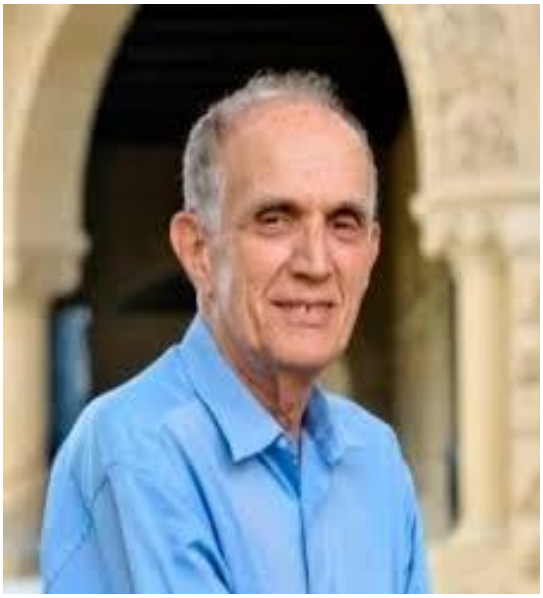
Unidad III

COMPROBACIÓN DE ADECUACIÓN DEL MODELO

"Those who ignore Statistics are condemned to reinvent it"
Bradley Efron

1. Introducción

Los modelos de regresión estudiados deben cumplir una sería de requisitos para que sus resultados tengan validez.
Esos requisitos son los supuestos que están asociados a la normalidad de errores, independencia y homogeneidad.
La verificación de estos supuestos se puede hacer de manera descriptiva mediante gráficos, sin embargo, el análisis gráfico nos podría llevar a conclusiones erróneas. Por lo que se debe pensar en análisis inferencial mediante pruebas de hipótesis.


Bradley Efron (1938-)

El incumplimiento de los supuestos puede tener consecuencias graves, como generar un modelo inestable, es decir que para una muestra distinta se obtiene un modelo totalmente diferente y obtener conclusiones opuestas.
En esta unidad se presentarán varios métodos de utilidad para diagnosticar violaciones de las premisas básicas de regresión. Esos métodos de diagnóstico se basan principalmente en el estudio de los residuales del modelo

Revisando bibliografía encontré este libro que me parece muy interesante

https://fhernanb.github.io/libro_regresion/index.html
https://rpubs.com/sebas_Alf/740734

2. Análisis de Residuales

2.1 Definición de residual

Un residual se define como:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

siendo y_i una observación, y $\hat{y}_i$ su valor ajustado correspondiente. Se puede considerar que un residual es la desviación entre los datos y el ajuste, también es una medida de la variabilidad de la variable de respuesta que no explica el modelo de regresión. Los residuales son los valores realizados, u observados, de los errores del modelo. El análisis de los residuales es una forma eficaz de descubrir diversos tipos de inadecuación del modelo.

Los residuales tienen varias propiedades importantes. Tienen media cero, y su varianza promedio aproximada se estima con:

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n-p} = \frac{SC_{Res}}{n-p} = CM_{Res}$$

Recordemos que:

El vector de valores ajustados $\hat{y}_i$ que corresponden a los valores observados y_i es:

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

```
modelo1<-lm(mpg~.,data=caso)
modelo1
summary(modelo1)

modelo2<-lm(mpg~sp+wt+hp,data=caso)
betas<-as.matrix(modelo2$coefficients)
X<-as.matrix(cbind(1,caso[,-c(1,4)]))
Y<-as.matrix(caso[,1])
#Estimados
estim<-X%*%betas
modelo2$fitted.values
H<-X%*%solve(t(X)%*%X)%*%t(X)
H%*%Y

#Residuos
resi<-caso$mpg-estim
modelo2$residuals
n<-nrow(caso)
I<-diag(1,n,n)
(I-H)%*%Y
mean(resi)
```

2.2 Métodos para escalar residuales

Los residuales escalados son útiles para determinar observaciones que sean atípicas o valores extremos, esto es, observaciones que en algún aspecto estén separadas del resto de los datos.

