Computación Científica Actuarial (746)
Cátedra: Departamento de Matemática
Curso: Del Rosso
1er Cuatrimestre de 2019

***"Remember that all models are wrong;***
***the practical question is how wrong***
***do they have to be to not be useful"***
**(George Box, 1987).**

# Modelos Lineales (*Linear Models*) en R

Temas: Modelos Lineales, Bondad del Modelo, Diagnóstico, Contrastes de Hipótesis

El ajuste de modelos emplea funciones cuyo primer argumento es un objeto fórmula que representa el modelo deseado según los nombres de vectores. Las fórmulas en R siguen la sintaxis,

***Variable Respuesta ~ Expresión de Covariables (Variables Predictoras)***

La virgulilla se interpreta como "*está modelada como una función de*".

Por ejemplo, para expresar un modelo de regresión lineal con "$y$" como respuesta (variable dependiente) y "$x$" como covariable (variable independiente) se usa la función $\boldsymbol{lm}()$

$> modelo.lineal < -lm(y \sim x)$

Otros ejemplos de fórmulas válidas,

$peso \sim talle + sexo + región$ $\quad \# \ modelo\ de\ regresión\ múltiple$

$peso \sim talle + sexo + talle{:}\,sexo$ $\quad \# \ modelo\ con\ interacción \rightarrow equivale\ a\ peso \sim talla * sexo$

$peso \sim sexo - 1$ $\quad \# \ modelo$ sin $término\ independiente$

$peso \sim poly(talle, 3)$ $\quad \# \ modelo\ que\ ajusta\ un\ polinomio\ de\ grado\ 3\ a\ talle$

La siguiente tabla exhibe los símbolos utilizados en una fórmula de R,

| Símbolo | Ejemplo | Significado |
|---|---|---|
| **+** | $+X$ | Incluye esta variable |
| **-** | $-X$ | Elimina esta variable |
| **:** | $X{:}\,Z$ | Incluye la interacción entre estas variables |
| ***** | $X * Z$ | Incluye estas variables y la interacción entre ellas |
| **\|** | $X\|Z$ | Incluye un condicionante |
| **^** | $(X + Y + Z)\text{^}3$ | Incluye estas variables y todas las interacciones posibles |
| **I** | $I(X * Z)$ | Incluye una nueva variable que surge del producto de estas variables |
| **1** | $X - 1$ | Elimina el intercepto (regresión a través del origen) |

Otros argumentos de interés de la función $\boldsymbol{lm}()$ son,
$\boldsymbol{data}$ = para indicar una data frame donde buscar los vectores de la fórmula

$\boldsymbol{subset}$ = para indicar una expresión lógica que defina una condición que selecciona las observaciones que deben emplearse para el modelo. Por ejemplo, para seleccionar sólo los hombres puede usarse: subset=(sexo=="hombre"). Los paréntesis son opcionales.

$\boldsymbol{weigths}$ = para indicar un vector con pesos si se desea un modelo de regresión ponderada.

$\boldsymbol{na.action}$ = función que indica que debe hacerse con los valores NA (perdidos). Por defecto es na.omit, por lo que se realiza un análisis con casos completos. Si se desea que el análisis acabe con un mensaje de error si hay NAs debe ponerse na.action=na.fail

**Modelos Lineales Generalizados (Generalized Linear Models – GLM -)**

Son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (Binomial, Poisson, Gamma, etc) y varianzas no constantes. Existen ciertos tipos de variables dependientes que sufren invariablemente la violación de estos supuestos de los modelos normales y los GLM ofrecen una buena alternativa para tratarlos, y se utilizará cuando la variable respuesta (dependiente o endógena) es,

i. Una variable de conteo. En concreto son casos (número de colisiones, accidentes, viviendas destruidas, etc.)
ii. Una variable de conteo de casos expresados éstos como proporciones (porcentaje de heridos graves en accidentes, etc.)
iii. Una variable binaria (vivo o muerto, hombre o mujer, carnet o no, etc.)

La función para estos modelos es $\boldsymbol{glm}()$. Además de los anteriores argumentos, precisa uno más que define la familia del modelo.

$\boldsymbol{family = normal \mid binomial \mid poisson \mid gamma}$

Cada familia está basada en una distribución de probabilidad del error tiene una transformación asociada. Otras transformaciones válidas pueden indicarse entre paréntesis: normal identity binomial logistic poisson log gamma inverse

**Exploración de los resultados**

Normalmente el resultado de ajustar un modelo con $\boldsymbol{lm}()$ o $\boldsymbol{glm}()$ se asigna a un objeto que es de clase lm o glm respectivamente. Las siguientes funciones permiten explorar diferentes aspectos del modelo,

| | | |
|---|---|---|
| $\boldsymbol{print}()$ | $\boldsymbol{summary}()$ | $\boldsymbol{coef}()$ |
| $\boldsymbol{resid}()$ | $\boldsymbol{fitted}()$ | $\boldsymbol{deviance}()$ |
| $\boldsymbol{anova}()$ | $\boldsymbol{predict}()$ | $\boldsymbol{plot}()$ |

**Ejemplo 1**

Con el fin de ejemplificar algunas opciones que se utilizarán ampliamente al estimar modelos de regresión vamos a considerar el caso siguiente. Generamos dos vectores con la siguiente información,

