Proyecto de Regresión Lineal Multiple

Romero Tapia Luis Donaldo
Rosas Vargas Pilar Issamara
Tapia Huerta Beatriz
Villegas Moctezuma Angel Alejandro

07 de Junio del 2019

Base de datos

Para realizar el análisis de las variables socioeconómicas primero se procedio a cargar ambas bases y limpiarlas para poder crear una conjunta.
Primero se trabajo con la base del INEGI, esta tenia los subtotales por entidad federativa y total del país, al estudiarla observamos que el codigo para decodificarlos era $CVE_DISTRITO = 0$ por lo que encuentra el número de coincidencias en toda la base y se retiran.

```
> INEGI<-as.data.frame(read.csv("~/INEGI/eiege_eic_2015.csv",sep = ",",stringsAsFactors = F))
> #sabemos que CVE_DISTRITO=0 implica un total, entonces retiramos esas filas
> table(INEGI$CVE_DISTRITO[]==0) #coincidencias con cero

FALSE  TRUE
  300    33

> aux=rep(0,33) #numero de incidencias de totales
> j=1
> for( i in 1:nrow(INEGI)){
+   if(INEGI$CVE_DISTRITO[i] == 0){
+     aux[j]=i
+     j=j+1
+   }
+ }
> INEGI=INEGI[-aux,] #quitamos subtotales
> table(INEGI$CVE_DISTRITO[]==0) #corrobora que ya no haya ningun cero

FALSE
  300

>
```

Después se realiza el ID unico para cada distrito, que comprende de dos digitos para el Estado y dos dígitos para el distrito interno, por lo que primero hacemos que todos distritos y Estados sean de dos dígitos y luego se juntan con el comando ”unite”

```
> for(i in 1:nrow(INEGI)){
+   if(as.numeric(INEGI$ï..CVE_ENT[i])< 10){
+     INEGI$ï..CVE_ENT[i]<-paste0(0,INEGI$ï..CVE_ENT[i])
+   }
+   if(as.numeric(INEGI$CVE_DISTRITO[i])<10){
+     INEGI$CVE_DISTRITO[i]<-paste0(0,INEGI$CVE_DISTRITO[i])
+   }
+ }
> #creamos el nuevo ID
> library(tidyr)
> INEGI<- unite(INEGI,ID,c(1:2))
```

Posteriormete se quitaron los intervalos de confianza de cada variable asi como los errores de estimación.

```
> nom_col<-names(INEGI)
> aux_nom_col<-strsplit(nom_col,"_")
> a=0
> for( i in 1:length(aux_nom_col)){
+   if(length(aux_nom_col[[i]])==3){
+     a<-c(a,i)
+   }
+ }
> a=a[-1] #elimina valor de inicialización
> INEGI<-INEGI[,-a] #base slos indices
```

Por último al estudiar las tres primeras variables se observo que no eran númericas, por lo que se les dio una decodificación númerica para hacerlas comparables con las demás.

```
> #Casos indigenas
> for ( i in 1:nrow(INEGI)){
+   if(INEGI$Indigena[i]=="NO"){
+     INEGI$Indigena[i]=0
+   }
+   else{
+     INEGI$Indigena[i]=1
+   }
+ }
> INEGI$Indigena<-as.numeric(INEGI$Indigena)
> for ( i in 1:nrow(INEGI)){
+   if(INEGI$MI[i]=="*"){
+     INEGI$MI[i]=1
+   }
+   else{
+     INEGI$MI[i]=0
+   }
+ }
> INEGI$MI<-as.numeric(INEGI$MI)
> categorias<-names(table(INEGI$Complejidad))
> for ( i in 1:nrow(INEGI)){
+  for (j in 1:length(categorias)){
+    if(INEGI$Complejidad[i]==categorias[j]){
+      INEGI$Complejidad[i]=j
+    }
+  }
+ }
> INEGI$Complejidad<-as.numeric(INEGI$Complejidad)
```

Una vez que la base del INEGI estaba preparada se procedio a trabajar con la del INE y se seleccionarón las variables que eran relevantes para el estudio.

```
> INE<-read.table("~/INEGI/Computos_Distritales_Presidente_2012.txt",sep="|",header = T) # cargamos la
> INE_votos=INE[,c(1,2,15:31)] #base con la que vamos a trabajar
> a=0
```

Después se quitaron los distritos cuyo número de identificación faltaba i.e. los valores faltantes.

```
> for( i in 1:nrow(INE_votos)){
+   if(is.na(INE_votos$DISTRITO_FEDERAL_2017[i]) == T){
+    a=c(a,i)
+   }
+ }
> a=a[-1] #quitamos el valor inicial fijado
> INE_votos=INE_votos[-a,] #quitamos los na
```

Puesto que las variables estaban por casilla y no por distrito se realizo el acumulado para que pudieran ser comparables con la base de INEGI

```
> INE_base_acumulada = data.frame() #definimos la nueva base
> k=1
> for(i in 1:32){
+   est<-subset(INE_votos,INE_votos$ID_ESTADO == i)
+   for(j in 1:length(table(est$DISTRITO_FEDERAL_2017))){
+     aux<-subset(est,est$DISTRITO_FEDERAL_2017 == j)
+     INE_base_acumulada[k,c(1,2)]<- c(est$ID_ESTADO[i],aux$DISTRITO_FEDERAL_2017[1])
+     INE_base_acumulada[k,3:length(est)]<- colSums(aux[3:19],na.rm = T)
+     k=k+1
+   }
+ }
> colnames(INE_base_acumulada)<-names(INE_votos) #Cambiar nombres de la base acumulada
```

Al tener los votos por distrito se agregaron las cuatro variables nuevas que contienen el acumulado de votos por candidato según las coaliciones realizadas en 2012.

```
> INE_base_acumulada$EPN<-INE_base_acumulada$PRI + INE_base_acumulada$PVEM + INE_base_acumulada$PRI_PVE
> INE_base_acumulada$JVM<-INE_base_acumulada$PAN
> INE_base_acumulada$AMLO<-INE_base_acumulada$PRD + INE_base_acumulada$PT + INE_base_acumulada$MC + INE
> INE_base_acumulada$Quadri<-INE_base_acumulada$PANAL
> #suma de las columnas
> Total<-colSums(INE_base_acumulada,na.rm = T)#vemos que las sumas coincidan
```

Por último se genero el ID de forma análoga a la base del INEGI

```
> #hacemos que todos los ID sean de dos valores
> for(i in 1:nrow(INE_base_acumulada)){
+   if(as.numeric(INE_base_acumulada$ID_ESTADO[i])< 10){
+     INE_base_acumulada$ID_ESTADO[i]<-paste0(0,INE_base_acumulada$ID_ESTADO[i])
+   }
+   if(as.numeric(INE_base_acumulada$DISTRITO_FEDERAL_2017[i])<10){
+     INE_base_acumulada$DISTRITO_FEDERAL_2017[i]<-paste0(0,INE_base_acumulada$DISTRITO_FEDERAL_2017[i]
+   }
+ }
> #creamos el nuevo ID
> library(tidyr)
> INE_base_acumulada<- unite(INE_base_acumulada,ID,c(1:2))
```

Finalmente con ambas bases preparadas se procedio a crear una nueva tomando como referencia los ID creados.

```
> #juntamos las dos bases
> INEGI_INE<-data.frame()
> k=1
> for( i in 1:nrow(INEGI)){
+   for(j in 1:nrow(INE_base_acumulada)){
+     if(INEGI$ID[i]==INE_base_acumulada$ID[j]){
+       INEGI_INE[k,1]<-INE_base_acumulada$ID[j]
+       for(l in 2:5){
+         INEGI_INE[k,l]<-INE_base_acumulada[j,l+17]
+       }
+       for(l in 2:length(INEGI)){
+         INEGI_INE[k,l+4]<-INEGI[i,l]
+       }
+       k=k+1
+     }
+   }
+ }
> colnames(INEGI_INE)<-c("ID","EPN","JVM","AMLO","Quadri",names(INEGI[c(2:length(INEGI))]))
> #En teoria cada entrada deberia de coincidir con el orgininal, corroboremos
> comprob<-matrix(FALSE,nrow = 300,ncol=4)
> for(i in 1:nrow(INEGI_INE)){
+   comprob[i,1]<-INEGI_INE$EPN[i] == INE_base_acumulada$EPN[i]
+   comprob[i,2]<-INEGI_INE$JVM[i] == INE_base_acumulada$JVM[i]
```

```
+   comprob[i,3]<-INEGI_INE$AMLO[i] == INE_base_acumulada$AMLO[i]
+   comprob[i,4]<-INEGI_INE$Quadri[i] == INE_base_acumulada$Quadri[i]
+ }
> table(comprob) #observamos que todas las entradas coinciden

comprob
TRUE
1200
```

Para realizar el estudio se separó en la muestra (distritos) en dos conjuntos, uno de entrenamiento(90%) y otro de predicción(10%). Habiendo separado la base conjunta se procede a realizar un Modelo de Regresión Lineal Múltiple para cada uno de los candidatos.

Ahora, haremos un modelo para cada candidato EPN

```
Call:
lm(formula = INEGI_INE_TRAIN$EPN ~ ., data = INE1)

Residuals:
     Min       1Q   Median       3Q      Max
-17240.1  -4675.6    229.2   4402.9  17403.2

Coefficients: (8 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.084e+05  1.590e+06  -0.571 0.568596
MI          -4.282e+03  4.376e+03  -0.979 0.329218
Indigena     1.838e+03  5.292e+03   0.347 0.728837
Complejidad  1.252e+03  6.088e+02   2.057 0.041231 *
IND_001     -4.034e-01  6.737e-01  -0.599 0.550131
IND_002     -2.531e+03  1.436e+03  -1.763 0.079814 .
IND_003     -2.227e+03  9.326e+03  -0.239 0.811590
IND_004      2.027e+03  2.667e+03   0.760 0.448180
IND_005      5.798e+02  1.510e+03   0.384 0.701400
IND_006      9.331e+03  3.530e+04   0.264 0.791860
IND_007             NA         NA      NA       NA
IND_818      7.302e+03  1.089e+04   0.670 0.503503
IND_819      5.885e+03  1.036e+04   0.568 0.570894
IND_820      1.223e+04  9.996e+03   1.224 0.222781
IND_821      9.911e+03  9.763e+03   1.015 0.311522
IND_822      1.138e+04  9.822e+03   1.158 0.248340
IND_823      7.852e+03  9.862e+03   0.796 0.427073
IND_824      9.395e+03  9.834e+03   0.955 0.340757
IND_825             NA         NA      NA       NA
IND_800      5.946e+00  9.630e+00   0.617 0.537753
IND_801     -1.387e+04  2.080e+04  -0.667 0.505822
IND_802     -2.976e+00  1.577e+01  -0.189 0.850544
IND_803      1.719e+04  4.622e+04   0.372 0.710381
IND_804             NA         NA      NA       NA
IND_805             NA         NA      NA       NA
IND_806     -3.813e+00  9.203e+00  -0.414 0.679204
IND_807      9.615e+03  1.930e+04   0.498 0.618999
IND_808     -3.053e-02  1.534e+01  -0.002 0.998415
IND_809     -6.304e+03  4.218e+04  -0.149 0.881367
IND_810             NA         NA      NA       NA
IND_811             NA         NA      NA       NA
IND_047      4.872e-02  3.332e-01   0.146 0.883931
IND_049     -4.511e+03  4.347e+03  -1.038 0.300899
IND_050     -7.826e+03  6.177e+03  -1.267 0.206936
IND_051     -6.421e+03  9.661e+03  -0.665 0.507224
IND_048      4.764e+03  8.699e+03   0.548 0.584628
IND_053     -4.822e+02  1.318e+04  -0.037 0.970856
```