Residuales Estandarizados

Ya que la varianza aproximada de un residual se estima con CMRes el cuadrado medio de los residuales, un escalamiento lógico de los residuales sería el de los residuales estandarizados

$$d_i = \frac{e_i}{\sqrt{CM_{Res}}} \quad , \qquad i = 1, 2, \ldots, n$$

Los residuales estandarizados tienen media cero y varianza aproximadamente unitaria, en consecuencia, un residual estandarizado grande indica que se trata de un valor atípico potencial.

```
#Residuos Estandarizados
sig<-summary(modelo2)$sigma
restan1<-resi/sig
head(restan1)
```

```
       1          2          3          4          5          6
 3.2341820  1.6667802  1.6393101  0.9751576 -1.0828789  0.2059943
```

Residuales Estudentizados

Se puede mejorar el escalamiento de residuales dividiendo e_i entre la desviación estándar exacta del i-ésimo residual. El vector de los residuales se puede escribir como sigue:

$$e = (I - H)y$$

donde H = X(X'X)$^{-1}$ es la matriz de sombrero. Esta matriz tiene varias propiedades útiles. Es simétrica (H ' = H) Y es idempotente (HH = H). De forma parecida, la matriz 1-H es simétrica e idempotente

$$e = (I - H)(X\beta + \varepsilon) = X\beta - HX\beta + (I - H)\varepsilon$$

$$= X\beta - X(X'X)^{-1}X'X\beta + (I - H)\varepsilon = (I - H)\varepsilon$$

Por lo anterior, los residuales son la misma transformación lineal de las observaciones *y* y los errores ε

$$Var(e) = Var[(I - H)\varepsilon] = (I - H)Var(\varepsilon)(I - H)' = \sigma^2(I - H)$$

La matriz de covarianza de los residuales es:

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

en donde h_{ij} es el i-ésimo elemento de la diagonal de la matriz de sombrero H y es llamado leverage. La covarianza entre los residuales e*i* y e_j es:

$$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Ahora bien, ya que $0 \leq h_{ij} \leq 1$, si se usa el cuadrado medio de los residuales, CMRes para estimar la varianza de los residuales, en realidad se sobreestima la Va(e_i).
Las violaciones de las premisas del modelo están, con más probabilidad, en los puntos remotos, y pueden ser difíciles de detectar por inspección de los residuales ordinarios e_i (o de los residuales estandarizados d_i), porque en general, sus residuales serán menores.
Entonces, un procedimiento lógico es examinar los residuales estudentizados es:

$$r_i = \frac{e_i}{\sqrt{\text{CM}_{\text{Res}}(1 - h_{ii})}} \quad , \qquad i = 1, 2, \ldots, n$$

Cuando la forma del modelo es correcta estos residuales estudentizados tiene varianza constante Var(r_i) = 1.

```
#Residuos Estudentizados
diag(H)
hii<-hatvalues(modelo2)
restuden1<-resi/sqrt(sig^2*(1-hii))
head(restuden1)
        1         2         3         4         5         6
 3.4344397  1.7334874  1.7049179  0.9940873 -1.1323904  0.2099930

restuden2<-rstandard(modelo2)
head(restuden2)
        1         2         3         4         5         6
 3.4344397  1.7334874  1.7049179  0.9940873 -1.1323904  0.2099930
```

Residuales PRESS

Los residuales estandarizados y los estudentizados son efectivos para detectar valores atípicos. Otro método para hacer que los residuales sean útiles en la determinación de valores atípicos consiste en examinar la cantidad que se calcula partiendo de $y_i - \hat{y}_{(i)}$, siendo $\hat{y}_{(i)}$, el valor ajustado de la i-ésima respuesta, basado en todas las observaciones excepto esa i-ésima. La lógica de este método es que si la i-ésima observación y_i realmente es atípica, el modelo de regresión basado en todas las observaciones estará demasiado influido por esta observación.

