Unidad 9: Modelamiento Estadístico 2: Requisitos

Requisitos del modelo lineal

Una serie de requisitos estadísticos nos prueban que el modelo lineal que hemos diseñado sea sólido matemáticamente. Estas pruebas se desarrollan debido a que nuestra base de datos puede presentar problemas en las variables, en los casos o teóricamente. Primero desarrollemos los requisitos relacionados con los casos: valores extremos, atípicos y palancas, con el modelo que ya se ha creado. A este proceso lo llamaremos *ubicación de influyentes*. La primera prueba de significancia que nos confirmaría la existencia de observaciones atípicas con un valor menor a 0.05 es el *outlierTest.*

Abriendo base de datos de indicadores financieros de la Unión Europea. Se utiliza el paquete *foreign* para exportar data de SPSS.

```
library(foreign)
data2<-read.spss("UE.sav", to.data.frame=TRUE, use.value.labels = TRUE)
outlierTest(modelo2) #detectando el caso outlier
```

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted     p-value    Bonferonni p
10 -2.599326                0.018706      0.41153
```

```
which.max(hatvalues(modelo2)) #o
6
4
which.max(cooks.distance(modelo2))
10
 8
```

Luego de localizar los influyentes -8 y 10- es necesario eliminarlos y crear un nuevo modelo sin ellos. Este modelo presentará cambios en su estructura.

```
ue<-data[-8,]  #Eliminando el Outlier
ue[7:11, ] #Observando los datos sin el Outlier
```

```
7  Lithuania
9  Malta
10 Portugal
11 Slovenia
12 Belgium
```

```
ue1<-data[-10,] #Eliminando el Outlier
ue1[7:11] #Observando los datos sin el Outlier
```

```
7    0.187
8    0.072
9    0.010
11   0.102
12   7.313
```

```
modelo3<- lm(GDPDeflator ~ TotalReservesMinusGold + ConsumerPrices +Governmentbonds,
data=ue1)
summary(modelo3)
```

```
Call:
lm(formula = GDPDeflator ~ TotalReservesMinusGold + ConsumerPrices +
   Governmentbonds, data = ue1)
Residuals:
   Min     1Q  Median     3Q    Max
-1.2115 -0.5592 -0.2232  0.5573  1.6405
Coefficients:
                            Estimate         Std. Error         t value   Pr(>|t|)
(Intercept)                 -2.792e+01       2.164e+01          -1.290    0.2142
TotalReservesMinusGold -2.521e-05            1.080e-05          -2.335    0.0321 *
ConsumerPrices              1.268e+00        2.117e-01          5.987     1.47e-05 ***
Governmentbonds             -8.518e-02       1.140e-01          -0.747    0.4652
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.8792 on 17 degrees of freedom

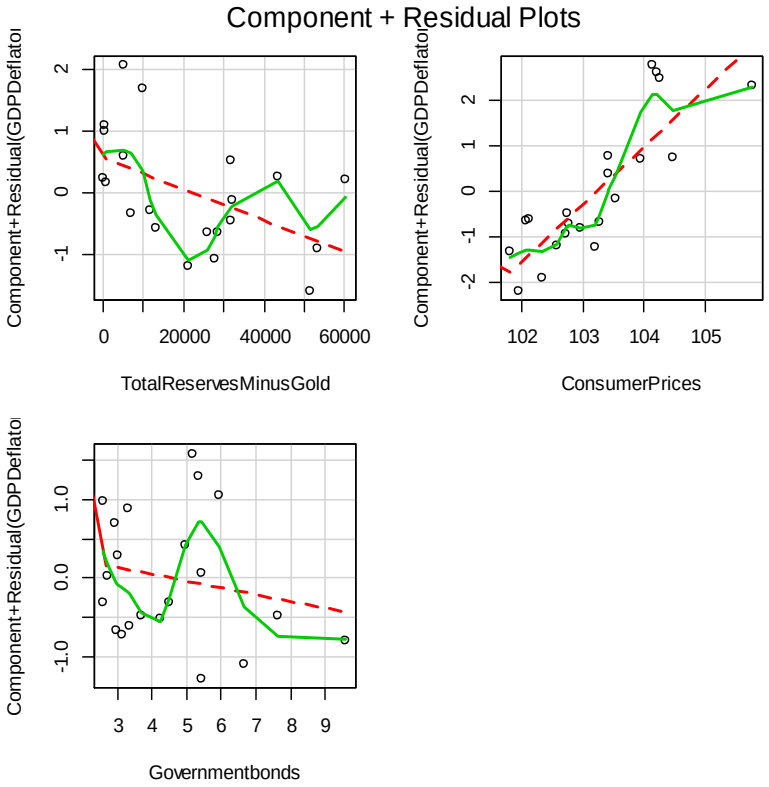
(8 observations deleted due to missingness)