IND_054	1.801e+03	1.733e+03	1.039	0.300139	
IND_055	-7.987e-01	6.856e-01	-1.165	0.245652	
IND_138	2.335e+03	1.431e+03	1.631	0.104732	
IND_056	1.234e+04	1.949e+04	0.633	0.527613	
IND_057	-3.945e+04	1.393e+04	-2.832	0.005197	**
IND_061	-3.038e+02	1.078e+02	-2.819	0.005404	**
IND_062	3.338e+02	1.904e+02	1.753	0.081356	.
IND_063	-1.251e+03	3.586e+02	-3.487	0.000624	***
IND_064	1.810e+03	1.278e+03	1.417	0.158395	
IND_075	3.461e+03	3.533e+03	0.979	0.328766	
IND_076	3.955e+03	3.538e+03	1.118	0.265185	
IND_077	3.472e+03	3.547e+03	0.979	0.329099	
IND_078	3.429e+03	3.733e+03	0.918	0.359763	
IND_059	-1.879e+02	3.478e+02	-0.540	0.589832	
IND_060	-8.256e+01	3.885e+02	-0.212	0.831988	
IND_058	-6.230e+01	4.635e+02	-0.134	0.893248	
IND_065	-2.415e+02	3.663e+02	-0.659	0.510570	
IND_066	1.620e+02	1.063e+02	1.524	0.129333	
IND_067	-4.088e+01	1.034e+02	-0.395	0.693044	
IND_068	-1.307e+02	4.747e+02	-0.275	0.783482	
IND_069	-3.216e+02	2.419e+02	-1.330	0.185477	
IND_070	-5.191e+02	2.172e+02	-2.390	0.017954	*
IND_071	2.873e+03	1.800e+03	1.596	0.112342	
IND_072	1.544e+02	2.664e+02	0.579	0.563090	
IND_073	-7.221e+01	1.195e+02	-0.604	0.546605	
IND_074	-1.312e+02	6.366e+01	-2.062	0.040801	*
IND_079	1.521e+03	1.687e+03	0.902	0.368397	
IND_080	8.407e+02	2.052e+03	0.410	0.682564	
IND_082	-9.877e+01	2.098e+03	-0.047	0.962515	
IND_083	4.394e+02	2.105e+03	0.209	0.834874	
IND_085	-2.043e+02	1.749e+03	-0.117	0.907162	
IND_086	2.744e+02	1.111e+03	0.247	0.805181	
IND_087	-2.459e+02	1.358e+02	-1.811	0.071901	.
IND_088	1.423e+03	1.433e+03	0.993	0.322116	
IND_089	-5.310e+02	6.355e+02	-0.836	0.404618	
IND_090	1.374e+03	5.122e+02	2.683	0.008041	**
IND_091	-9.550e+02	8.768e+02	-1.089	0.277633	
IND_092	5.448e+02	8.581e+02	0.635	0.526368	
IND_093	-1.546e+02	4.328e+02	-0.357	0.721407	
IND_094	4.550e+02	2.619e+02	1.737	0.084176	.
IND_095	-5.889e+03	7.962e+03	-0.740	0.460527	
IND_096	4.782e+02	8.746e+02	0.547	0.585246	
IND_097	NA	NA	NA	NA	
IND_098	3.482e+03	1.211e+04	0.288	0.774083	
IND_099	-9.786e+02	8.372e+03	-0.117	0.907086	
IND_100	-3.532e+03	4.094e+03	-0.863	0.389557	
IND_139	-7.084e+03	7.969e+03	-0.889	0.375287	
IND_104	-4.821e+02	9.223e+02	-0.523	0.601875	
IND_105	-3.329e+02	5.852e+02	-0.569	0.570238	
IND_106	-8.247e+02	7.406e+02	-1.114	0.267036	
IND_107	5.472e+02	1.536e+03	0.356	0.722035	
IND_108	NA	NA	NA	NA	
IND_812	-3.932e+03	1.579e+03	-2.490	0.013760	*
IND_813	-4.120e+03	1.581e+03	-2.606	0.009977	**
IND_814	-3.859e+03	1.612e+03	-2.394	0.017776	*
IND_815	-2.613e+03	1.546e+03	-1.691	0.092769	.
IND_816	-3.876e+03	1.636e+03	-2.369	0.018989	*
IND_112	4.376e+03	9.319e+03	0.470	0.639304	
IND_113	4.425e+03	9.363e+03	0.473	0.637105	
IND_114	4.499e+03	9.320e+03	0.483	0.629955	
IND_115	1.664e+03	9.510e+03	0.175	0.861275	
IND_116	1.415e+04	9.848e+03	1.437	0.152536	
IND_117	9.390e+03	9.896e+03	0.949	0.344077	

```
IND_119      9.629e+02  2.420e+02   3.979 0.000103 ***
IND_123     -2.022e+02  7.483e+02  -0.270 0.787309
IND_121     -1.755e+01  7.435e+02  -0.024 0.981199
IND_122     -1.330e+02  7.453e+02  -0.178 0.858631
IND_124     -5.606e+02  8.157e+02  -0.687 0.492842
IND_120     -1.031e+03  7.146e+02  -1.443 0.150778
IND_125      2.667e+02  1.043e+03   0.256 0.798471
IND_141      2.450e+01  1.052e+02   0.233 0.816140
IND_126      1.273e+02  2.333e+02   0.545 0.586181
IND_127      9.098e+01  2.279e+02   0.399 0.690295
IND_128      6.910e+01  4.071e+02   0.170 0.865428
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7876 on 167 degrees of freedom
Multiple R-squared:  0.7841,        Adjusted R-squared:  0.6522
F-statistic: 5.945 on 102 and 167 DF,  p-value: < 2.2e-16
```

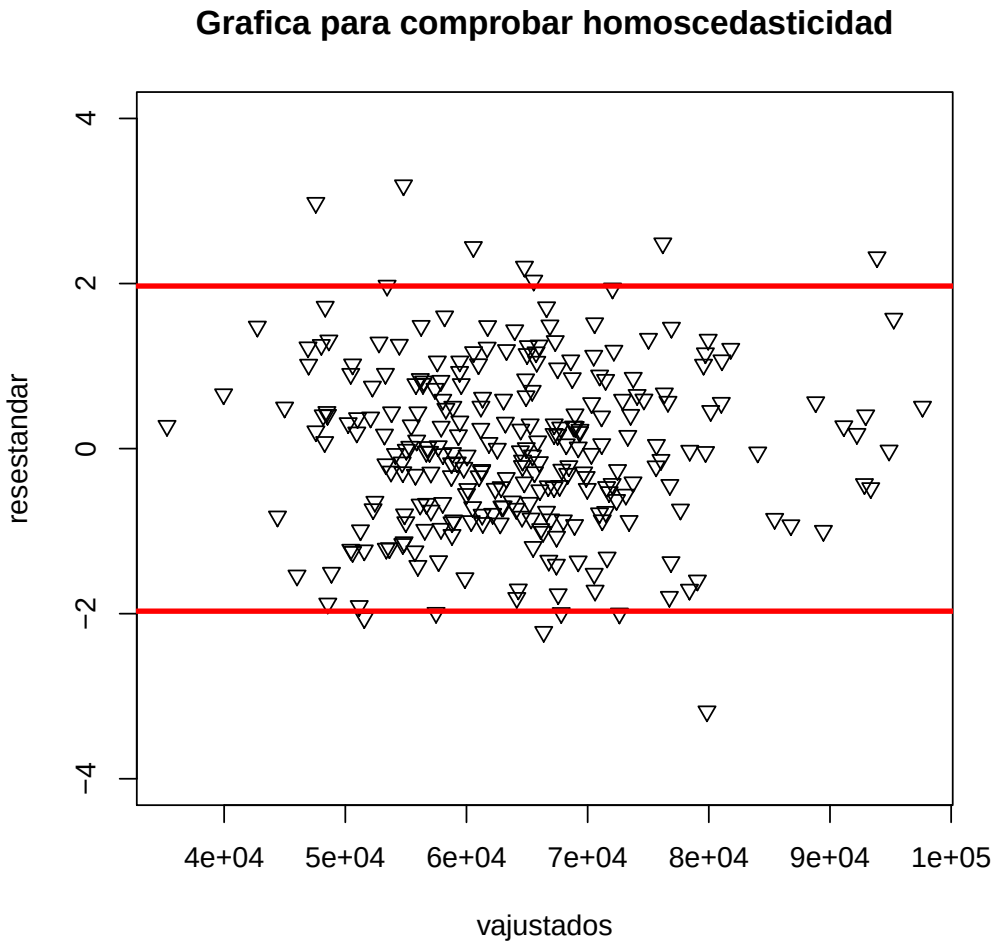
Notamos que esta prueba no es de ayuda en este momento pues existen variables que son combinaciones lineales de otras,así tenemos que construir EPN.vacio para tener el modelo vacio y el completo donde se integran todas las variables explicativas.Despues utilizaremos el metodo stepward,veamos el resultado del modelo obtenido y su respecivo VIF.

Podemos quitar mas variables, aunque teniendo cuidado en que mejore el modelo, quitemos de los que tienen mayor vif a los de menor, pero solo los que son mayores a 10.Cabe aqui recalcar una cosa mas, haciendo este metodo de ir quitando poco a poco las variables con ms VIf para as obtener un mejor modelo se llega a uno que ademsde tener la variable $IND07$ no significativa en el modelo , la mayoria de los supuestos no se cumplian , ni con transformaciones se pudo arreglar ese problema ,pues empezaan a aparecer mas variables no significativas, asi decidimos quitar la variable IND_107 desde el principio y ver a que se llega.
Eso nos dio la idea de que conforme obteniamos el mejor modelo por el metodo de ir quitando variables con vif mayor que 10 y teniendo en cuenta lo antes descrito ,nuevamente veiamos que tanto mejoraba los supuestos con y sin transformaciones , y nuevamente probando que variables podian ser retiradas desde el principio y nuevamnete repetir el proceso hasta llegar a este punto.
Asi en este primer paso veremos el Vif de $ENP1$ y quitaremos las variables que nos dieron un mejor modelo , despues procederemos como siempre quitando la variable de mayor vif y mayor que 10,entonces las variablesque que quitaremos de ENP son IND_107 y al final IND_088

En donde ya no tenemos variables con vif mayor a 10,exceptuando IND_070 con 10.3 , sin embargo tener esa variable mejora mucho nuestro modelo y tampoco es tan alejado de 10, veremos a continuacion la validacion de los supuestos.
Usamos nuevamente el metodo stepward, para encontrar un modelo mas chiquito con tope $EPN8\_52$.

$EPN9$ sera nuestro modelo final.Vemos que el vif de IND_070 es de 10.3 , creemos un precio bajo para que todo este en orden,ya con esto podemos ver que no hay muticolinealidad.
El modelo que nos quedo al final al final consta de las siguientes variables:
IND_115....Porcentaje de la poblacion de 12 años y mas separada.
IND_063....Porcentaje de viviendas con disponibilidad de servicio sanitario en la vivienda.
IND_802....Estimador total de población de 15 años y mas(Hombres).
IND_117....Porcentaje de la población de 12 años y más viuda.
IND_055....Estimador total de viviendas particulares habitadas.
IND_120....Porentaje de la pobacion afiliada a servicios medicos por seguro privado.
IND_057....Promedio de ocupantes por cuartos.
IND_070....Porcentaje de viviendas con telefono fijo.
IND_119....Porcentaje de poblacion afiliada a servicios de salud.
IND_087....Porcentaje de aistencia escolar dela población de $3a5$ años.
IND_813....Porcentaje de la poblacion ocupada que labora en el sector económico de minería,industriasmanufactureras,electric
y agua.
IND_125....Porcentaje de la población afiliada a servicios medicos por otra institución.
IND_076....Porcetaje de viviendas alquiladas. IND_061....Porcentaje de viviendas con disponibilidad de agua entuada en la vivienda. IND_062...Porcentaje de viviendas con disponibilidad de drenaje.
IND_126...Porcentaje de la poblacion de 3 años y mas que habla alguna legua indigena.
Complejidad....Grupo de complejidad electoral.
IND_815....Porcentaje de la poblacion ocupada que labora en el sector economico del comercio.
IND_124....Porcentaje de la población afiliada a Pmex,Defensa o Marina.
IND_002....Porcentaje estatal de la poblacion.