Si se elimina la i-ésima observación, se ajusta el modelo de regresión a las n - 1 observaciones restantes, y se calcula el valor predicho de y_i correspondiente a la observación omitida, el error de predicción correspondiente es:

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

Al principio, parecería que el cálculo de PRESS residuales requiere el ajuste de n regresiones diferentes. Sin embargo, es posible calcularlos a partir de los resultados de un solo ajuste por mínimos cuadrados de todas las n observaciones:

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \quad \text{i} = 1, 2, \ldots, \text{n}$$

El residual PRESS no es más que el residual ordinario ponderado por los elementos diagonales de la matriz de sombrero h_{ij}. Los residuales asociados con puntos para los que h_{ii} es grande tendrán PRESS residuales grandes. Esos puntos, en general, serán puntos de gran influencia. En general, una gran diferencia entre el residual ordinario y el PRESS residual indica un punto donde el modelo se ajusta bien a los datos, pero un modelo formado sin ese punto hace malas predicciones
La varianza del i-ésimo PRESS residual es:

$$\text{Var}\left[e_{(i)}\right] = Var\left[\frac{e_i}{1-h_{ii}}\right] = \frac{1}{(1-h_{ii})^2}[\sigma^2(1 - h_{ii})] = \frac{\sigma^2}{1-h_{ii}}$$

Por lo que un residual PRESS estandarizado es:

$$\frac{e_{(i)}}{\sqrt{Var[e_{(i)}]}} = \frac{e/(1-h_{ii})}{\sqrt{\sigma^2(1-h_{ii})}} = \frac{e}{\sqrt{\sigma^2(1-h_{ii})}}$$

```
#Residuo PRESS
rpress1<-resi/(1-hii)
head(rpress1)
        1         2         3         4         5         6
13.2765976  6.5630014  6.4548369  3.6890425 -4.3107406  0.7792807

rpress2<-rstandard(modelo2, type="pred")
head(rpress2)
        1         2         3         4         5         6
13.2765976  6.5630014  6.4548369  3.6890425 -4.3107406  0.7792807

#Otra forma eliminando la observación i
modelo.1<-lm(mpg ~ sp+wt+hp, data = caso[-1, ])
y.pred.1 <- predict(modelo.1, data.frame(caso[1, ]))
caso$mpg[1]-y.pred.1
```