$> y < -c(1{,}2{,}3, -1{,}0, -1{,}2{,}1{,}2)$

$> x < -c(0{,}1{,}2, -2{,}1, -2{,}0. -1{,}1)$

Ahora es posible correr la regresión para el modelo lineal simple, $Y_i = a_0 + a_1 X_i + \varepsilon_i$

Por el momento no nos preocupamos por las características del modelo, ni de la comprensión del método de estimación ya que se abordará más adelante. Aquí simplemente debe aprender que para correr esa regresión se utiliza la función lineal model o $\boldsymbol{lm}()$

```
> lm(y ~ x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     1.0000       0.8125
```

Los resultados de la regresión se pueden obtener con el comando $\boldsymbol{summary}()$,

```
> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8125 -0.3750  0.1875  0.3750  1.0000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0000     0.2938   3.404  0.01138 *
x             0.8125     0.2203   3.688  0.00778 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8814 on 7 degrees of freedom
Multiple R-squared:  0.6602,    Adjusted R-squared:  0.6116
F-statistic:  13.6 on 1 and 7 DF,  p-value: 0.007782
```

**Ejemplo 2 – PWT (Penn World Table)**

Ahora ya estamos en condiciones de preparar nuestros datos para utilizarlos en el paquete. La manera más fácil de manejar los archivos de datos en R es crearlos en una planilla de cálculo como Excel y guardarlos como archivo de texto delimitado por tabulaciones. Para que los datos puedan ser cargados en R habrá que utilizar el comando para leer tablas (read.table) e indicar que la primer línea de su cuadro de datos contiene los nombres de las variables (header=TRUE) y que las columnas están separadas por tabulaciones (sep=). Las instrucciones son las siguientes,

$> datos < -read.table(\text{PWT\_2000.txt}, header = TRUE, sep = \text{""})$

Los datos de la tabla ahora están cargados en un objeto llamado "datos", sin embargo R no puede reconocer cada una de las variables que están en el cuadro. Para indicar que las variables están en las columnas se debe usar la siguiente instrucción,

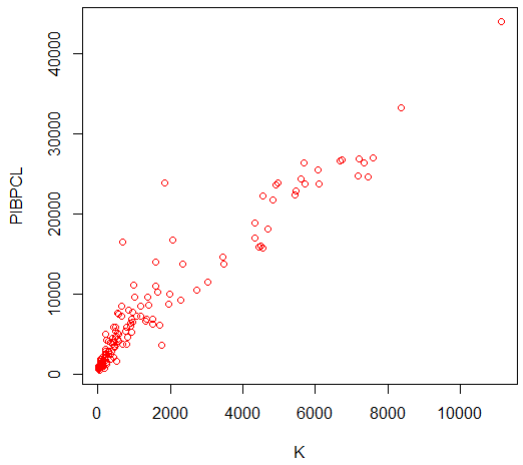
$> attach(datos)$

La leyenda que arroja el software estadístico R es el siguiente,

**The following object(s) are masked from 'datos (position 3)':  K, PAIS, PIBPCL**

Una herramienta gráfica que utilizaremos frecuentemente es un diagrama de dispersión. Por ejemplo, se puede solicitar una nube de puntos para visualizar la relación entre el esfuerzo de inversión de los países $(K)$ y su ingreso per cápita $(PIBPCL)$,

$> plot(K, PIBPCL, col = "red")$




En la gráfica se puede observar claramente una relación positiva entre el esfuerzo de inversión y el PIB per cápita de los países de la muestra de datos. Sin embargo, es posible estimar el coeficiente de correlación lineal entre ambas variables, mediante la siguiente instrucción,

$> cor(K, PIBPCL)$

$[1]\ 0.9568068$

Asimismo es posible efectuar un contraste de hipótesis de significatividad del coeficiente de correlación lineal. El planteo de hipótesis es el siguiente,

$H_0: \rho(K, PIBPCL) = 0$

$H_1: \rho(K, PIBPCL) \neq 0$

La sintaxis en R para efectuar dicho test es,

$> cor.test(K, PIBPCL)$

```
        Pearson's product-moment correlation

data:  K and PIBPCL
t = 37.668, df = 131, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9396197 0.9691793
sample estimates:
      cor
0.9568068
```

Teniendo en cuenta que se rechazó la hipótesis nula y que bajo la evidencia empírica se puede afirmar que el coeficiente de correlación lineal entre ambas variables es distinto a cero, es posible plantear un contraste para evaluar si el grado de asociación lineal es positiva o negativa. En este caso se testea la posibilidad que la hipótesis alternativa sea menos a cero, es decir,

$H_0: \rho(K, PIBPCL) \geq 0$

$H_1: \rho(K, PIBPCL) < 0$

La sintaxis en R para efectuar dicho test es,

$> cor.test(K, PIBPCL, alternative = "less")$

Dado que ya sabemos utilizar el comando "lm" de regresión lineal, podemos ahora estimar un modelo para explicar el ingreso per cápita de los países en función de su capital, pero ahora guardaremos el resultado en un objeto bajo el nombre PWT (Penn World Table),

$> PWT < -lm(PIBPCL \sim K)$

Un comando útil que proveerá los coeficientes estimados del modelo es,

$> coef(PWT)$

```
(Intercept)           K
2363.858449    3.641078
```

Para obtener un resumen estadístico más completo del modelo, por ejemplo, utilizamos el comando $\boldsymbol{summary}()$