MI....Indicador de muestra insuficiente.
IND_116....Porcentaje de la población de 12 años y mas divorciada.
IND_094...Porcentaje de la poblacion de $15a24$ años que asiste a la escuela en un municipio o delegación distinto al de residencia.
Ahora verifiquemos los supuestos, veamos si la varianza es constante.Notemos que los residuos estandarizados son los res estandar y los valores ajustados son los valjustados.Graficamos para ver si la varianza es constante.



Podemos observar varianza constante ,a excepcion de algunos puntos fuera de la banda.Ahora hagamos la Prueba Homosedasticidad

```
        studentized Breusch-Pagan test

data:  EPN9
BP = 25.534, df = 23, p-value = 0.3233
```

Notemos que se cumple la homeostacidad deade que $p\_value = 0.3233 > 0.05$.Revisemos mas a fondo la multicolinealidad.

```
    IND_115         IND_063         IND_802          IND_117
 Min.   :2.016   Min.   :73.64   Min.   :103952   Min.   :2.138
 1st Qu.:3.478   1st Qu.:96.28   1st Qu.:129485   1st Qu.:4.123
 Median :4.438   Median :98.21   Median :136435   Median :4.706
 Mean   :4.377   Mean   :96.90   Mean   :138021   Mean   :4.720
 3rd Qu.:5.185   3rd Qu.:99.29   3rd Qu.:146446   3rd Qu.:5.449
 Max.   :7.447   Max.   :99.94   Max.   :181858   Max.   :7.133
    IND_055          IND_120            IND_057          IND_070
 Min.   : 71619   Min.   : 0.03301   Min.   :0.5967   Min.   : 2.823
 1st Qu.: 97937   1st Qu.: 0.99126   1st Qu.:0.9272   1st Qu.:22.014
 Median :105360   Median : 1.95799   Median :1.0334   Median :32.724
 Mean   :106464   Mean   : 3.22525   Mean   :1.0535   Mean   :35.727
 3rd Qu.:114598   3rd Qu.: 3.70488   3rd Qu.:1.1553   3rd Qu.:47.163
 Max.   :159700   Max.   :25.57733   Max.   :1.8221   Max.   :83.292
    IND_119         IND_087         IND_813          IND_125
```

```
 Min.   :66.68   Min.   :36.88   Min.   : 2.151   Min.   :0.1060
 1st Qu.:78.09   1st Qu.:57.96   1st Qu.: 9.739   1st Qu.:0.6428
 Median :82.59   Median :63.05   Median :14.371   Median :1.1821
 Mean   :82.10   Mean   :63.38   Mean   :15.973   Mean   :1.5510
 3rd Qu.:86.19   3rd Qu.:68.91   3rd Qu.:20.392   3rd Qu.:2.1583
 Max.   :93.87   Max.   :87.00   Max.   :53.606   Max.   :5.9547
    IND_076          IND_061         IND_062         IND_126
 Min.   : 1.381   Min.   :15.35   Min.   :39.08   Min.   : 0.09395
 1st Qu.: 9.785   1st Qu.:56.63   1st Qu.:91.79   1st Qu.: 0.68481
 Median :14.653   Median :78.16   Median :96.86   Median : 1.38537
 Mean   :15.644   Mean   :72.56   Mean   :92.57   Mean   : 6.66060
 3rd Qu.:20.251   3rd Qu.:90.95   3rd Qu.:98.61   3rd Qu.: 4.10891
 Max.   :43.432   Max.   :99.21   Max.   :99.80   Max.   :80.57921
  Complejidad       IND_815          IND_124           IND_002
 Min.   :1.000   Min.   : 5.486   Min.   : 0.01148   Min.   : 1.955
 1st Qu.:3.000   1st Qu.:15.538   1st Qu.: 0.19655   1st Qu.: 4.672
 Median :5.000   Median :17.809   Median : 0.49515   Median : 8.145
 Mean   :4.663   Mean   :18.067   Mean   : 1.25571   Mean   :10.826
 3rd Qu.:6.000   3rd Qu.:20.306   3rd Qu.: 1.26197   3rd Qu.:13.535
 Max.   :9.000   Max.   :28.512   Max.   :15.13966   Max.   :53.789
      MI_1          IND_116         IND_094
 Min.   :1.000   Min.   :0.1883   Min.   : 0.03129
 1st Qu.:1.000   1st Qu.:0.9072   1st Qu.: 2.91447
 Median :1.000   Median :1.4431   Median : 6.71988
 Mean   :1.033   Mean   :1.5776   Mean   : 9.65374
 3rd Qu.:1.000   3rd Qu.:2.1404   3rd Qu.:15.16074
 Max.   :2.000   Max.   :4.6544   Max.   :42.77394
```

	IND_115	IND_063	IND_802	IND_117	IND_055
IND_115	1.00000000	0.36381514	0.12703169	-0.018787418	0.179799899
IND_063	0.36381514	1.00000000	0.29941085	-0.145436701	0.284393825
IND_802	0.12703169	0.29941085	1.00000000	-0.228966498	0.880448633
IND_117	-0.01878742	-0.14543670	-0.22896650	1.000000000	-0.095797325
IND_055	0.17979990	0.28439382	0.88044863	-0.095797325	1.000000000
IND_120	0.05458384	0.31909984	0.31623671	0.005467049	0.346212588
IND_057	-0.16056226	-0.60043339	-0.42436581	-0.083821798	-0.460915285
IND_070	0.28039907	0.51179075	0.23916369	0.116533839	0.203046983
IND_119	-0.42006816	-0.29355644	0.09186241	-0.043092715	0.107615008
IND_087	-0.09885919	-0.09028746	-0.01071099	0.385671410	0.013031821
IND_813	-0.17639761	0.14866084	0.21067698	-0.413338898	0.136069575
IND_125	0.19572105	0.30575974	0.23343989	-0.101517593	0.173506988
IND_076	0.30705240	0.52324551	0.19745978	-0.132703477	0.255209344
IND_061	0.18237472	0.56070458	0.42992409	-0.284810207	0.409647408
IND_062	0.40213404	0.57847965	0.39917637	-0.327555746	0.325049835
IND_126	-0.26502092	-0.42719654	-0.29502068	0.131227967	-0.263095993
Complejidad	-0.36174474	-0.45392483	-0.18351985	0.043659470	-0.079774776
IND_815	0.47769086	0.46592594	0.14181094	-0.075568724	0.043395180
IND_124	0.23006454	0.09199288	0.03332581	0.102184064	0.111500978
IND_002	-0.04057303	-0.06300091	0.16137290	-0.229679809	0.206663133
MI_1	-0.04685818	-0.20261459	-0.07357361	0.009836568	0.007785231
IND_116	0.23348308	0.46324976	0.40871091	0.017852166	0.486035738
IND_094	0.21741896	0.19726972	0.07395577	0.130459828	0.007660409

	IND_120	IND_057	IND_070	IND_119	IND_087
IND_115	0.054583839	-0.16056226	0.280399072	-0.42006816	-0.098859189
IND_063	0.319099839	-0.60043339	0.511790751	-0.29355644	-0.090287456
IND_802	0.316236709	-0.42436581	0.239163687	0.09186241	-0.010710993
IND_117	0.005467049	-0.08382180	0.116533839	-0.04309271	0.385671410
IND_055	0.346212588	-0.46091528	0.203046983	0.10761501	0.013031821
IND_120	1.000000000	-0.60290352	0.676993496	-0.11665881	0.181624376
IND_057	-0.602903524	1.00000000	-0.774541295	0.16562445	-0.051940179
IND_070	0.676993496	-0.77454129	1.000000000	-0.36103483	0.196756884
IND_119	-0.116658813	0.16562445	-0.361034825	1.00000000	0.113897276
IND_087	0.181624376	-0.05194018	0.196756884	0.11389728	1.000000000
IND_813	0.019400737	-0.22504702	0.054861111	0.11955494	-0.434681767

```
IND_125      0.372805723 -0.39818205  0.450900001 -0.20218303 -0.085821327
IND_076      0.532628995 -0.54625062  0.709812529 -0.40122672  0.036393657
IND_061      0.511692443 -0.75372345  0.717611523 -0.16046530 -0.075131531
IND_062      0.367353437 -0.58529187  0.602374152 -0.30638953 -0.079130795
IND_126     -0.247503898  0.60060230 -0.469833437  0.15764729  0.194778526
Complejidad -0.407287894  0.45796470 -0.699337973  0.45999876 -0.084336580
IND_815      0.122151216 -0.46706506  0.591466668 -0.57047727 -0.056540613
IND_124      0.040944611 -0.04250351  0.003287463 -0.08320115  0.126732758
IND_002     -0.138129328  0.03179338 -0.200798096  0.37362299  0.004194826
MI_1        -0.056319536 -0.01043992 -0.109235304 -0.06551099 -0.115813648
IND_116      0.636226843 -0.79315528  0.783149190 -0.09232495  0.135552879
IND_094      0.338117803 -0.29624537  0.453753141 -0.31258655  0.195163867
                IND_813     IND_125     IND_076     IND_061     IND_062
IND_115     -0.17639761  0.19572105  0.30705240  0.18237472  0.40213404
IND_063      0.14866084  0.30575974  0.52324551  0.56070458  0.57847965
IND_802      0.21067698  0.23343989  0.19745978  0.42992409  0.39917637
IND_117     -0.41333890 -0.10151759 -0.13270348 -0.28481021 -0.32755575
IND_055      0.13606957  0.17350699  0.25520934  0.40964741  0.32504984
IND_120      0.01940074  0.37280572  0.53262899  0.51169244  0.36735344
IND_057     -0.22504702 -0.39818205 -0.54625062 -0.75372345 -0.58529187
IND_070      0.05486111  0.45090000  0.70981253  0.71761152  0.60237415
IND_119      0.11955494 -0.20218303 -0.40122672 -0.16046530 -0.30638953
IND_087     -0.43468177 -0.08582133  0.03639366 -0.07513153 -0.07913079
IND_813      1.00000000  0.14524630  0.04460811  0.37191682  0.25039402
IND_125      0.14524630  1.00000000  0.30901367  0.36936029  0.32067860
IND_076      0.04460811  0.30901367  1.00000000  0.71381887  0.61842448
IND_061      0.37191682  0.36936029  0.71381887  1.00000000  0.76418305
IND_062      0.25039402  0.32067860  0.61842448  0.76418305  1.00000000
IND_126     -0.28927000 -0.25359545 -0.39778054 -0.60028198 -0.73622804
Complejidad -0.09561109 -0.42232356 -0.58931398 -0.53361549 -0.53772841
IND_815      0.05713080  0.32390125  0.52990734  0.54299820  0.58837701
IND_124     -0.10774113  0.03053262  0.05365456 -0.03093136  0.10670120
IND_002     -0.01436655 -0.32178070 -0.05996305  0.06999095  0.03399314
MI_1         0.06874131  0.02694263 -0.11866201 -0.07401355 -0.14713107
IND_116      0.11058362  0.36211229  0.65340377  0.78555153  0.57207275
IND_094     -0.10084612  0.27158851  0.16697728  0.08113046  0.20090438
                IND_126 Complejidad       IND_815       IND_124      IND_002
IND_115     -0.26502092 -0.36174474  0.4776908604  0.2300645426 -0.040573032
IND_063     -0.42719654 -0.45392483  0.4659259365  0.0919928836 -0.063000909
IND_802     -0.29502068 -0.18351985  0.1418109383  0.0333258076  0.161372904
IND_117      0.13122797  0.04365947 -0.0755687245  0.1021840638 -0.229679809
IND_055     -0.26309599 -0.07977478  0.0433951800  0.1115009778  0.206663133
IND_120     -0.24750390 -0.40728789  0.1221512158  0.0409446114 -0.138129328
IND_057      0.60060230  0.45796470 -0.4670650615 -0.0425035131  0.031793383
IND_070     -0.46983344 -0.69933797  0.5914666684  0.0032874633 -0.200798096
IND_119      0.15764729  0.45999876 -0.5704772695 -0.0832011464  0.373622994
IND_087      0.19477853 -0.08433658 -0.0565406127  0.1267327578  0.004194826
IND_813     -0.28927000 -0.09561109  0.0571307960 -0.1077411318 -0.014366548
IND_125     -0.25359545 -0.42232356  0.3239012484  0.0305326189 -0.321780697
IND_076     -0.39778054 -0.58931398  0.5299073383  0.0536545629 -0.059963054
IND_061     -0.60028198 -0.53361549  0.5429981971 -0.0309313642  0.069990950
IND_062     -0.73622804 -0.53772841  0.5883770130  0.1067011951  0.033993139
IND_126      1.00000000  0.30133810 -0.5027470004 -0.0704170796  0.012404580
Complejidad  0.30133810  1.00000000 -0.6263134601 -0.0238510789  0.286952707
IND_815     -0.50274700 -0.62631346  1.0000000000  0.0006430932 -0.204513151
IND_124     -0.07041708 -0.02385108  0.0006430932  1.0000000000 -0.032093634
IND_002      0.01240458  0.28695271 -0.2045131512 -0.0320936343  1.000000000
MI_1         0.06917795  0.19363455 -0.0693590465  0.0475325278 -0.010249864
IND_116     -0.43773500 -0.48514444  0.3762464758  0.0650830321  0.122898200
IND_094     -0.13376418 -0.44572241  0.2245983089  0.0480580245 -0.296762988
                   MI_1     IND_116      IND_094
IND_115     -0.046858177  0.23348308  0.217418961
IND_063     -0.202614595  0.46324976  0.197269720
```