3. Gráfico de residuales y Pruebas Estadísticas

3.1 Normalidad

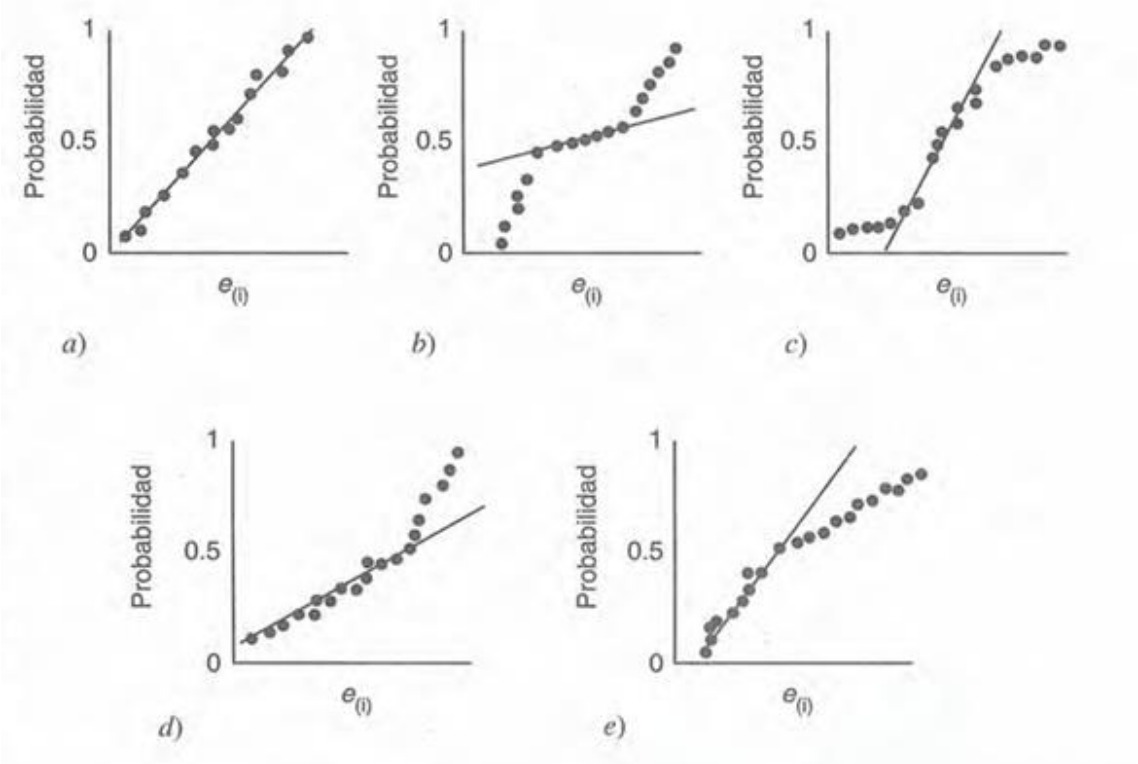
Gráfica de probabilidad normal

Las pequeñas desviaciones respecto a la hipótesis de normalidad no afectan mucho al modelo, pero una no. normalidad grande es potencialmente más seria, porque los estadísticos t o F y los intervalos de confianza y de predicción dependen de la suposición de normalidad.

Un método muy sencillo de comprobar la suposición de normalidad es trazar una gráfica de probabilidad normal de los residuales.

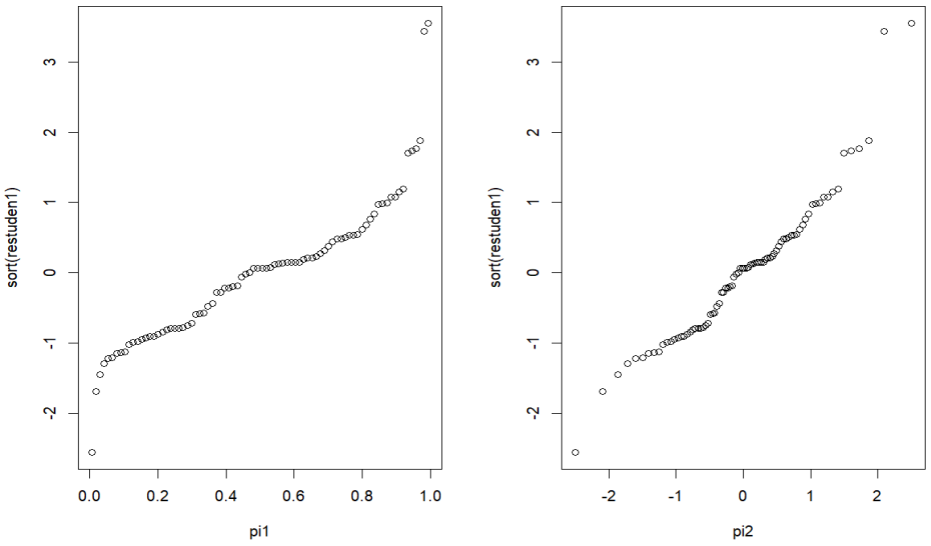
Sean $e_{[1]} < e_{[2]} < \ldots < e_{[n]}$ los residuales ordenados en orden creciente. Si se grafican $e_{[i]}$ en función de la probabilidad acumulada $P_i = (i - 1/2)/n$, i = 1,2, ... , n, en papel de probabilidad normal, los puntos que resulten deberían estar aproximadamente sobre una línea recta. Esa recta se suele de terminar en forma visual, con énfasis en los valores centrales. A veces, las gráficas de pro babilidad normal se trazan graficando el residual clasificado $e_{[i]}$ en función del "valor normal esperado", $\phi^{-1}(i - 1/2)/n]$, donde ϕ^{-1} representa la distribución acumulada normal estándar.

La gráfica a muestra una gráfica de probabilidad normal "idealizada". La parte b muestra una curva que va bruscamente hacia arriba y hacia abajo en los dos extremos, lo que indica que las colas de esta distribución son demasiado gruesas para poder considerarla como normal. Al contrario, la parte c muestra un aplanamiento en los extremos, que es un comportamiento característico de las muestras to madas de una distribución con colas más delgadas que la normal. Las partes d y e de la gráfica muestran patrones asociados con asimetría positiva y negativa, respectivamente



Fuente: Introducción al Análisis de Regresión (Montgomery, Peck y Vining)

```
n<-length(restuden1)
pi1<-(1:n-1/2)/n
pi2<-qnorm(pi1)
par(mfrow=c(1,2))
plot(pi1,sort(restuden1))
plot(pi2,sort(restuden1))
```