$> summary(PWT)$

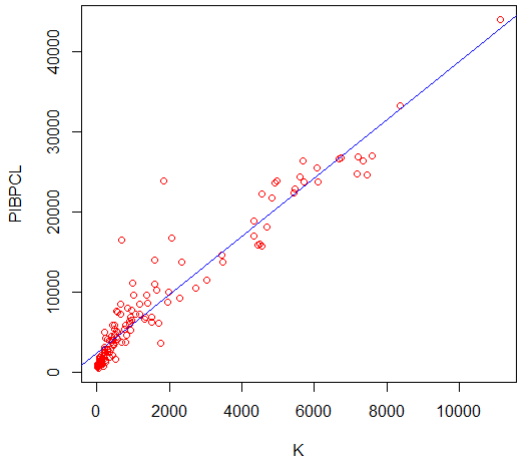
Los resultados del modelo indican que al incrementarse la inversión en un dólar el ingreso de los países se incrementa en 3.64 dólares, tal y como se aprecia en el siguiente cuadro.

```
Call:
lm(formula = PIBPCL ~ K)

Residuals:
    Min      1Q  Median      3Q     Max
-5180.3 -1553.1  -591.4   825.3 14757.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.364e+03  2.800e+02   8.443 5.06e-14 ***
K           3.641e+00  9.666e-02  37.668  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2564 on 131 degrees of freedom
Multiple R-squared:  0.9155,    Adjusted R-squared:  0.9148
F-statistic:  1419 on 1 and 131 DF,  p-value: < 2.2e-16
```

**Adición de Líneas de Regresión a los Gráficos**

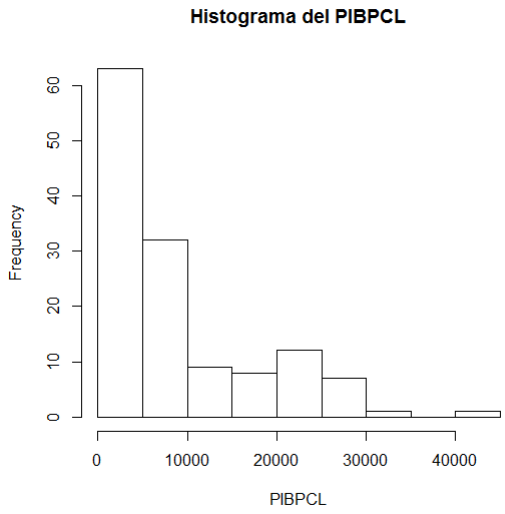
La recta de regresión estimada la podemos añadir al diagrama de dispersión que ya habíamos generado con la siguiente opción,

$> abline(PWT, col = "blue")$




Otra gráfica que nos va a ser de utilidad es el histograma, en el cual podemos relacionar intervalos de los datos con sus frecuencias. Con la siguiente instrucción generaremos el histograma para los datos del PIB per cápita de los países,

$> hist(PIBPCL, main = "Histograma\ del\ PIBPCL")$

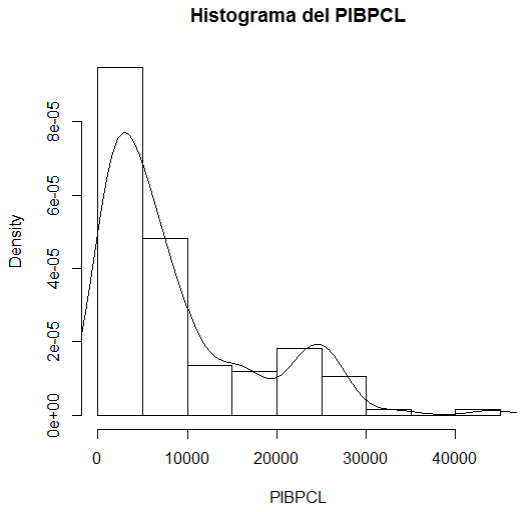



Claramente el histograma muestra que la mayoría de los países se encuentran en los ingresos más bajos de la distribución. Resulta útil visualizar el histograma en densidades (área bajo la curva igual

a la unidad) y añadirle funciones de densidad kernel, lo cual se puede hacer con la instrucción siguiente,

$> hist(PIBPCL, main = "Histograma del PIBPCL", freq = FALSE)$

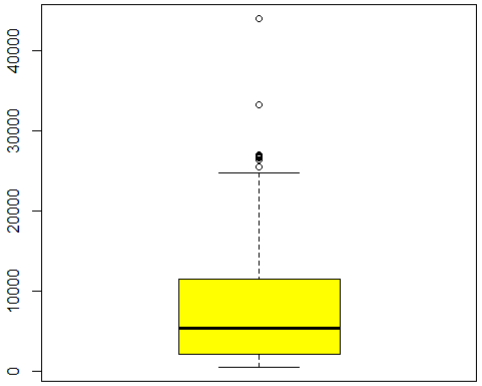
$> lines(density(PIBPCL))$




Para observar la distribución de los datos es utilizar cajas de box, en las cuales la caja muestra los umbrales para los cuartiles inferior y superior, además de la mediana. Las líneas abajo y arriba de la caja permiten identificar las observaciones extremas. Para obtener este tipo de gráficas se utiliza la instrucción siguiente,

$> boxplot(PIBPCL, col = "yellow")$

La gráfica resultante exhibe un grupo de países con ingresos extremos.