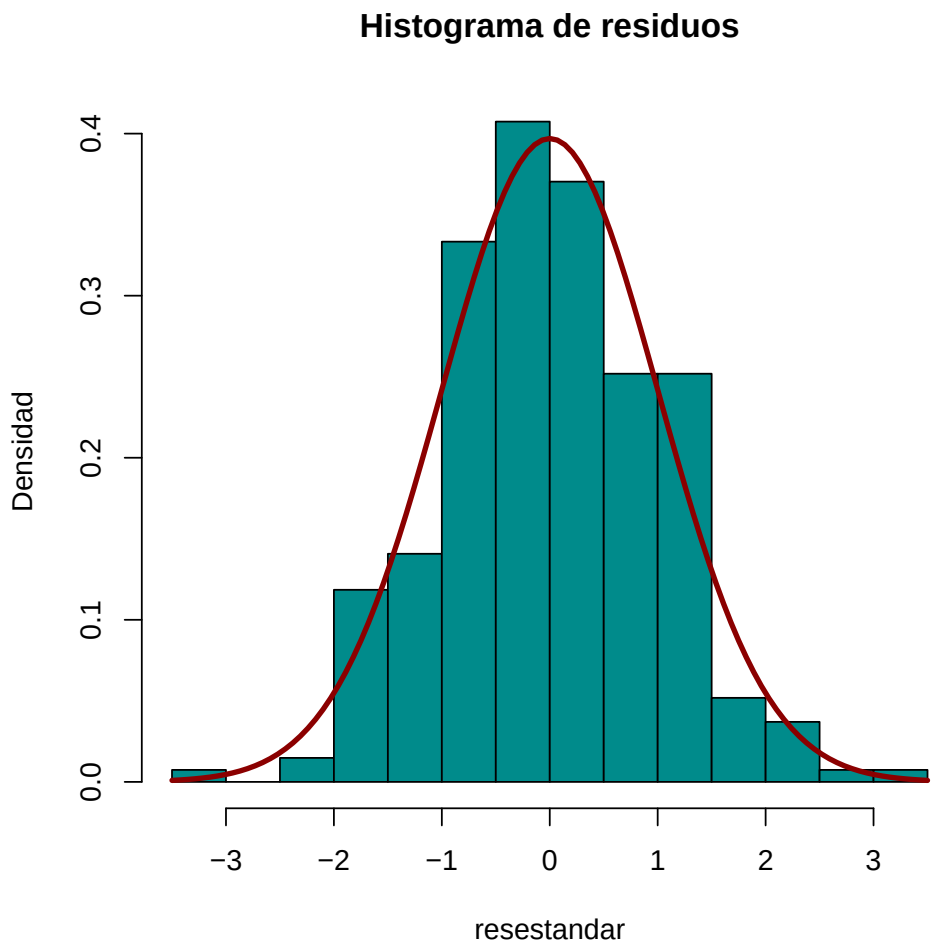
```
IND_802     -0.073573609  0.40871091  0.073955766
IND_117      0.009836568  0.01785217  0.130459828
IND_055      0.007785231  0.48603574  0.007660409
IND_120     -0.056319536  0.63622684  0.338117803
IND_057     -0.010439925 -0.79315528 -0.296245369
IND_070     -0.109235304  0.78314919  0.453753141
IND_119     -0.065510988 -0.09232495 -0.312586548
IND_087     -0.115813648  0.13555288  0.195163867
IND_813      0.068741310  0.11058362 -0.100846116
IND_125      0.026942632  0.36211229  0.271588509
IND_076     -0.118662008  0.65340377  0.166977276
IND_061     -0.074013551  0.78555153  0.081130464
IND_062     -0.147131072  0.57207275  0.200904382
IND_126      0.069177945 -0.43773500 -0.133764183
Complejidad  0.193634549 -0.48514444 -0.445722414
IND_815     -0.069359047  0.37624648  0.224598309
IND_124      0.047532528  0.06508303  0.048058025
IND_002     -0.010249864  0.12289820 -0.296762988
MI_1         1.000000000 -0.01797170 -0.055526688
IND_116     -0.017971699  1.00000000  0.156416704
IND_094     -0.055526688  0.15641670  1.000000000
```

```
    IND_115     IND_063     IND_802     IND_117     IND_055     IND_120
   2.184592    2.330014    7.321896    2.498231    8.647193    3.093053
    IND_057     IND_070     IND_076     IND_119     IND_125     IND_813
   7.363589   10.365763    3.784667    2.374110    1.586390    2.055705
    IND_087     IND_094     IND_116     IND_815     IND_061     IND_062
   1.852706    1.945425    6.939612    3.968731    8.074125    5.432403
    IND_126 Complejidad     IND_124     IND_002          MI
   3.292070    3.148247    1.192571    1.832953    1.273289
```

Hacemos lo de MI para que m?s adelante no nos de problemas en las transformaciones,ver con el summary si las variables son positivas y sacamos la correlacion.
Observamos nuevamente que las correlaciones mas altas estan en IND_070 , aunque nuevamente no esta tan alejado del 10.

Histograma para comprobar el supuesto de normalidad junto con la curva normal asociada.

```
        Anderson-Darling normality test

data:  resestandar
A = 0.20508, p-value = 0.8715
```



La grafica se adapta muy bien a la cura normal,exceptuando por un pico que sobresale por 1 en el eje x, ya nos da buenos indicios de normalidad.Hacemos la Grafica probabilidad normal (QQ plot), la cual nos muestra que verdaderamente se ajusta muy bien,con pequeños problemas en las colas.Por ultimo hacimos la Prueba Anderson-Darling.
Como el $p_valor = 0.8715 > 0.05$, entonces podemos aceptar la normalidad en nuestro modelo.

INDEPENDENCIA
Prueba Durbin-Watson de Autocorrelacion

```
        Durbin-Watson test

data:  EPN9
DW = 1.6555, p-value = 0.0004532
alternative hypothesis: true autocorrelation is greater than 0
```

Esta es de las pruebas mas dificiles, pues con nada pudimos hace que se cumpliera , aunque si nos pudimos acercar mucho mas a dos con la transformada.Para el modelo sin transformar tenemos un dw=1.655 cercano a dos , pero podria indicar correlacion positiva.
Ahora veamos cuales son las potencias que nos recomienda usar, donde nos dices la funcion powerTransform la mejor lambda y testTransform que lo mejor seria transformar nuestro modelo para mejorarlo.

Este es el modelo con todas lasvariables transformadas el cual vemos que es un desastre , rompiendo significancia de betas y seguro de pruebas de normalidad ,etc.

En cambio este modelo se consiguio siguiendo las sugerencias y se gano un mayor acercamento en la prueba dw, con $dw = 1.7546$, con todos los demas supuestos bien , y con variables significativas, en este caso ViF de $IND\_070 = 10.46$,cercano a 10,las variables transformadas son:
$IND_0 63, IND_8 02, IND_0 70, IND_1 19, IND_0 87, IND_8 13, IND_1 25\ IND_0 76, IND_0 62, COMPLEJIDAD, IND_8 15, IND_1 24$

Por ultimo veamos el Aic de estos tres modelos y veamos sus pros y contras.

```
              df      AIC
EPN9          25 5663.764
EPN9_trans    25 5655.318
EPN9_trans_1 25 5678.894
```

Ahora veamos los pros y contras del modelo transformado y sin trnsformar .En primera el modelo no tranformado casi todas sus variables son significativas a cualquier nivel con un $R2$ ajustada de 0.59,todas sus varibles tienen vif menor que $1a0$ excepciÓn de IND_070 con 10.3 , psa la prueba de A_D con un

$$p\_value = 0.8715$$

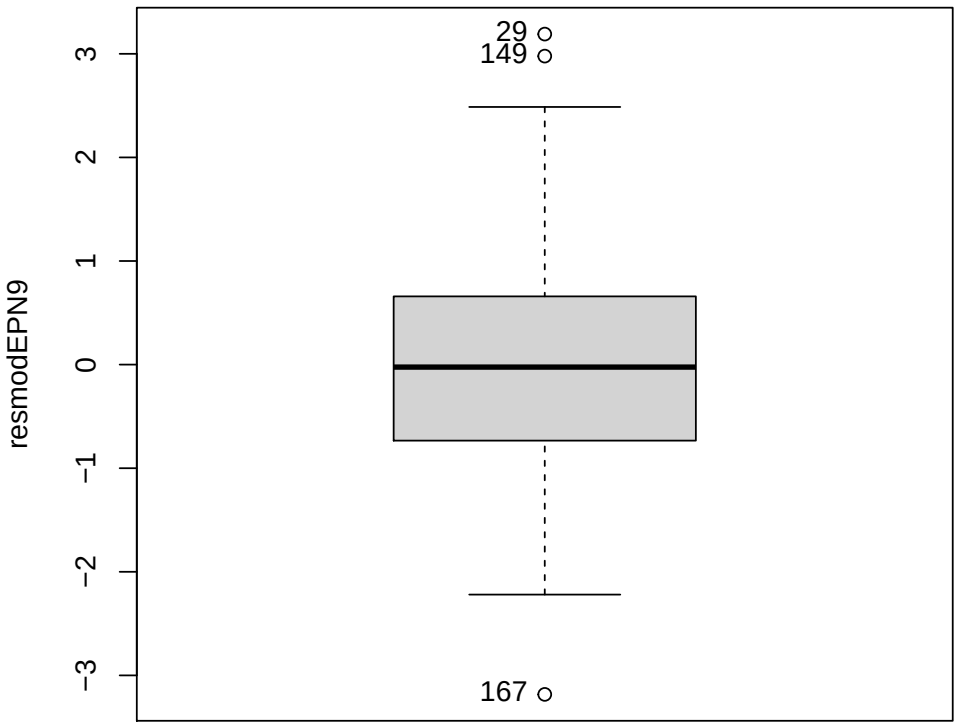
. Y la prueba de homosedasticidad con $p\_value = 0.3233$,mientras que en la prueba de $DW = 1.655$, lo cual nos dice que podria tener correlacion cero ya que es cercano a 2 , ademas sus variables son interpretables pues no han sufrido transformaciones.Por otra parte el modelo transformado sufre de una interpretcion menos fuerte de sus variables , la mayoria de sus variables son significativas a cualquier nivel , nuevamente los vifs son vajos a exepcion de IND_070 con 10.46,mejora en la prueba de $DW = 1.7546$, diciendonos que aqui lo mas seguro que la correlacion sea cero,la prueba de Anderson Darling la pasa con $p\_value = 0.8715$ y la de Homosedasticidad la pasa con un $p_value = 0.3672$.Si fuera por mi me quedaria con el modelo sin transformar, aunque tampoco esta de más analizar estos dos modelos tan interesantes.