Prueba de Normalidad

H_0: Los errores se distribuyen normalmente
H_1: Los errores no se distribuyen normalmente

En la literatura en estadística existen muchas pruebas de normalidad, una de ellas es la Prueba de Shapiro Wilk, cuyo estadístico es dado por:

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i \left(X_{(n-i+1)} - X_{(i)}\right)\right]^2$$

$$D = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

Donde $a_1, a_2, \ldots, a_k$, son los coeficientes obtenidos de la distribución de Shapiro Wilk, donde k es aproximadamente $n/2$ y $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ son las estadísticas de orden.

```
shapiro.test(resi)

        Shapiro-Wilk normality test
data:  resi
W = 0.94851, p-value = 0.002442
```

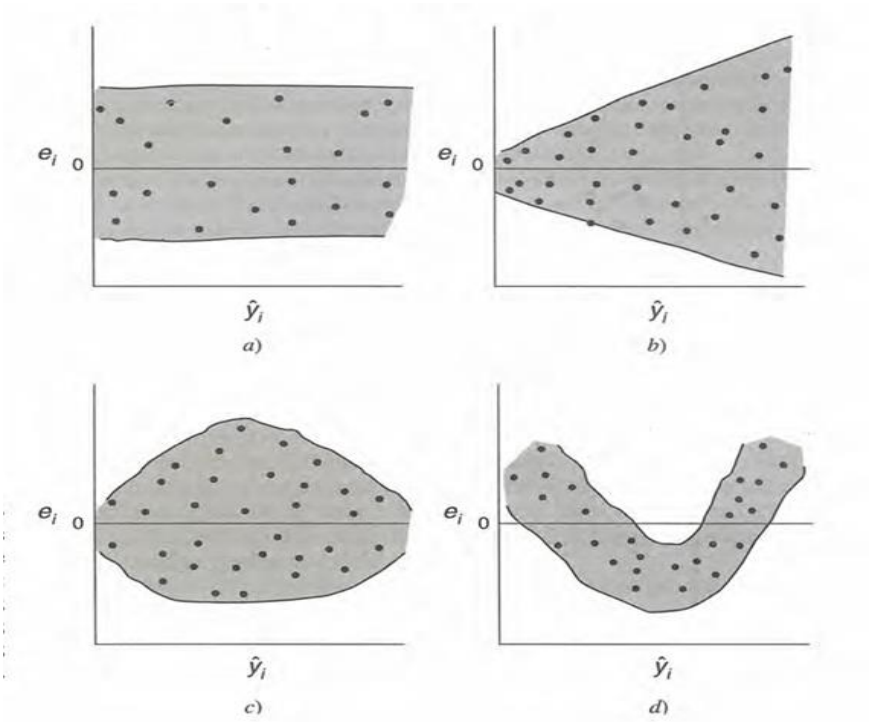
```
shapiro.test(resi[-c(1,8,29)])

        Shapiro-Wilk normality test

data:  resi[-c(1, 8, 29)]
W = 0.97424, p-value = 0.1102
```

Otras pruebas de normalidad son: Anderson-Darling, Jarque-Bera, Shapiro-Francia, Cramer von Mises, Pearson, entre otras.
El paquete nortest del R ofrece varias pruebas de normalidad.

3.2 Homogeneidad de varianzas

Gráfica de residuales en función de los valores ajustados

Las distribuciones en las partes b y e indican que la varianza de los errores no es constante. La figura de embudo abierto hacia afuera en la parte b implica que la varianza es función creciente de y. También es posible un embudo abierto hacia dentro, que indica que Var(ε) aumenta a medida que *y* disminuye.
La distribución en doble arco en la parte e se presenta con frecuencia cuando y es una proporción entre 0 y 1.
El método común pa ra manejar la no constancia de la varianza es aplicar una transformación adecuada ya sea a la variable regresora o a la de respuesta o usar el método de mínimos cuadrados ponderados.
Una gráfica en curva, como la de la parte d, indica no linealidad. Esto podría indicar que se necesitan otras variables regresoras en el modelo. Por ejemplo, podría ser necesario un término al cuadrado. Las transformaciones de la variable regresora y/o la de res puesta también podrían ayudar en estos casos.