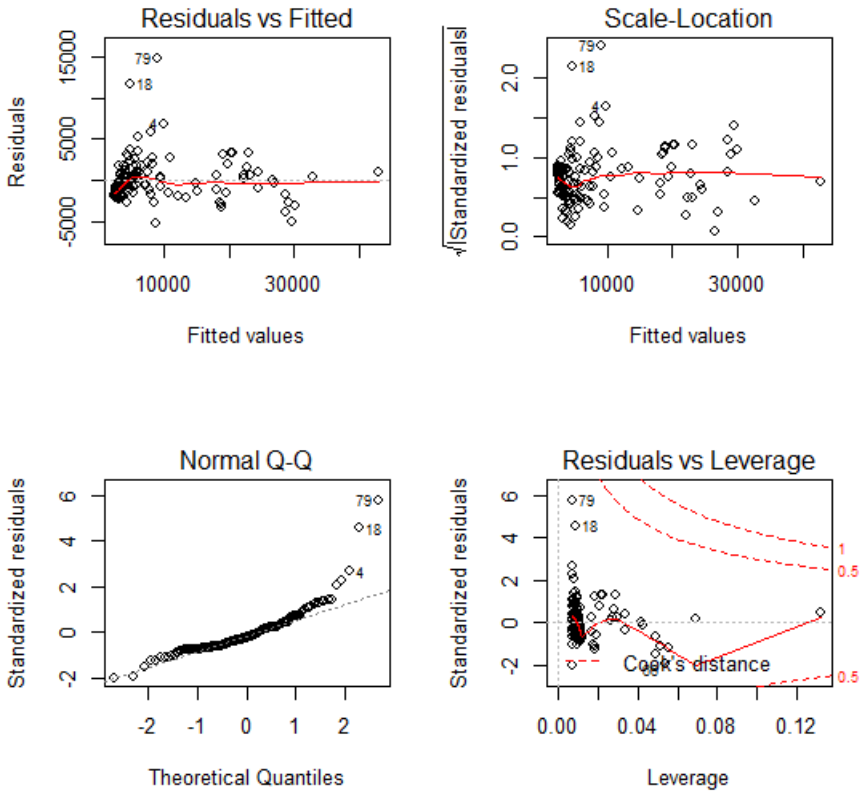
Computación Científica Actuarial (746)
Cátedra: Departamento de Matemática
Curso: Del Rosso
1er Cuatrimestre de 2019

**Evaluación de los Resultados de un Modelo Lineal**

Antes de aceptar el resultado de un modelo lineal, es importante evaluar su idoneidad para explicar los datos. Una de las varias formas de hacer esto es examinar visualmente los residuos.
Si el modelo es apropiado, entonces los errores residuales deben ser aleatorios y normalmente distribuidos. Además, la eliminación de un caso no debe afectar significativamente la idoneidad del modelo. El programa estadístico suministra 4(cuatro) aproximaciones gráficas para evaluar un modelo, mediante el comando $\boldsymbol{plot}()$

$> layout(matrix(1{:}4,2,2))$

$> plot(PWT)$




El gráfico en extremo superior izquierdo exhibe los residuos del modelo versus los valores estimados (fitted) por el modelo. Los residuos deberían ser distribuidos aleatoriamente alrededor de la línea horizontal, que representa un error residual de cero, lo que implica, que no debería existir una tendencia distinta en la distribución de los puntos. El gráfico en el extremo inferior izquierdo es un gráfico QQ Plot, que debería sugerir que los errores residuales son distribuidos normalmente. El gráfico "scale – location" en el extremo superior derecho exhibe la raiz cuadrada de los residuos estandarizados como una función de los valores ajustados. Y el ultimo gráfico en el extremo inferior derecho exhibe cada punto de apalancamiento, que representa una medida de la importancia en la determinación de la regresión resultante.

**Utilización de los Resultados de una Regresión para hacer Predicciones**

El objetivo de un análisis de regresión, por supuesto, es para desarrollar un modelo que pueda ser utilizado para predecir los resultados de experimentos futuros. En nuestro ejemplo, la ecuación de calibración es,

$$\widehat{PIBPCL} = 2364 + 3{,}641 * K$$

La recta ajustada se utiliza tanto para estimar la respuesta media como para predecir una respuesta individual. La distinción entre la estimación de una media y la predicción de un valor futuro se aclara cuando utilizamos intervalos que proveen una medida del error. Una vez hallada la recta de regresión, si la regresión es significativa, podría interesarnos predecir la variable respuesta para un valor fijo x de la variable independiente, o estimar la respuesta media. Tiene sentido realizar predicciones y estimaciones dentro del rango de la variable independiente. El rango de la variable independiente es,

$> range(K)$

$[1]\ 25.02441 \quad 11130.66452$

Podemos hacer predicciones y estimaciones dentro de este rango de valores.