Multiple R-squared: 0.6972, Adjusted R-squared: 0.6437

F-statistic: 13.05 on 3 and 17 DF, p-value: 0.0001141

Ahora que se han eliminado los influyentes, el modelo ha pasado los requisitos para los casos y podemos continuar con los requisitos relacionados con las variables: linealidad y multicolinealidad. El primer requisito se analiza gráficamente y el segundo, se verifica con un valor ubicado en una escala entre el 1 y el 10 con los comandos *cr.plots* y *vif.*

```
cr.plots(modelo3, one.page=TRUE, ask=FALSE) #pidiendo el requisito de linealidad
```




```
vif(requisitos1.m2) #pidiendo el requisito de multicolinealidad
```

TotalReservesMinusGold	ConsumerPrices	Governmentbonds
1.074507	1.155637	1.149094

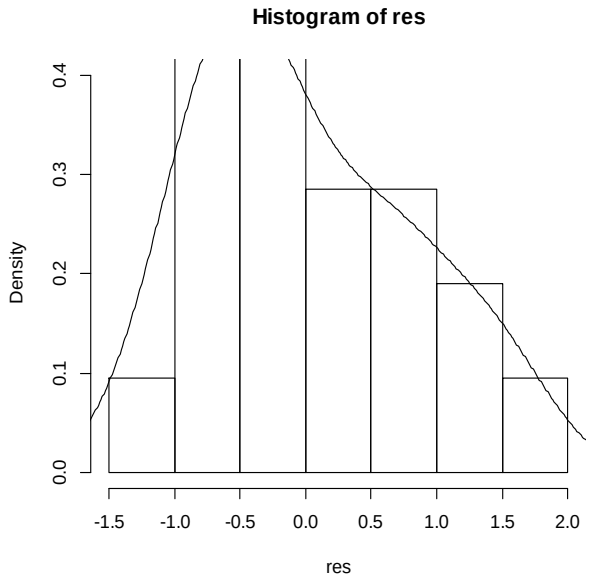
```
sqrt(vif(requisitos1.m2))) #verificando el requisito
```

TotalReservesMinusGold	ConsumerPrices	Governmentbonds

FALSE FALSE FALSE

Se puede verificar que la relación entre cada variable independiente y la dependiente tiene una forma lineal. Mientras tanto, para la multicolinealidad el resultado debe ubicarse entre 1 y 10. En nuestro modelo, los resultados cumplen este requisito -1.074507. 1.155637.1.149094- y la prueba siguiente nos confirma que no existe colinealidad entre las variables. Por último, la normalidad de los requisitos se puede comprobar con una prueba de significancia y apoyándonos también en un gráfico.

```
res<-residuals(modelo3)
hist(res,prob=TRUE,ylim=c(0,0.4)) ; lines(density(res)) #pidamos un histograma de los residuos
```




```
install.packages("nortest") #instalemos el paquete nortest para observar la normalidad
library(nortest) #llamemos a la librería
sf.test(res) #pidamos el test de Shapiro, el tipo de prueba depende de la cantidad de casos.
```