Fuente: Introducción al Análisis de Regresión (Montgomery, Peck y Vining)

Prueba de Breusch-Pagan
H_0: Los errores son homocedasticos
H_1: Los errores no son homocedasticos

Consiste en ajustar un modelo de regresión lineal con variable respuesta dada por residuales del modelo original al cuadrado e_i^2 y como covariables las variables del modelo original.

$$\hat{e}_i^2 = \delta_0 + \delta_1 X_1 + \cdots + \delta_k X_k + u$$

Si se concluye que $\delta_1 = \cdots = \delta_k = 0$ significa que los residuales no son función de las covariables del modelo. El estadístico en esta prueba está dado por:

$$nxR^2 \sim \chi^2_{(1-\alpha,k)}$$

```
n<-nrow(caso)
ei <- resid(modelo2)
fit <- lm(ei^2 ~ sp + wt + hp, data=caso)
R2 <- summary(fit)$r.squared
k <- 3
estadistico <- n * R2
valorP <- pchisq(q=estadistico, df=k, lower.tail=FALSE)
cbind(estadistico, valorP)
     estadistico       valorP
[1,]    22.07817 6.283304e-05

bptest(modelo2)

        studentized Breusch-Pagan test

data:  modelo2
BP = 22.078, df = 3, p-value = 6.283e-05


library(lmtest)
bptest(modelo2)
modelo2s<-lm(mpg~sp+wt+hp,data=caso[-c(1,2,3,4,7,8,29,30),]
)
bptest(modelo2s)


studentized Breusch-Pagan test

data:  modelo2s
BP = 7.6315, df = 3, p-value = 0.05428
```

Otras pruebas de homogeneidad son: White, Prueba de Score (ncvtest del paquete car), Goldfeld-Quandt (gqtest del paquete lmtest) , Harrison-McCabe (hmctest del paquete lmtest), entre otras.

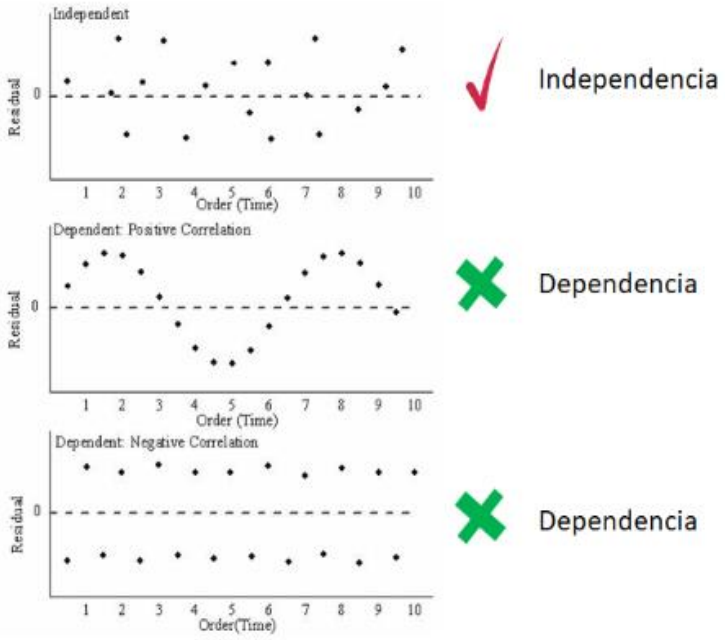
3.3 Independencia

Gráfico de residuales en el tiempo

Si se conoce la secuencia temporal de recolección de los datos, se aconseja graficar los residuales en función de su orden en el tiempo.

La gráfica de residuales en secuencia temporal puede indicar que los errores en un periodo se correlacionan con los de otros periodos. La correlación entre los errores del modelo en distintos periodos se llama autocorrelación.

En el caso ideal, esa gráfica se parecerá a la de la figura a (del gráfico anterior), esto es, una banda horizontal que abarca todos los residuales, y los residuales varían en forma más o menos aleatoria dentro de esa banda. Sin embargo, si la gráfica se asemeja a las pautas de la figura b-d, puede indicar que la varianza está cambiando con el tiempo, o que se deben agregar al modelo términos lineales o cuadráticos.