**Intervalos de predicción para un valor dado de la variable independiente**

Por ejemplo, podría interesarnos predecir el valor del ingreso per cápita entre los países para valores promedio del esfuerzo de inversión igual a 370, 390, 400. Podemos hacer la cuenta,

$> 2364 + 3.641 * 370$
$[1]\ 3711.17$

$> 2364 + 3.641 * 390$
$[1]\ 3783.99$

$> 2364 + 3.641 * 400$
$[1]\ 3820.4$

En R podemos usar la sentencia "$predict$", que nos permite hallar los valores predichos y sus intervalos de predicción respectivos. El modelo ajustado era:

$> PWT < -lm(PIBPCL \sim K)$

Si hacemos,

$> predict(PWT)$

Obtenemos los mismos valores haciendo,

$> PWT\$fitted$

Lo chequeamos restando:

$> round(predict(PWT) - PWT\$fitted, 0)$

Si queremos predecir el valor del ingreso per capital para valores del Esfuerzo de Inversión =370, 390, 400 y sus respectivos intervalos de predicción:

$> new < -data.frame(x = c(370,390,400))$

$> predict(PWT, new, interval = "prediction")$

El default es 0.95, si queremos otro nivel lo tenemos que especificar:

$> predict(PWT, new, interval = "prediction", level = 0.99)$

**Observación:** Necesitamos definir el conjunto de datos que queremos predecir como un data frame y tiene que tener la variable el mismo nombre que nuestra variable independiente, en nuestro cado x. Si no ponemos un nuevo conjunto de datos, lo hace para los valores ajustados por el modelo.

$> predict(PWT, interval = "prediction")$

**Intervalos de Confianza para la respuesta media**

Si fijamos un valor de la variable independiente x, y queremos hallar el valor esperado de Y y su respectivo intervalo de confianza usamos,

$> predict(PWT, new, interval = "confidence")$

El default es 0.95, si queremos otro nivel lo tenemos que especificar:

$> predict(PWT, new, interval = "confidence", level=0.99)$

Si comparamos los resultados obtenidos vemos que los valores predichos, y los esperados coinciden, pero los intervalos de predicción tienen mayor longitud que los intervalos de confianza.

Debido a que existe incertidumbre tanto en la pendiente calculada como en la intercepción, habrá incertidumbre en el el esfuerzo de inversión (K) calculados. Suponga que deseamos predecir el ingreso per cápita para distintos niveles de inversión junto con el intervalo de confianza para cada uno. Podemos utilizar el comando $\boldsymbol{predict}()$ para hacer esto. La sintaxis es,

$$\boldsymbol{predict}(modelo, data.frame(pred = new\ pred), level = 0.95, interval = "confidence")$$

Donde "*pred*" es el objeto que contiene los valores de la variable independiente y "*new pred*" es el objeto que contiene los nuevos valores para las predicciones que son deseadas y "*level*" es el nivel de confianza. Asimismo "*lwr*" es el límite inferior del intervalo de confianza y "*upr*" es el límite superior de dicho intervalo. Una de las limitaciones de R, es que no contiene una característica para encontrar los intervalos de confianza para los valores predichos de la variable independiente para valores específicos de la variable dependiente.

**Ejemplo 3 – Deuda**

Para comenzar, se utilizarán datos de la Economía Mexicana para el periodo comprendido de Enero de 2009 a Diciembre de 2013, con frecuencia mensual y cuya fuente provienen de la página web del Banco de México (www.banxico.gob.mx), con dicha información permitirá estimar el siguiente modelo,

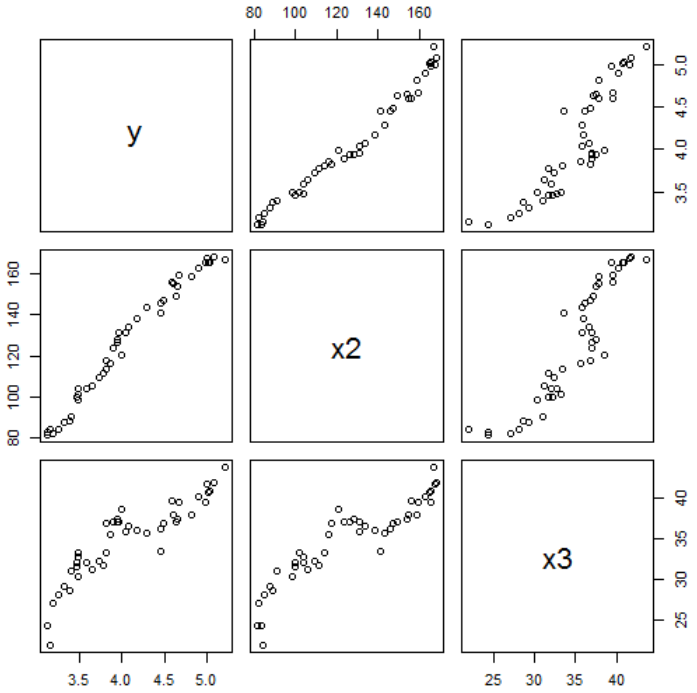
$$Deuda\ Pública_t = a_0 + a_1 RRII_t + a_2 Indice\ Bursatil_t + \varepsilon_t$$

La variable dependiente, y, es el nivel de deuda pública del gobierno mexicano (expresada en miles de millones de pesos) que es explicada por el nivel de Reservas Internacionales, X2, (expresada en miles de millones de dólares) y por el índice bursátil de la Bolsa Mexicana de Valores, X3 (miles de unidades). Para encontrar el modelo que explique el comportamiento de la Deuda Externa en función de las Reservas Internacionales y del Índice Bursátil se utilizarán los datos que se encuentran en el archivo "*deuda_mex*" con extensión CSV (delimitado por comas). Para ejecutarlo en R se hace uso del siguiente código,