Veamos también si hay valores atipicos o influyentes.

```
[1] 167  29 149

 25  27  29  35  38  39  44  46  94  95 134 159 166 190 201 203 205 273 276
 20  21  23  29  32  33  38  40  82  83 119 143 149 167 176 178 180 244 247

Potentially influential observations of
         lm(formula = EPN ~ IND_115 + IND_063 + IND_802 + IND_117 + IND_055 +      IND_120 + IND_057 + I

    dfb.1_ dfb.IND_115 dfb.IND_063 dfb.IND_80 dfb.IND_117 dfb.IND_055
25   -0.03    -0.01        0.01        0.00       0.00        0.00
27    0.00     0.00        0.00        0.00       0.00        0.00
29   -0.04     0.02        0.07       -0.13       0.03        0.07
35    0.25     0.10        0.51        0.12      -0.56       -0.34
38    0.00     0.00        0.01       -0.03       0.00        0.03
39    0.01     0.03       -0.01       -0.02       0.01        0.01
44    0.01     0.00        0.00        0.04      -0.01       -0.03
46   -0.25    -0.05        0.23       -0.03       0.06        0.07
94   -0.41    -0.20        0.73        0.18       0.16       -0.13
95    0.00     0.00        0.00        0.00       0.00        0.00
134   0.03    -0.01       -0.02        0.02      -0.02       -0.01
159   0.04    -0.04        0.01        0.28      -0.09       -0.26
166   0.02    -0.24        0.03       -0.36      -0.44        0.51
190   0.38     0.23       -0.15       -0.45      -0.41        0.27
201   0.02     0.01       -0.03        0.00       0.00       -0.01
203   0.01     0.00       -0.01        0.00       0.00        0.00
205  -0.07     0.00        0.07       -0.01       0.02        0.00
273   0.01     0.01        0.03        0.02      -0.01       -0.01
276   0.00     0.02       -0.01        0.00      -0.01        0.01
    dfb.IND_120 dfb.IND_057 dfb.IND_070 dfb.IND_076 dfb.IND_119 dfb.IND_125
25     0.00        0.01        0.00       -0.01        0.00        0.00
27     0.00        0.01        0.01        0.00        0.00        0.00
29     0.10       -0.11       -0.17       -0.03        0.09       -0.03
35    -0.22       -0.06        0.02       -0.06       -0.16        0.08
38    -0.01        0.01        0.01       -0.02        0.00        0.02
39     0.00        0.01        0.02        0.00       -0.01        0.03
44    -0.01        0.02        0.01        0.00       -0.05       -0.03
46     0.04        0.17        0.02       -0.08        0.03       -0.01
94     0.03        0.03       -0.24       -0.05       -0.01       -0.01
95     0.00        0.01        0.00        0.00        0.00        0.00
134   -0.01       -0.01        0.00        0.01        0.02       -0.01
159   -0.30       -0.12        0.01        0.17       -0.04        0.03
166   -0.11        0.11        0.27       -0.27       -0.15        0.11
190   -0.16       -0.29       -0.28        0.24       -0.12       -0.67
```

```
201  0.00       -0.01        0.00       -0.01        0.00        0.01
203  0.00        0.00        0.00        0.00       -0.01        0.00
205  0.00        0.04        0.00        0.00        0.05        0.01
273 -0.01       -0.02        0.00        0.01       -0.01        0.00
276 -0.01       -0.01       -0.01        0.01        0.01        0.00
    dfb.IND_813 dfb.IND_08 dfb.IND_09 dfb.IND_116 dfb.IND_815 dfb.IND_061
25  -0.01        0.00       0.00       0.01       -0.01        0.00
27  -0.01       -0.01       0.00       0.01       -0.01        0.00
29  -0.10       -0.24      -0.11       0.01        0.16        0.01
35  -0.37        0.08      -0.12       0.21       -0.62        0.16
38   0.04       -0.01      -0.01       0.00       -0.02        0.00
39   0.06       -0.03       0.00       0.00       -0.04        0.02
44   0.03        0.04      -0.01       0.01       -0.05       -0.01
46   0.05        0.08       0.01       0.00        0.02       -0.01
94  -0.25       -0.13       0.10       0.06       -0.06        0.01
95   0.00        0.00       0.00       0.00        0.00        0.00
134 -0.04       -0.05       0.00       0.01       -0.01       -0.01
159 -0.06        0.07      -0.11       0.16       -0.25       -0.05
166 -0.14        0.13       0.07      -0.19        0.23        0.11
190 -0.09        0.08       0.12       0.07       -0.03       -0.03
201  0.00       -0.02      -0.01       0.00       -0.01        0.01
203  0.00        0.00       0.00       0.01        0.00        0.00
205  0.00       -0.02       0.02       0.02        0.02        0.01
273  0.00        0.02       0.00       0.00        0.00        0.00
276 -0.01        0.02       0.00      -0.01       -0.02        0.00
    dfb.IND_062 dfb.IND_126 dfb.Cmpl dfb.IND_124 dfb.IND_00 dfb.MI dffit
25   0.04        0.05        0.02    -0.01       -0.02       0.00   0.08
27   0.01        0.02        0.01     0.00       -0.01       0.00   0.04
29   0.20        0.67       -0.22    -0.05       -0.04      -0.03   0.94_*
35  -0.66       -0.11       -0.20     0.02       -0.19      -0.03   1.51_*
38   0.00        0.00       -0.03     0.00        0.00       0.00   0.09
39   0.01        0.01        0.03    -0.02        0.00       0.14   0.23
44   0.00       -0.01       -0.04     0.02        0.01      -0.13  -0.20
46   0.09       -0.01       -0.04    -0.01        0.01      -0.18  -0.46
94   0.03       -0.34       -0.07    -0.04        0.00       0.24  -1.15_*
95   0.00        0.00        0.00     0.00        0.00       0.00   0.01
134 -0.03       -0.03       -0.03     0.01       -0.02      -0.01   0.09
159  0.03       -0.01        0.06    -0.06       -0.27       0.00   0.58
166 -0.10       -0.01        0.14     0.10       -0.11      -0.10   0.86
190 -0.05       -0.05       -0.03    -0.04       -0.29      -0.10  -1.28_*
201  0.02        0.02        0.01     0.01       -0.01      -0.07  -0.08
203 -0.01       -0.01        0.00    -0.01        0.00      -0.03  -0.03
205 -0.02       -0.03        0.02    -0.02       -0.01      -0.07  -0.15
273 -0.07       -0.01        0.00     0.00        0.00      -0.01   0.11
276  0.00        0.00       -0.01    -0.09       -0.01       0.01  -0.10
    cov.r   cook.d hat
25   1.34_*  0.00   0.18
27   1.42_*  0.00   0.22
29   0.70_*  0.04   0.13
35   0.49_*  0.09   0.18
38   1.29_*  0.00   0.15
39   1.30_*  0.00   0.17
44   1.30_*  0.00   0.17
46   1.32_*  0.01   0.22
94   0.86    0.05   0.21
95   1.33_*  0.00   0.17
134  1.39_*  0.00   0.21
159  0.63_*  0.01   0.05
166  0.49_*  0.03   0.08
190  0.46_*  0.07   0.13
201  1.36_*  0.00   0.19
203  1.41_*  0.00   0.22
205  1.41_*  0.00   0.22
```

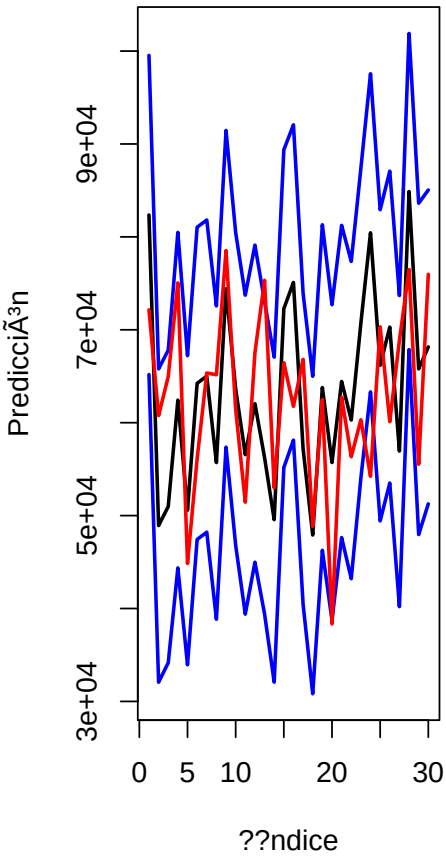
```
273  1.41_*  0.00   0.22
276  1.41_*  0.00   0.22
```



Podemos ver que aparentemente hay 4 puntos discrepantes $(33, 230, 231, 253)$, el Diagnostico de influencia esta representado por influence.measures y el summary nos muestra que observaciones se pueden considerar como influyentes,aunque al final ninguna la podemos considerar como tal.
Ahora veamos que tal nuestro modelo precide los datos de prediccion

```
         fit      lwr       upr
17  82356.51 65179.12  99533.89
4   48926.74 32065.55  65787.93
132 50985.29 34162.93  67807.64
26  62430.71 44372.74  80488.69
47  50578.95 33942.11  67215.79
181 64237.25 47431.84  81042.65
19  65015.34 48215.34  81815.34
174 55704.76 38843.81  72565.71
184 74416.28 57371.72  91460.84
59  63633.35 46806.36  80460.34
188 56560.05 39393.89  73726.20
16  62049.58 44986.36  79112.81
176 56242.93 39413.91  73071.94
178 49562.10 32071.57  67052.64
79  72269.52 55169.16  89369.88
78  75095.06 58125.47  92064.64
139 57200.38 40461.38  73939.37
281 47910.57 30825.13  64996.01
192 63786.00 46267.37  81304.63
218 55710.43 38725.87  72694.99
234 64437.40 47636.68  81238.12
223 60284.30 43209.58  77359.02
88  70708.58 54042.27  87374.89
```

```
163 80438.82 63318.40  97559.25
121 66174.77 49426.00  82923.54
197 70290.83 53507.08  87074.57
5   56943.24 40202.18  73684.29
100 84879.89 67864.39 101895.38
54  65790.62 47958.50  83622.75
210 68158.12 51266.23  85050.01
```



Vemos que los resultados son muy bueno , asi que podemos considerar muy bueno nuestro modelo sin transformar, hagmos lo mismo con el modelo transformado.

Josefina Vázquez Mota Primero realizaremos un modelo tomando como variable dependiente JVM y las demas como independientes Los datos de las variables provendrán de INEGI_INE_TRAIN Podemos ver que este modelo cubre una ADJ $R^2 = 74.52$ sin embargo podemos notar que el p-value de la mayoria de las variables es > 0.05 asi como en algunas tenemos datos NA Realizando vif, concluimos que este no nos es de utilidad en este momento, ya que no hemos corroborado que todas las variables sean Linealmente Independientes

Construimos un modelo vacio "JVM.vacio" y un modelo donde iremos integrando las variables explicativas "JVM.completo"

Utilizamos el algoritmo Stepwise para escoger un modelo con AIC. Revisaremos el vif del modelo por AIC y procederemos a retirar una variable cuyo resultado sea > 10, en este caso IND_810 que fue la mas alta con un valor de 1232.187721 .