Prueba de Durbin-Watson

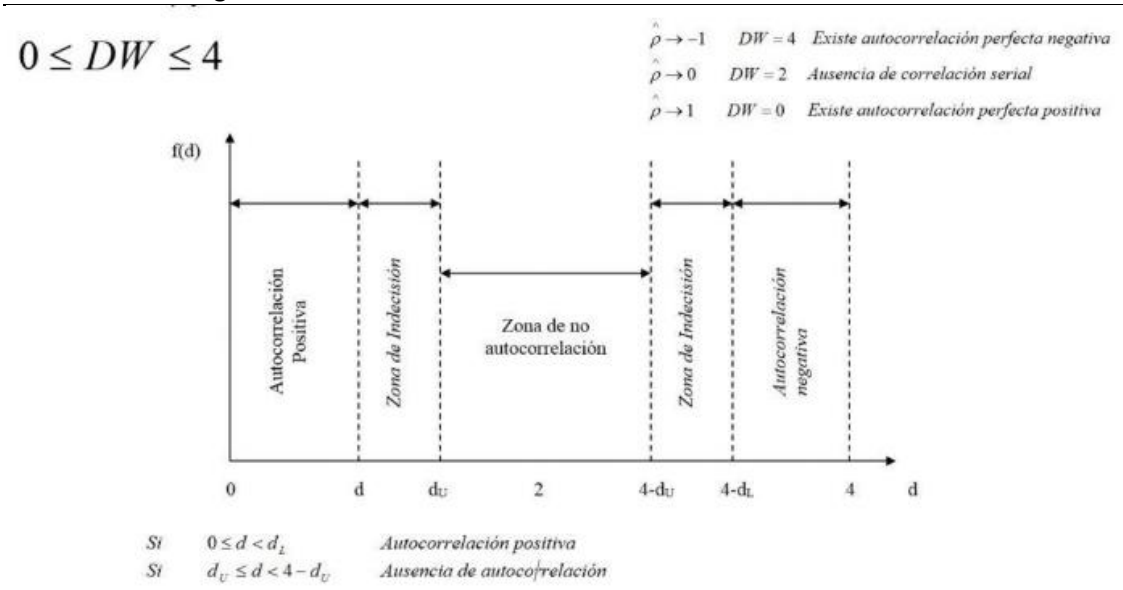
H_0: Los errores son independientes

H_1: Los errores no son independientes

Permite ver si los valores presentan algún tipo de dependencia en cuanto al orden de obtención

Es un estadístico que varía entre 0 y 4.

$$DW = \frac{\sum_{i=2}^{n}(\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n}\hat{e}_i^2}$$

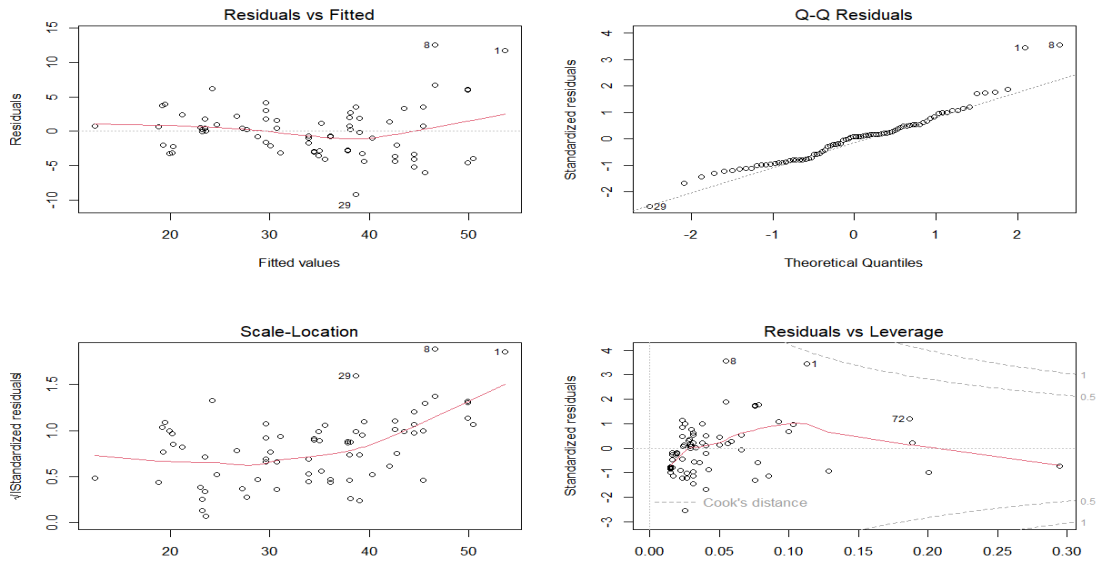



```
library(lmtest)
dwtest(modelo2,alternative="t")
```

```
        Durbin-Watson test

data:  modelo2
DW = 1.1846, p-value = 4.224e-05
alternative hypothesis: true autocorrelation is not 0
```