$deuda < -read.csv("C:/data/deuda\_mex.csv", header = T)$

$attach(deuda)$

$pairs(deuda)$




El comando para obtener el vector de residuos del modelo estimado es,

$> residuos < -modelo\$residuals$

Otra posibilidad alternativa es mediante la siguiente sintaxis,

$> residuos < -resid(modelo)$

En nuestro ejemplo, la matriz de varianzas-covarianzas de las estimaciones de los parámetros del modelo se obtiene mediante el siguiente comando,

$> vcov(modelo)$

| | (Intercept) | x2 | x3 |
|---|---|---|---|
| (Intercept) | 1.739391e-02 | 9.807685e-05 | -8.501880e-04 |
| x2 | 9.807685e-05 | 1.702452e-06 | -9.002337e-06 |
| x3 | -8.501880e-04 | -9.002337e-06 | 5.715120e-05 |

El comando para obtener la tabla ANOVA del ejemplo que se desarrollo es el siguiente,

$> anova(modelo)$

**Contrastes de Hipótesis**

**Test de Student**

La instrucción en R para aplicar el test de Student es "$\boldsymbol{t.test}$". Por default es un test bilateral (a dos colas) y con un intervalo de confianza del 95%.

**Ejemplo 1: Para la media de una población normal**

Generamos una muestra de tamaño 30 de una distribución normal con media 5 y desvío 1.

$> set.seed(34)$

$> x < -rnorm(30,5,1)$

$> t.test(x, mu = 5)$

**Ejemplo 2: Podemos hacerlo con más argumentos**

$> t.test(x, alternative = "two.sided", mu = 5, conf.level = 0.95)$

Si planteamos un test unilateral a izquierda

$> t.test(x, alternative = "less", mu = 5, conf.level = 0.95)$

**¿Qué es lo único que cambia? ¿Por qué?**

**Ejemplo 3: Si planteamos un test unilateral a derecha para mu=4 y nivel de confianza 0.99**

$> t.test(x, alternative = "greater", mu = 4, conf.level = 0.95)$

**Para comparar la media de dos poblaciones normales e independientes**

Por default es un test bilateral, y con un intervalo de confianza del 95%, diferencia 0 y varianzas distintas.

**i. Asumiendo varianzas iguales**

Generamos otra muestra de tamaño 25 de una distribución normal con media 6 y desvío 1.

$> set.seed(35)$

$> y < -rnorm(25,6,1)$

$> t.test(x, y, var.equal = TRUE)$

**Ejemplo 5: Si queremos ver si la media de x es menor en una unidad que la media de y**

$> t.test(x, y, mu = 1, alternative = "less", var.equal = TRUE)$

**Ejemplo 6: Cuando la variable de interés está definida mediante un factor**

**Generamos un data frame denominado "datos"**

$> set.seed(20)$

$> variable < -rnorm(40,20,2)$

$> set.seed(20)$

$> sexo < -sample(c("F","M"),40,replace=TRUE)$

$> datos < -data.frame(variable, sexo)$

$> t.test(variable \sim sexo, data = datos, var.equal = TRUE)$

**ii. Asumiendo varianzas distintas (Test de Welch)**

**Ejemplo7**

$> set.seed(29)$

$> x1 < -rnorm(25,5,0.5)$

$> set.seed(30)$

$> x2 < -rnorm(20,5.2,1)$

$> t.test(x1, x2)$

**Para muestras apareadas**

**Ejemplo 8**

$> set.seed(29)$

$> w1 < -rnorm(25,5,0.5)$

$> set.seed(30)$

$> w2 < -rnorm(25,3,1)$

$> t.test(w1, w2, paired = TRUE)$

**Test F para igualdad de varianzas**

La instrucción en R para aplicar el test F es "$var.test$". Por default es un test bilateral para cociente 1 y con un intervalo de confianza del 95%. Chequeamos el supuesto para los test realizados anteriormente.

**Ejemplo 9**

$> var.test(x, y)$

**Ejemplo 10**

$> var.test(x1, x2)$

**Ejemplo 11**

$> var.test(variable \sim sexo, data = datos)$

*Ver el help para otras opciones*

**Test de Shapiro Wilk (SW)**

Se utiliza para contrastar la normalidad de un conjunto de datos. La hipótesis nula postula que una muestra de $x_1, \dots, x_n$ proviene de una población que se encuentra normalmente distribuida. Es considerado uno de los test más potentes para el contraste de normalidad, sobre todo para muestras pequeñas (n<50). El estadístico utilizado es,

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Se rechazará si $W$ es demasiado pequeño y puede oscilar entre 0 y 1. Si el P-Valor es menor al nivel de significación entonces la hipótesis nula es rechazada (se concluye que los datos no provienen de una distribución normal). Ahora bien, si el P-Valor es mayor al nivel de significación, no se rechaza la hipótesis y se concluye que los datos siguen una distribución normal. Para los ejemplos vistos,

