los dos supuestos que nos falta analizar.

Verificación del supuesto (d) Las X_1, X_2, \dots, X_p son de valores predeterminados, o $\rho_{X_i X_k} = 0$ para todo par de variables independientes X_j y X_k (supuesto de “no multicolinealidad”).

En este caso se aplica el comando **vif** del paquete **faraway**. En general si la v. independiente X_i no correlaciona con las demás v. independientes del modelo, si hacemos una regresión múltiple tomando a X_i como dependiente de las demás v. independientes del modelo, debiéramos obtener un R_i^2 cero o cercano a cero. Por diferencia la cantidad $h_i = 1 - R_i^2$ debiera estar cercana a 1. La estadística h_i se llama “Tolerancia” de X_i y si es muy cercana a cero indicaría que X_i depende excesivamente de las demás v. independientes, o sea que está muy correlacionada con alguna(s) de ella(s). Lo anterior se llama “multicolinealidad”. Una regla empírica es considerar que no hay multicolinealidad que si $h_i > 0.1$; como con valores decimales puede ser algo pesado saber cuán pequeños son, una alternativa es trabajar con el inverso de la tolerancia. Esta estadística se llama *Factor de inflación de varianza* (de siglas *vif* en inglés), o sea $vif_i = 1/h_i$ y en este contexto la regla empírica es considerar que no hay multicolinealidad si $vif_i < 10$. En nuestro ejemplo, una vez cargado el paquete **faraway**, usamos el comando **vif** como se muestra abajo:

vif(modelo2)

Obtenemos de R, el siguiente resultado:

```
HorasW      Edad ExpMeses
1.211122 1.062340 1.189082
```

El *vif* de Horas de trabajo (HorasW) es 1.211122; el de Edad es 1.06234 y el de Experiencia es 1.189082. Se ve que todos los *vif* son menores que 10 y se concluye que no hay problema de multicolinealidad.

Verificación del supuesto de normalidad de residuos $\varepsilon_i \sim N(0, \sigma^2)$.

La normalidad de residuos se verifica con el comando **shapiro.test** que contrasta la hipótesis nula H_0 : Los residuos sí tienen distribución normal. Para aplicarlo primero tenemos que crear una variable que guarde los residuos estimados $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ para cada caso. Esto se hace con la orden siguiente:

residuos=residuals(modelo2)

Esta orden graba los residuos que están en memoria en una nueva variable que hemos llamado “residuos” (podríamos usar cualquier otro nombre, no es obligatorio el que hemos usado en este ejemplo). Una vez guardados los residuos estimados del modelo, usamos el comando **shapiro.test** aplicado a la variable **residuos**:

shapiro.test(residuos)

Se obtiene de R:

```
Shapiro-Wilk normality test
```

```
data:  residuos
W = 0.98156, p-value = 0.6833
```

R aplica el test de Shapiro-Wilk que como ya dijimos contrasta la hipótesis H_0 : Los residuos sí tienen distribución normal. La estadística W toma valores entre 0 y 1 y si H_0 es cierta debe tomar valores cercanos a 1, por ello si W cae cerca de 0 se debe rechazar H_0 , en caso contrario se acepta H_0 y se concluye que sí hay distribución normal de los residuos. R proporciona la probabilidad de obtener un W menor o igual que el muestral, esta probabilidad que R llama “p-value” se compara con 0.05, si es menor que 0.05 se rechaza H_0 . En nuestro ejemplo, el valor p (p-value) es $0.6833 > 0.05$ así que no rechazamos H_0 y por concluimos que sí hay normalidad de residuos. El supuesto de normalidad sí se cumple.

En resumen, en este ejemplo todos los supuestos se cumplen y podemos confiar en los resultados del análisis de regresión múltiple aplicado.

Observaciones:

- (1) Los procedimientos aplicados no son los únicos existentes, hay otros procedimientos gráficos y tests útiles, alternativos o complementarios a los mostrados en estas notas pero su explicación excede el objetivo de estos apuntes que son sólo una introducción.
- (2) Cuando algunos de los supuestos no se cumplen hay que buscar medidas correctivas, que son variadas y dependen del caso específico. Estas medidas se ven en cursos más avanzados como los de Econometría y Estadística aplicada.