Otra prueba de independencia (autocorrelación) es: Breusch-Godfrey que se encuentra en R en la función gbtest del paquete lmtest.

```
par(mfrow=c(2,2))
plot(modelo2)
```



4. La estadística PRESS

Anteriormente se definieron los residuales PRESS como:

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

Siendo $\hat{y}_{(i)}$ el valor predicho dela i-ésima respuesta observada, basado en un ajuste de modelo con los n - 1 puntos de muestra.
Los residuales PRESS grandes tienen una utilidad potencial para identificar observaciones donde el modelo no se ajusta bien a los datos, o para observaciones en las que es probable que el modelo produzca malas predicciones

$$PRESS = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2$$

$$= \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Se considera que PRESS es una medida de lo bien que funciona un modelo de regresión para predecir nuevos datos. Lo deseable es tener un modelo con valor pequeño de PRESS

R^2 para predicción basado en PRESS

Con la estadística PRESS se puede calcular un estadístico parecido a la R2 para predicción,

$$R^2_{predicción} = 1 - \frac{PRESS}{SS_T}$$

Una aplicación muy importante de la estadística PRESS es comparar modelos de regresión. En general, un modelo con pequeño valor de PRESS es preferible a uno con PRESS grande.
Una función en R que permite obtener ols_pred_rsq, del paquete olsrr

```
modelo2 <- lm(mpg~sp+wt+hp, data = caso)
olsrr::ols_pred_rsq(modelo2)

[1] 0.8544887

rpress <- rstandard(modelo2, type = "pred")
scp <- sum(rpress^2)
sct <- sum(anova(modelo2)$'Sum Sq')
(R2_pred <- 1 - scp/sct)

[1] 0.8544887
```

5. Detección de valores atípicos

Un valor atípico es una observación extrema. Los residuales cuyo valor absoluto es bastante mayor que los demás, digamos de tres a cuatro desviaciones estándar respecto a la media, indican que hay valores atípicos potenciales en_el espacio de y. Los valores atípicos son puntos que no son representativos del resto de los datos.

Las gráficas de residuales en función de $\hat{y}_i$ y la gráfica de probabilidad normal son útiles para identificar puntos atípico.

Los valores atípicos se deben investigar con cuidado, para ver si se puede encontrar una razón de su comportamiento extraordinario. A veces, los valores atípicos son "malos" y se deben a eventos desacostumbrados, pero explicables. Es claro que el eliminar valores malos es conveniente, porque los mínimos cuadrados jalan la ecuación ajustada hacia el valor atípico, ya que eso minimiza la suma de cuadrados de residuales, sin embargo, se hace notar que debe contarse con una fuerte evidencia no estadística de que el valor atípico es malo, para entonces descartarlo.

El efecto de los valores atípicos sobre el modelo de regresión se puede comprobar con facilidad eliminándolos y volviendo a ajustar la ecuación de regresión. Se podrá encontrar que los valores de los coeficientes de regresión, o de los estadísticos de resumen como t, F o R^2, y que el cuadrado medio de residuales pueden ser muy sensibles a los valores atípicos.

```
residuos_estandarizados <- rstandard(mod1)
#Identificar casos con residuos estandar. mayores a |3|
casos_atipicos <- which(abs(residuos_estandarizados) > 3)
# Visualizar los valores de los residuos estandarizados
residuos_atipicos<- residuos_estandarizados[casos_atipicos]
print(residuos_atipicos)
```