El modelo se reduce utilizando AIC, esto lo realizamos con el algoritmo Stepwise Seguido de esto, analizamos el vif para cada modelo en busca de uno > 10, se escoge el valor mas grande y este es eliminado del modelo simpre y cuando la varianza predecida por el modelo no disminuya abruptamente Una vez ya no haya vif > 10 procederemos a correr el algoritmo step para ver si se puede reducir más nuestro modelo

Utilizamos el algoritmo Stepwise para escoger un modelo con AIC. Revisaremos el vif del modelo por AIC y procederemos a retirar una variable cuyo resultado sea > 10, en este caso $\mathrm{IND}_8 10 que fue la mas alta con un valor de 1232.187721.$

Tenemos nuestro primer candidato a modelo final Podemos ver que este modelo cubre una ADJ $R^2 = 70.04$ notando una leve disminución Tambien podemos recalcar que ya se presenta un buen p-value en las variables Nuestro modelo está constituido por las siguientes variables:
IND_806—- Estimador del total de población de 18 años y más
IND_115—- Porcentaje de la población de 12 años y más separada
IND_818—- Porcentaje de población de 0 a 9 años
IND_122—- Porcentaje de la población afiliada al ISSSTE
IND_094—- Porcentaje de la población de 15 a 24 años que asiste a la escuela en un municipio o delegación distinto al de residencia
IND_047—- Densidad de población (hab/km2)
IND_061—- Porcentaje de viviendas con disponibilidad de agua entubada en la vivienda
IND_082—- Porcentaje de población de 15 años y más con nivel de escolaridad media superior
IND_108—- Porcentaje de la PNEA en otras actividades no económicas
IND_062—- Porcentaje de viviendas con disponibilidad de drenaje en la vivienda
IND_141—- Porcentaje de la población que se considera indígena
IND_049—- Porcentaje de la población que tiene acta de nacimiento
Indigena—- Identificador de distrito indígena
IND_058—- Porcentaje de viviendas con piso de tierra
IND_092—- Porcentaje de la población de 6 a 11 años que asiste a la escuela en un municipio o delegación distinto al de residencia
IND_064—- Porcentaje de viviendas con disponibilidad de electricidad en la vivienda
IND_048—- Porcentaje de la población que no tiene nacionalidad mexicana
IND_072—- Porcentaje de viviendas con calentador solar
IND_116—- Porcentaje de la población de 12 años y más divorciada

Transformaciones Primero crearemos una base de datos utilizando las variables necesarias para nuestro modelo base que es "JVM6" estos datos proveendrán de INEGI$_I NE_T RAIN$

Veremos que las variables sean > 0, de no ser asi las retiraremos

```
INEGI_INE_TRAIN.JVM    IND_806          IND_115         IND_818
Min.   :  5849      Min.   :190518   Min.   :2.016   Min.   : 8.738
1st Qu.: 26984      1st Qu.:249906   1st Qu.:3.478   1st Qu.:16.463
Median : 40440      Median :265281   Median :4.438   Median :18.251
Mean   : 42272      Mean   :267362   Mean   :4.377   Mean   :17.991
3rd Qu.: 53770      3rd Qu.:284533   3rd Qu.:5.185   3rd Qu.:19.779
Max.   :114801      Max.   :352001   Max.   :7.447   Max.   :28.129
   IND_122          IND_094            IND_047             IND_061
Min.   : 1.053   Min.   : 0.03129   Min.   :    5.542   Min.   :15.35
1st Qu.: 4.661   1st Qu.: 2.91447   1st Qu.:   53.336   1st Qu.:56.63
Median : 6.700   Median : 6.71988   Median :  215.169   Median :78.16
Mean   : 7.755   Mean   : 9.65374   Mean   : 2453.452   Mean   :72.56
3rd Qu.: 9.648   3rd Qu.:15.16074   3rd Qu.: 2039.717   3rd Qu.:90.95
Max.   :22.473   Max.   :42.77394   Max.   :19736.661   Max.   :99.21
   IND_082          IND_108          IND_062         IND_141
Min.   : 8.858   Min.   : 5.008   Min.   :39.08   Min.   : 2.118
1st Qu.:17.588   1st Qu.: 8.526   1st Qu.:91.79   1st Qu.: 8.448
Median :21.978   Median :10.878   Median :96.86   Median :13.756
Mean   :21.498   Mean   :11.511   Mean   :92.57   Mean   :21.856
3rd Qu.:24.922   3rd Qu.:13.712   3rd Qu.:98.61   3rd Qu.:26.192
Max.   :34.062   Max.   :24.747   Max.   :99.80   Max.   :85.445
   IND_049         Indigena          IND_058            IND_092
Min.   :82.45   Min.   :0.00000   Min.   : 0.01756   Min.   : 0.0000
1st Qu.:97.83   1st Qu.:0.00000   1st Qu.: 0.92865   1st Qu.: 0.3897
Median :98.28   Median :0.00000   Median : 1.93368   Median : 1.3315
Mean   :97.90   Mean   :0.09259   Mean   : 3.81351   Mean   : 3.1948
3rd Qu.:98.59   3rd Qu.:0.00000   3rd Qu.: 4.99397   3rd Qu.: 4.9532
Max.   :99.38   Max.   :1.00000   Max.   :24.70547   Max.   :18.0394
   IND_064          IND_048            IND_072            IND_116
Min.   : 78.12   Min.   :0.007377   Min.   : 0.08967   Min.   :0.1883
1st Qu.: 98.47   1st Qu.:0.075133   1st Qu.: 0.55350   1st Qu.:0.9072
Median : 99.33   Median :0.210156   Median : 1.74095   Median :1.4431
Mean   : 98.62   Mean   :0.404859   Mean   : 3.17818   Mean   :1.5776
3rd Qu.: 99.72   3rd Qu.:0.404465   3rd Qu.: 4.00241   3rd Qu.:2.1404
Max.   :100.00   Max.   :4.394647   Max.   :36.10789   Max.   :4.6544
```

Observamos que tanto Indigena como $IND_092 tienenalmenosunvalor = 0porloquequitaremosestas2$

Utilizaremos la funcion powerTransform para obtener una lambda para transformar cada variable

```
Call:
lm(formula = JVM ~ IND_806 + IND_115 + IND_818 + IND_122 + IND_094 +
    IND_047_tr + IND_061 + IND_082 + IND_108 + IND_062_tr + IND_141_tr +
    IND_049_tr + IND_058 + IND_064 + IND_048_tr + IND_072 + IND_116_tr,
    data = INEGI_INE_TRAIN)

Residuals:
   Min     1Q Median     3Q    Max
-30233  -6742   -218   7395  36995

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept)   1.937e+05   4.717e+04   4.107 5.42e-05 ***
IND_806       1.857e-01   3.280e-02   5.663 4.04e-08 ***
IND_115      -3.637e+03   9.385e+02  -3.875 0.000136 ***
IND_818      -4.871e+03   4.645e+02 -10.487  < 2e-16 ***
IND_122      -1.731e+03   2.244e+02  -7.714 2.84e-13 ***
IND_094      -4.603e+02   1.058e+02  -4.351 1.97e-05 ***
IND_047_tr   -1.510e+03   5.739e+02  -2.631 0.009033 **
IND_061       3.981e+02   9.917e+01   4.014 7.88e-05 ***
IND_082      -1.108e+03   3.291e+02  -3.366 0.000881 ***
IND_108      -1.328e+03   2.836e+02  -4.680 4.68e-06 ***
IND_062_tr   -3.728e-20   1.300e-20  -2.869 0.004466 **
IND_141_tr   -4.635e+03   1.255e+03  -3.693 0.000272 ***
IND_049_tr   7.352e-140  2.430e-140   3.026 0.002739 **
IND_058       1.999e+02   2.866e+02   0.698 0.486127
IND_064      -2.698e+02   4.492e+02  -0.601 0.548646
IND_048_tr    1.391e+03   1.079e+03   1.289 0.198537
IND_072      -3.498e+02   2.276e+02  -1.536 0.125693
IND_116_tr   -1.483e+04   9.812e+03  -1.511 0.131955
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11170 on 252 degrees of freedom
Multiple R-squared:  0.6914,        Adjusted R-squared:  0.6706
F-statistic: 33.21 on 17 and 252 DF,  p-value: < 2.2e-16
```

Este modelo explica una R^2 de 67.07% Tambien podemos notar que IND_058, IND_064, IND_048_tr, IND_072, IND_116_tr presentan un p-value alto por lo que no rechazan la Hipótesis nula i.e. pareceque $beta\_13 = 0$, $beta\_14 = 0$, $beta\_15 = 0$, $beta\_16 = 0$, $beta\_17 = 0$.
Proseguimos a crear una base de datos con las variables de JVM1_trans para luego poder analizar si se correlacionan.

```
Call:
lm(formula = JVM ~ IND_806_tr + IND_115_tr + IND_818_tr + IND_122_tr +
    IND_094_tr + IND_047_tr + IND_061_tr + IND_082_tr + IND_108_tr +
    IND_062_tr + IND_141_tr + IND_049_tr + IND_058_tr + IND_064_tr +
    IND_048_tr + IND_072_tr + IND_116_tr, data = INEGI_INE_TRAIN)

Residuals:
   Min     1Q Median     3Q    Max
-29508  -7236   -318   7442  40781

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept)   1.881e+05   2.659e+04   7.076 1.47e-11 ***
IND_806_tr    1.992e-01   3.289e-02   6.057 5.02e-09 ***
IND_115_tr   -1.677e+04   4.164e+03  -4.027 7.47e-05 ***
IND_818_tr   -4.818e+03   4.929e+02  -9.774  < 2e-16 ***
IND_122_tr   -1.377e+04   1.863e+03  -7.391 2.14e-12 ***
IND_094_tr   -4.557e+03   1.388e+03  -3.283  0.00117 **
```

```
IND_047_tr   -6.922e+02   6.641e+02  -1.042  0.29829
IND_061_tr    3.270e+01   7.627e+00   4.287 2.58e-05 ***
IND_082_tr   -8.619e+02   3.365e+02  -2.561  0.01101 *
IND_108_tr   -9.644e+03   3.294e+03  -2.928  0.00372 **
IND_062_tr   -1.587e-20   1.406e-20  -1.129  0.26017
IND_141_tr   -5.035e+03   1.226e+03  -4.107 5.43e-05 ***
IND_049_tr   9.971e-140  2.247e-140   4.439 1.35e-05 ***
IND_058_tr    9.580e+03   7.017e+03   1.365  0.17337
IND_064_tr  -1.445e-175   0.000e+00    -Inf  < 2e-16 ***
IND_048_tr    1.663e+03   1.054e+03   1.577  0.11594
IND_072_tr   -1.375e+03   7.844e+02  -1.753  0.08089 .
IND_116_tr   -1.025e+04   9.789e+03  -1.047  0.29628
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11220 on 252 degrees of freedom
Multiple R-squared:  0.6883,        Adjusted R-squared:  0.6673
F-statistic: 32.74 on 17 and 252 DF,  p-value: < 2.2e-16
```

Realizamos un modelo con las variables transformadas y lo llamamos JVM_trans Podemos ver que tvalue de IND_064_tr tiende a Inf, por lo que la retiraremos ya que impide realizar vif Este modelo explica el 66.73%.

```
Call:
lm(formula = JVM ~ IND_806_tr + IND_115_tr + IND_818_tr + IND_122_tr +
    IND_094_tr + IND_047_tr + IND_061_tr + IND_082_tr + IND_108_tr +
    IND_062_tr + IND_141_tr + IND_049_tr + IND_058_tr + IND_048_tr +
    IND_072_tr + IND_116_tr, data = INEGI_INE_TRAIN)

Residuals:
   Min     1Q Median     3Q    Max
-29222  -6719    735   6872  39926

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept)   1.904e+05   2.688e+04   7.083 1.39e-11 ***
IND_806_tr    1.950e-01   3.322e-02   5.869 1.37e-08 ***
IND_115_tr   -1.535e+04   4.175e+03  -3.677 0.000289 ***
IND_818_tr   -4.954e+03   4.957e+02  -9.992  < 2e-16 ***
IND_122_tr   -1.443e+04   1.867e+03  -7.727 2.59e-13 ***
IND_094_tr   -4.301e+03   1.400e+03  -3.071 0.002366 **
IND_047_tr   -1.544e+03   5.852e+02  -2.639 0.008836 **
IND_061_tr    3.077e+01   7.678e+00   4.007 8.08e-05 ***
IND_082_tr   -8.357e+02   3.402e+02  -2.457 0.014691 *
IND_108_tr   -9.998e+03   3.329e+03  -3.004 0.002934 **
IND_062_tr   -3.361e-20   1.246e-20  -2.699 0.007431 **
IND_141_tr   -4.791e+03   1.236e+03  -3.875 0.000136 ***
IND_049_tr   9.095e-140  2.247e-140   4.048 6.87e-05 ***
IND_058_tr    1.220e+04   7.024e+03   1.737 0.083576 .
IND_048_tr    1.710e+03   1.066e+03   1.604 0.110039
IND_072_tr   -1.395e+03   7.933e+02  -1.759 0.079829 .
IND_116_tr   -1.114e+04   9.895e+03  -1.126 0.261176
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11350 on 253 degrees of freedom
Multiple R-squared:  0.6799,        Adjusted R-squared:  0.6597
F-statistic: 33.59 on 16 and 253 DF,  p-value: < 2.2e-16
```