**Ejemplo 12**

$> shapiro.test(x)$

**Ejemplo 13**

$> shapiro.test(y)$

**Ejemplo 14**

Para muestras apareadas la diferencia debe tener distribución normal

$> shapiro.test(w1 - w2)$

**Ejemplo 15**

$> tapply(datos\$variable, datos\$sexo, shapiro.test)$

**Test de Anderson Darling (AD)**

Es una prueba estadística no paramétrica que permite determinar si una muestra de datos se extrae de una distribución de probabilidad. El Estadístico de Prueba es,

$$A^2 = -N - S$$

$$S = \sum_{k=1}^{N} \frac{2k-1}{N} [lnF(X_k) + \ln(1 - F(X_{N+1-k}))]$$

En su forma más sencilla, el test asume que no existen parámetros a estimar en la distribución que se está probando, en cuyo caso la prueba y su conjunto de valores críticos siguen una distribución libre. Sin embargo, la prueba se utiliza con mayor frecuencia en contextos en los que se está probando una familia de distribuciones, en cuyo caso deben ser estimados los parámetros de esa familia y debe tenerse estos en cuenta a la hora de ajustar la prueba estadística y sus valores críticos. Cuando se aplica para probar si una distribución normal describe adecuadamente un conjunto de datos, es una de las herramientas estadísticas más potentes para la detección de la mayoría de las desviaciones de la normalidad. Si el P-Valor es inferior al nivel de significación se rechaza la hipótesis nula de normalidad.

**Test de Kolmogorov – Smirnov (KS)**

Esta prueba conocida como KS es una prueba no paramétrica que determina la bondad de ajuste de una distribución empírica con una teórica, en este caso con la Distribución Normal.

El estadístico de prueba para un test unilateral o a una sola “cola” es la siguiente expresión,

$$D = max|F_n(x) - F_0(x)|$$

Cuando la prueba es bilateral o a dos colas, el estadístico surge del cálculo de dos diferencias absolutas,

$$D^+ = max|F_n(x) - F_0(x)|$$

$$D^- = max|F_0(x) - F_n(x)|$$

donde $F_n(x)$ representa la Función de Distribución Muestral que se calcula de la forma siguiente,

$$F_n(x) = \sum_{i=1}^{n} x_i \begin{cases} 1 & si \quad x_i \leq x \\ 0 & en\ otra\ alternativa \end{cases}$$

y $F_0(x)$ corresponde a la función teórica o correspondiente a la población normal especificada en la hipótesis nula.

La distribución del estadístico de Kolmogorov-Smirnov es independiente de la distribución poblacional especificada en la hipótesis nula y los valores críticos de este estadístico están tabulados. Si la distribución postulada es la normal y se estiman sus parámetros, los valores críticos se obtienen aplicando la corrección de significación propuesta por Lilliefors.

Si el P-Valor es menor al nivel de significación entonces la hipótesis nula es rechazada (se concluye que los datos no provienen de una distribución normal). Ahora bien, si el P-Valor es mayor al nivel de significación, no se rechaza la hipótesis y se concluye que los datos siguen una distribución normal.

## Contrastes de Homocedasticidad

El Contraste de **Breusch - Pagan (BP)** es uno de los contrastes más conocido para detectar la heterocedasticidad. La idea es comprobar si se puede encontrar un conjunto de variables, que permitan determinar la dinámica de la varianza de las perturbaciones, estimada a partir del cuadrado de los errores del modelo inicial. El proceso a seguir para llevar a cabo este contraste es el siguiente,

i. Estimar el modelo inicial, sobre el que se pretende saber si hay o no heterocedasticidad, empleando MCO y determinar los errores.

ii. Calcular una serie con los errores del modelo anterior al cuadrado estandarizados,

$$\tilde{\epsilon}_i^2 = \frac{\epsilon_i^2}{\hat{\sigma}^2}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n}$$

iii. Estimar una regresión sobre los determinantes de los errores mediante la incorporación de variables independientes "$Z$", mediante las cuales se busca establecer si este conjunto de variables explican el proceso de heterocedasticidad de las perturbaciones en el modelo original; la estimación propuesta es la siguiente,

$$\tilde{\epsilon}_i^2 = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \cdots + \alpha_p Z_{pi} + \varepsilon_t$$

iv. El modelo es ineficiente si la varianza de la variable dependiente estimada y su error estimado es grande. Entonces, podría afirmarse que el poder explicativo del conjunto de variables Z sobre la representación de la varianza de las perturbaciones aleatorias es escaso. Mediante el diseño de un contraste calculado con la sumatoria de los residuales de la estimación planteada en el paso 3, cuando este se encuentre cercano a cero, la probabilidad de que el proceso sea homocedástico es alta. El contraste propuesto sería el siguiente,

$$\frac{\sum_{i=1}^{n} \widehat{\tilde{\epsilon}_i^2} * n}{2}$$

Breusch y Pagan, mostraron que el contraste se distribuye como una Chi-cuadrada, cuando el proceso del modelo es homocedástico, al revisar el contraste tablas, se toman en cuenta las siguientes hipótesis,

$$H_0: Presencia\ de\ Homocedasticidad$$

$$H_1: Presencia\ de\ Heterocedasticidad$$

Cuando la probabilidad de cometer el Error Tipo I, es muy alta no se puede rechazar la hipótesis nula, entonces, la varianza de los errores aleatorios es constante, por lo tanto, homocedásticos.