Shapiro-Francia normality test

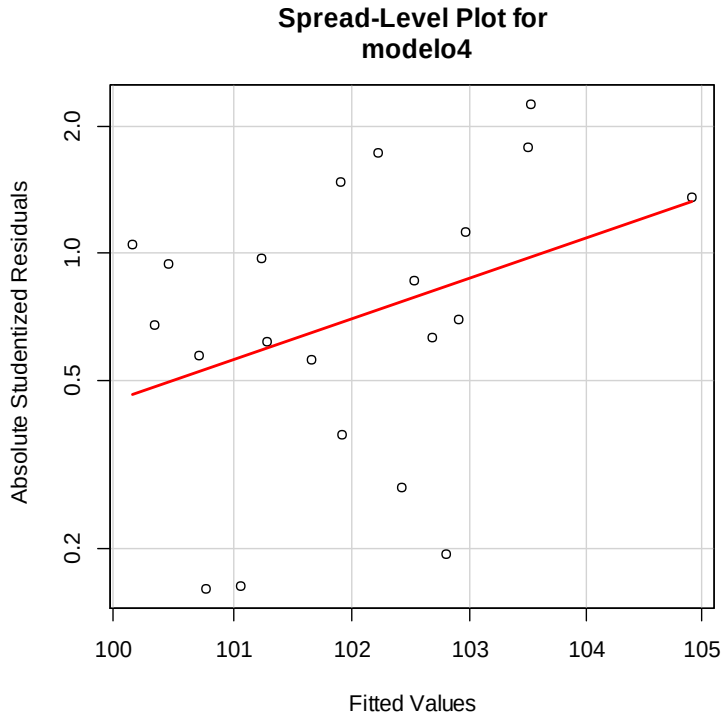
data: res

W = 0.9555, **p-value = 0.3648**

Podemos afirmar se acepta la H0 de normalidad debido a que la significancia es mayor a 0.05. De esta forma, concluimos que existe normalidad en los errores del modelo. Ahora probemos

algunos problemas teóricos como la *homocedasticidad* y la *independencia de errores* con los comandos *spread.level.plot* y *dwtest*.

```
spread.level.plot(modelo3)
```




```
ncvTest(modelo3)
```

Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 2.306116 Df = 1 **p = 0.1288657**

```
dwtest(modelo3)
```

Durbin-Watson test

data: modelo4

DW = 1.7343, **p-value = 0.2069**

alternative hypothesis: true autocorrelation is greater than 0

La primera prueba sobre *homocedasticidad* nos señala que la significancia es mayor a 0.05; por lo tanto, podemos aceptar la H0. Concluimos, entonces, que existe homocedasticidad en las variables. La segunda prueba sobre *independencia de errores* nos muestra también una significancia mayor a 0.05; por lo tanto, se acepta la H0 y se concluye que no existe

autocorrelación. Estos son los clásicos requisitos que un modelo lineal debe superar para considerársele como uno válido.

Ejercicios

1. La literatura sobre comportamiento electoral, nos muestra distintas perspectivas para entender el comportamiento de los votantes. En el 2013, Andy Baker y Kenneth F. Greene analizaron la existencia de la relación entre algunos issues importantes y los candidatos a la presidencia de algunos países de América Latina y detectaron que los votantes latinoamericanos suelen generar relaciones entre estos a la hora de votar. En ese sentido, y aterrizando, a las elecciones regionales y municipales 2014 en el Perú, podemos observar el caso de la municipalidad de Lima, cuyo candidato ganador fue el exalcalde Luis Castañeda Lossio, con resultados abrumadores respecto a sus más cercanos competidores.

Ahora bien, en el presente ejercicio usted debe buscar los issues que más se relacionan al candidato Castañeda, para ello utilizará la base de datos "VotaciónCastañeda", que proviene de la encuesta "Lima como vamos 2014" del IOP-PUCP, cuyo muestreo nos permite observar los resultados por cada distrito. La pregunta que usaremos para observar los issues prioritarios por cada distrito es: ***¿Cómo califica su nivel de satisfacción según aspectos que influyen en la calidad de vida en la ciudad de Lima?***

Cómo podemos observar esta pregunta busca dilucidar cuál es el nivel de satisfacción de los encuestados respecto a los problemas más comunes en Lima. Asimismo, observamos el informe de Ipsos en Setiembre del 2014 para ver qué issues se relacionan con las propuestas de cada candidatos, en el caso de la votación para Castañeda se observa que se le asocia primero, el transporte como el área principal en que se enfocan sus propuestas; seguido de Seguridad Ciudadana; Limpieza Pública; Comercio Ambulatorio e informalidad; Cultura y poder; y, finalmente, medio ambiente y áreas verdes. De esta forma, identificaremos algunos issues que coinciden, tanto en la relación con las propuestas de Castañeda y la insatisfacción por cada distrito sobre este punto.

Los cinco issues que tomaremos en cuenta son tráfico; seguridad; parques y áreas verdes; economía e informalidad y actividades culturales, que cumplirán las funciones de variables independientes, y tendrá que realizar una regresión lineal para observar si es que existe respuestas de causalidad entre el voto a Castañeda y la insatisfacción por los issues en cada distrito. De esta forma, tendrá que observar las variables significativas y cuanto explica la regresión.

2. Después de observar los resultados del ejercicio 1, usted tendrá que ver si la regresión lineal logra pasar los requisitos de casos y variables. De esta forma podrá identificar un caso outlier y tendrá que realizar la nueva regresión sin la presencia de esté; y observar cuanto explican, si existen algunos cambios en las variables significativas y, además, verificar nuevamente los requisitos.