Podemos ver que el modelo tiene una R^2 de 65.97 notando que disminuye, pero no considerablemente

Volvemos a correr el algoritmo Stepwise, ahora para JVM_trans, llegando a un nuevo modelo, posteriormente corremos vif para buscar valores > 10 notando que no hay ninguno.
LLamamos a este modelo JVM.trans.
Este modelo explica el 65.93 Notando una leve disminucion Tambien podemos notar que IND_058_tr, IND_048_tr, $IND\_072\_trpresentanunpvaluealtoporloquenorechazanlaHiptesisnulai.e.parecequebeta\_13 = 0$,beta_14 = 0,beta_15

$= 0$.
$Proseguimosacrearunabasededatosconlasvariables de JVM1\_transparaluegopoderanalizarsisecorrelacionan$

```
Call:
lm(formula = log(JVM + 4.1186) ~ IND_806 + IND_115 + IND_818 +
    IND_122 + IND_094 + IND_047 + IND_061 + IND_082 + IND_108 +
    IND_062 + IND_141 + IND_049 + Indigena + IND_058 + IND_092 +
    IND_064 + IND_048 + IND_072 + IND_116, data = INEGI_INE_TRAIN)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24379 -0.12726  0.01908  0.18213  0.76278

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.969e+00  1.509e+00   5.944 9.32e-09 ***
IND_806      4.354e-06  8.957e-07   4.861 2.07e-06 ***
IND_115     -8.847e-02  2.595e-02  -3.409 0.000761 ***
IND_818     -1.150e-01  1.208e-02  -9.519  < 2e-16 ***
IND_122     -3.299e-02  6.020e-03  -5.481 1.03e-07 ***
IND_094     -1.793e-02  5.285e-03  -3.393 0.000804 ***
IND_047     -2.770e-05  5.834e-06  -4.749 3.45e-06 ***
IND_061      8.619e-03  2.452e-03   3.516 0.000520 ***
IND_082     -3.051e-02  8.203e-03  -3.720 0.000246 ***
IND_108     -2.548e-02  7.086e-03  -3.596 0.000389 ***
IND_062     -1.242e-02  4.722e-03  -2.630 0.009071 **
IND_141     -6.902e-03  1.872e-03  -3.687 0.000279 ***
IND_049      1.362e-01  2.565e-02   5.310 2.42e-07 ***
Indigena     3.111e-01  1.132e-01   2.748 0.006434 **
IND_058     -3.122e-02  8.342e-03  -3.742 0.000226 ***
IND_092      2.365e-02  1.047e-02   2.259 0.024749 *
IND_064     -8.330e-02  2.024e-02  -4.115 5.26e-05 ***
IND_048      1.991e-01  5.066e-02   3.929 0.000110 ***
IND_072     -1.031e-02  6.000e-03  -1.718 0.087054 .
IND_116     -1.092e-01  5.895e-02  -1.852 0.065143 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2955 on 250 degrees of freedom
Multiple R-squared:  0.6864,        Adjusted R-squared:  0.6625
F-statistic: 28.79 on 19 and 250 DF,  p-value: < 2.2e-16
```

Corremos el algoritmo logtrans para encontrar un $\alpha para realizar la transformacion logaritmica$.
$Este modelo explica el 66.25 notando un leve aumento de R^2$ Podemos notar que IND_072 y IND_116 presenta un $pvalue$ alto por lo que no rechaza la Hipótesis nula i.e. parece que $beta\_18 = 0$, $beta\_19 = 0$.

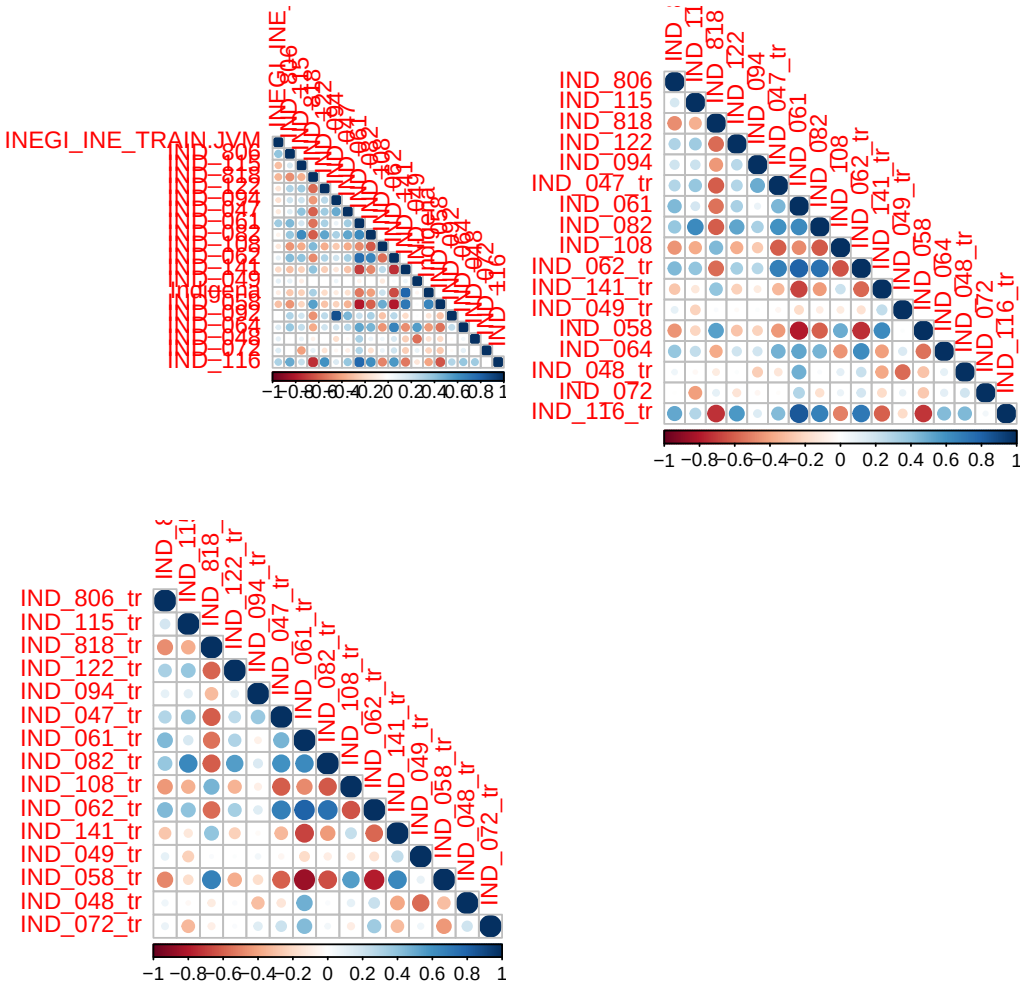
```
Call:
lm(formula = (JVM)^0.5580808 ~ IND_806 + IND_115 + IND_818 +
    IND_122 + IND_094 + IND_047 + IND_061 + IND_082 + IND_108 +
    IND_062 + IND_141 + IND_049 + Indigena + IND_058 + IND_092 +
    IND_064 + IND_048 + IND_072 + IND_116, data = INEGI_INE_TRAIN)

Residuals:
     Min       1Q   Median       3Q      Max
-195.168  -30.026    2.242   36.731  143.597

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.309e+02  2.752e+02   0.475 0.634859
IND_806      8.538e-04  1.634e-04   5.227 3.64e-07 ***
IND_115     -1.671e+01  4.733e+00  -3.531 0.000492 ***
IND_818     -2.429e+01  2.203e+00 -11.025  < 2e-16 ***
IND_122     -7.037e+00  1.098e+00  -6.410 7.19e-10 ***
IND_094     -3.970e+00  9.638e-01  -4.120 5.16e-05 ***
```

```
IND_047     -5.363e-03  1.064e-03  -5.040 8.92e-07 ***
IND_061      1.704e+00  4.471e-01   3.811 0.000174 ***
IND_082     -6.445e+00  1.496e+00  -4.308 2.37e-05 ***
IND_108     -5.265e+00  1.292e+00  -4.074 6.20e-05 ***
IND_062     -2.394e+00  8.611e-01  -2.780 0.005851 **
IND_141     -1.482e+00  3.414e-01  -4.342 2.06e-05 ***
IND_049      2.468e+01  4.678e+00   5.276 2.87e-07 ***
Indigena     6.682e+01  2.065e+01   3.237 0.001373 **
IND_058     -5.002e+00  1.521e+00  -3.288 0.001155 **
IND_092      4.671e+00  1.909e+00   2.447 0.015111 *
IND_064     -1.453e+01  3.692e+00  -3.936 0.000107 ***
IND_048      3.162e+01  9.239e+00   3.423 0.000724 ***
IND_072     -2.315e+00  1.094e+00  -2.116 0.035319 *
IND_116     -2.059e+01  1.075e+01  -1.915 0.056647 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.88 on 250 degrees of freedom
Multiple R-squared:  0.7221,        Adjusted R-squared:  0.7009
F-statistic: 34.18 on 19 and 250 DF,  p-value: < 2.2e-16
```

Realizamos el algoritmo de boxcox para encontrar una lambda para poder realizar una transformacion del tio Y^l
Este modelo explica el 70.09 notando un leve aumento de R^2
Notemos que el intercepto tiene un p-value alto por lo que no rechaza la Hipótesis nula i.e. parece ser que $beta\_0 = 0$
Tambien podemos notar que IND_116 presenta un p-value alto por lo que no rechaza la Hipótesis nula i.e. parece que $beta\_19 = 0$.

Comprobaremos los supuestos para los 5 modelos MODELOS: JVM6 JVM6.inv JMS.invi JVM1_trans JVM.trans

MULTICOLINEALIDAD Ya que JVM6, JVM6.inv, JVM6.invi presentan variables sin transformar, basta con analizar la correlacion entre las variables de BD_JVM0 Para JVM1_trans y JVM.trans analizaremos por sep-