El **test de White** es considerado una prueba robusta al no requerir supuestos previos como, por ejemplo, la normalidad de las perturbaciones. De igual manera, no es necesario determinar a priori las variables explicativas que determinan heterocedasticidad.
El objetivo de esta prueba es determinar si las variables explicativas del modelo, pueden determinar la evolución de los errores al cuadrado. Es decir; si la dinámica de las variables explicativas en relación a las varianzas y covarianzas es significativa para determinar el valor de la varianza muestral de los errores.
El proceso de estimación es el siguiente,

i. Estimar el modelo original por MCO, para obtener los errores en la estimación.
ii. Estimar una regresión sobre los determinantes de los errores, con la incorporación de todas las variables incluidas en el estimación del primer modelo, estas elevados al cuadrado y sus combinaciones no repetidas,

$$\epsilon_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \cdots + \alpha_p X_{pi} + \varepsilon_t$$

## Contrastes de Incorrelación o Ausencia de Correlación Serial

Para detectar la autocorrelación se pueden utilizar métodos gráficos y contrastes de hipótesis. Con frecuencia un examen visual de las perturbaciones nos permitirá conocer la presencia de la autocorrelación. Aunque es una forma subjetiva de probar la existencia de la autocorrelación, existen pruebas formales para detectarla.

El Contraste de **Durbin Watson (DW)** es uno de los contrastes más conocido para detectar la existencia de autocorrelación serial. El planteo de hipótesis es el siguiente,

$$H_0: No\ existe\ autocorrelación\ serial$$

$$H_1: Existe\ autocorrelación\ serial$$

El Estadístico de Prueba se puede expresar de la siguiente manera,

$$DW = \frac{\sum_{t=2}^{n}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})}{\sum_{t=1}^{n}\hat{\epsilon}_t^2} = \frac{\sum_{t=2}^{n}\hat{\epsilon}_t^2 + \sum_{t=2}^{n}\hat{\epsilon}_{t-1}^2 + 2\sum_{t=2}^{n}\hat{\epsilon}_t\hat{\epsilon}_{t-1}}{\sum_{t=1}^{n}\hat{\epsilon}_t^2}$$

Para muestras grandes se puede considerar que las sumatorias de los residuos del modelo en el periodo $t$ y en $t-1$ son casi iguales. De este modo, el estadístico de DW puede expresase de la siguiente forma,

$$DW \cong \frac{2\sum_{t=2}^{n}\hat{\epsilon}_t^2 - 2\sum_{t=2}^{n}\hat{\epsilon}_t\hat{\epsilon}_{t-1}}{\sum_{t=1}^{n}\hat{\epsilon}_t^2} = 2(1-\hat{\rho})$$

Donde $\hat{\rho}$ denota el coeficiente de correlación lineal muestral de orden 1. El Estadístico asumirá distintos valores y se puede comprobar la existencia de la autocorrelación serial de primer orden,

Si $\hat{\rho} = -1 \Longrightarrow DW \cong 4 \Longrightarrow Existe\ autocorrelación\ negativa$

Si $\hat{\rho} = 0 \Longrightarrow DW \cong 2 \Longrightarrow No\ existe\ autocorrelación\ serial$

Si $\hat{\rho} = 1 \Longrightarrow DW \cong 0 \Longrightarrow Existe\ autocorrelación\ positiva$

Los criterios de rechazo y de indecisión de la Hipótesis Nula son,

Si $DW < DW_L \Longrightarrow Existe\ evidencia\ de\ autocorrelación\ serial\ positiva$

Si $DW > 4 - DW_L \Longrightarrow Existe\ evidencia\ de\ autocorrelación\ serial\ negativa$

Si $DW_U < DW < 4 - DW_U \Longrightarrow No\ hay\ evidencia\ de\ autocorrelación$

Si $DW_L < DW < DW_U \Longrightarrow La\ prueba\ no\ es\ concluyente$

Si $4 - DW_U < DW < 4 - DW_L \Longrightarrow La\ prueba\ no\ es\ concluyente$

A pesar de ser la prueba más conocida y más utilizada para detectar la autocorrelación, sólo permite detectar la autocorrelación serial de primer orden y carece de interpretación cuando incluimos rezagos dentro del modelo, además no permite obtener conclusiones en las regiones de indecisión.

El Contraste de **Breusch-Godfrey (BG)** determina si existe o no autocorrelación de orden superior a uno y consiste en estimar una regresión auxiliar mediante el Método de Mínimos Cuadrados Ordinarios (MCO) y efectuar una prueba de hipótesis sobre los parámetros de esta regresión.

Supongamos que se estima el siguiente modelo, $Y_t = X_t B + \epsilon_t$

La regresión auxiliar para el contraste de autocorrelación hasta de orden p en los residuos tiene la siguiente forma, $\epsilon_t = X_t \theta + \sum_{j=1}^{p} \epsilon_{t-j} + v_t$

Donde el estadístico $LM = n * R^2\ (n \to \infty) \sim \chi_p^2$ y el P-Valor se calcula de la siguiente manera,

$$P - Valor = P(\chi_p^2 \geq LM)$$

Las ventajas de la prueba Breusch-Godfrey son las siguientes,

i) Implementación relativamente fácil
ii) se puede generalizar para detectar autocorrelación de orden superior
iii) la distribución asintótica del estadístico LM para la prueba de autocorrelación hasta de orden p tiene una Distribución $\chi_p^2$