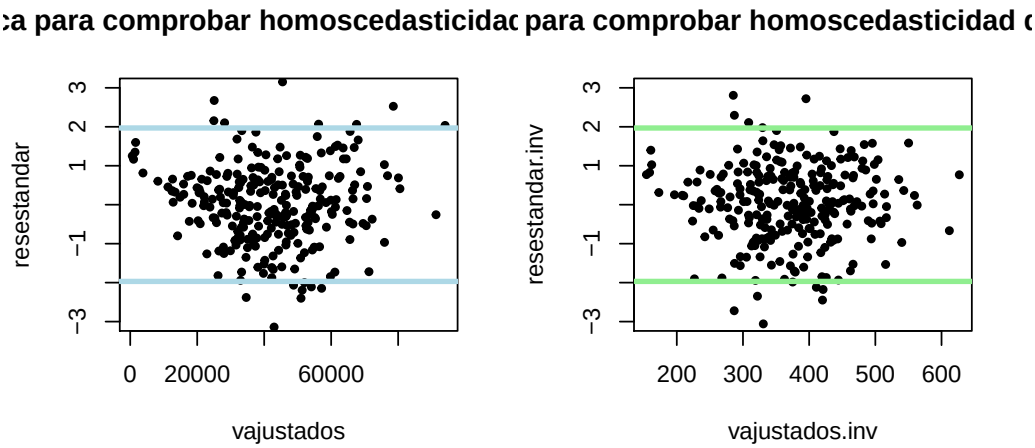
arado su correlacion

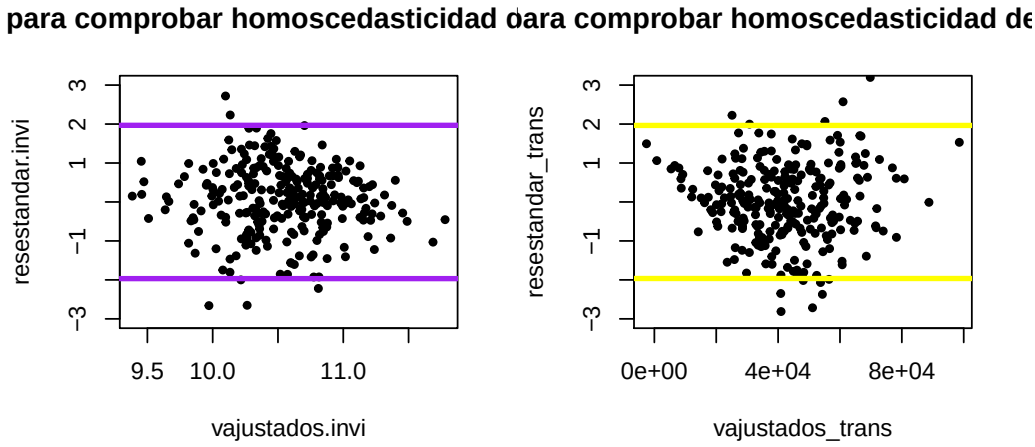
Observemos que hay correlaciones muy altas entre las variables para los 3 casos por lo que analizaremos sus vif

```
 IND_806   IND_115   IND_818   IND_122   IND_094   IND_047   IND_061   IND_082
1.717252 2.712193 3.816762 2.235057 6.156686 2.181570 8.349869 4.597724
 IND_108   IND_062   IND_141   IND_049 Indigena   IND_058   IND_092   IND_064
2.277978 6.412170 4.318623 5.103766 3.330402 4.589407 5.280539 6.179048
 IND_048   IND_072   IND_116
3.084242 1.928008 7.608271

 IND_806   IND_115   IND_818   IND_122   IND_094   IND_047   IND_061   IND_082
1.717252 2.712193 3.816762 2.235057 6.156686 2.181570 8.349869 4.597724
 IND_108   IND_062   IND_141   IND_049 Indigena   IND_058   IND_092   IND_064
2.277978 6.412170 4.318623 5.103766 3.330402 4.589407 5.280539 6.179048
 IND_048   IND_072   IND_116
3.084242 1.928008 7.608271

 IND_806   IND_115   IND_818   IND_122   IND_094   IND_047   IND_061   IND_082
1.717252 2.712193 3.816762 2.235057 6.156686 2.181570 8.349869 4.597724
 IND_108   IND_062   IND_141   IND_049 Indigena   IND_058   IND_092   IND_064
2.277978 6.412170 4.318623 5.103766 3.330402 4.589407 5.280539 6.179048
 IND_048   IND_072   IND_116
3.084242 1.928008 7.608271

   IND_806    IND_115    IND_818    IND_122    IND_094 IND_047_tr    IND_061
  1.611407   2.482157   3.949267   2.174333   1.726336   3.541353   9.562470
   IND_082    IND_108 IND_062_tr IND_141_tr IND_049_tr    IND_058    IND_064
  5.180981   2.554695   6.396474   2.214520   2.446209   3.791600   2.129014
IND_048_tr    IND_072 IND_116_tr
  3.505365   1.942778   9.828141

IND_806_tr IND_115_tr IND_818_tr IND_122_tr IND_094_tr IND_047_tr IND_061_tr
  1.576353   2.798889   3.436722   2.024904   1.730788   3.378828   8.443206
```

```
IND_082_tr IND_108_tr IND_062_tr IND_141_tr IND_049_tr IND_058_tr IND_048_tr
  5.291477   2.415459   5.687885   2.073879   2.003237   6.518931   3.175759
IND_072_tr
  2.063143
```

Analizando los vif para cada uno de los 5 modelos, vemos que no hay ninguno superior a 10 Analizando esto podemos intuir que no hay multicolinealidad





HOMOSEDASTICIDAD
Podemos ver que los 5 modelos tienen varianza mas o menos constante segun sus gráficas Realizamos una prueba Breusch-Pagan para comprobar la Homocedasticidad, se busca que obtengamos un $pvalue > 0.05$
Prueba Homosedasticidad

```
        studentized Breusch-Pagan test

data:  JVM6
BP = 38.532, df = 19, p-value = 0.005075

        studentized Breusch-Pagan test

data:  JVM6.inv
BP = 40.274, df = 19, p-value = 0.003011

        studentized Breusch-Pagan test

data:  JVM6.invi
BP = 33.326, df = 19, p-value = 0.02204

        studentized Breusch-Pagan test

data:  JVM1_trans
BP = 25.373, df = 17, p-value = 0.08667
```

```
        studentized Breusch-Pagan test

data:  JVM.trans
BP = 23.542, df = 15, p-value = 0.07329
```

Vemos que solo $JVM1\_trans$ y JVM.trans cumple con Homocedasticidad, aun cuando JVM6.invi en la grafica parece cumplirlo de igual manera
NORMALIDAD Realizaremos una prueba Anderson-Darling para corroborar normalidad Buscamos que la prueba sobre los residuos nos arroje un $p - value > 0.05$

```
        Anderson-Darling normality test

data:  resestandar
A = 0.33109, p-value = 0.5114

        Anderson-Darling normality test

data:  resestandar.inv
A = 0.93444, p-value = 0.0176

        Anderson-Darling normality test

data:  resestandar.invi
A = 2.5272, p-value = 2.138e-06

        Anderson-Darling normality test

data:  resestandar_trans
A = 0.30351, p-value = 0.5704

        Anderson-Darling normality test

data:  resestandar.trans
A = 0.25308, p-value = 0.7322
```

Estos valores, junto con las graficas QQ plot, nos indican que solo los residuos de JVM6, JVM_trans y JVM.trans se distribuyen Normal
INDEPENDENCIA Prueba Durbin-Watson de Autocorrelación

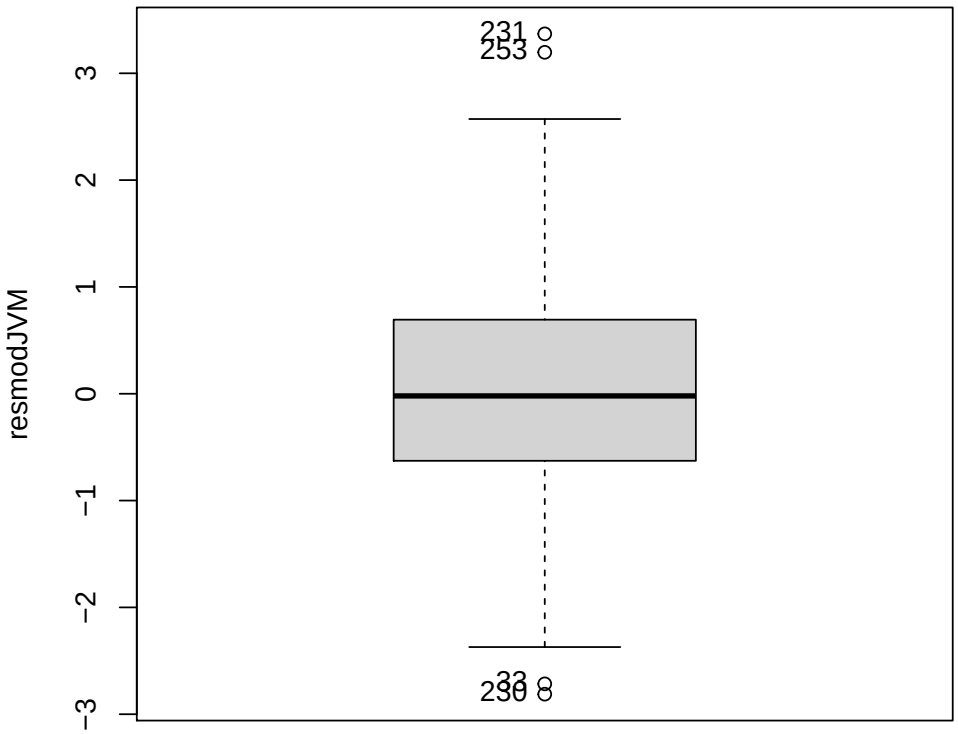
```
        Durbin-Watson test

data:  JVM6
DW = 1.5413, p-value = 1.964e-05
alternative hypothesis: true autocorrelation is greater than 0

        Durbin-Watson test

data:  JVM6.inv
DW = 1.4751, p-value = 1.572e-06
alternative hypothesis: true autocorrelation is greater than 0

        Durbin-Watson test

data:  JVM6.invi
DW = 1.3892, p-value = 3.852e-08
alternative hypothesis: true autocorrelation is greater than 0

        Durbin-Watson test

data:  JVM1_trans
DW = 1.5123, p-value = 6.539e-06
alternative hypothesis: true autocorrelation is greater than 0

        Durbin-Watson test

data:  JVM.trans
DW = 1.4606, p-value = 8.993e-07
alternative hypothesis: true autocorrelation is greater than 0
```

Estos DW's son relativamente cercanos a 2 en especial para JVM1_trans, por lo que podemos decir que no hay autocorrelacion residual por lo que entonces son independientes
Ya que solo JVM1_trans y JVM.trans cumplio con todos los supuestos, nos quedaremos con JVM1_trans ya que es el que tiene una R^2 mayor y pasa todos los supuestos

Puntos discrepantes e influyentes

```
[1]  33 230 231 253

  6  27  29  39  46  61  92  95  96 103 108 193 205 259 260 267 283
  4  21  23  33  40  52  80  83  84  90  95 169 180 230 231 238 253

Potentially influential observations of
         lm(formula = JVM ~ IND_806 + IND_115 + IND_818 + IND_122 + IND_094 +      IND_047_tr + IND_061

    dfb.1_ dfb.IND_80 dfb.IND_115 dfb.IND_81 dfb.IND_12 dfb.IND_09 dfb.IND_047
6    0.01   0.00      -0.02        0.00       0.01       0.00       0.00
27   0.11  -0.05      -0.07        0.07       0.01       0.04       0.00
29  -0.06  -0.01       0.05        0.07       0.02       0.01       0.05
39  -0.30   0.09      -0.10       -0.09       0.17      -0.11       0.17
46   0.01   0.00       0.00        0.00       0.00       0.00       0.00
61  -0.01   0.12       0.01       -0.02       0.06       0.22       0.02
92   0.17  -0.19      -0.15        0.02       0.22       0.02       0.05
95   0.02   0.00      -0.01       -0.03       0.00      -0.01      -0.01
96  -0.04  -0.01      -0.03        0.01       0.04      -0.02      -0.03
103  0.06  -0.17      -0.01        0.00       0.27      -0.12       0.00
108  0.00   0.10       0.11       -0.25      -0.03      -0.12      -0.16
193  0.00  -0.02       0.00       -0.02       0.00      -0.01       0.01
205 -0.98  -0.23       0.06        0.16      -0.23       0.14      -0.02
259 -0.04  -0.34      -0.30        0.10      -0.26       0.17       0.19
260 -0.10  -0.03      -0.06        0.20      -0.04      -0.07       0.15
267  0.04  -0.10      -0.01       -0.15      -0.07      -0.03      -0.01
283  0.22   0.14       0.31       -0.44      -0.06       0.01      -0.30
    dfb.IND_061 dfb.IND_08 dfb.IND_10 dfb.IND_062 dfb.IND_14 dfb.IND_049
6    0.00        0.00       0.00       0.01        0.00       0.01
27   0.07        0.05      -0.18       0.02       -0.03      -0.03
29   0.03       -0.06      -0.01      -0.05        0.03      -0.01
39  -0.17        0.12       0.07       0.11        0.00       0.18
46   0.00        0.00       0.00       0.00        0.00       0.00
61   0.02       -0.19      -0.04       0.02        0.04      -0.03
92   0.13       -0.18      -0.15       0.11        0.16      -0.08
95  -0.03        0.03      -0.02      -0.01        0.00      -0.01
96   0.01        0.01       0.00       0.01       -0.01      -0.05
103  0.23       -0.22       0.08       0.22        0.01      -0.06
108 -0.02        0.07      -0.10      -0.05       -0.05      -0.18
193 -0.01        0.00      -0.01      -0.01        0.00       0.00
205 -0.02        0.09       0.07      -0.19       -0.07       0.17
259 -0.24        0.23       0.16      -0.13       -0.18       0.14
260 -0.03        0.06       0.08      -0.03       -0.03      -0.12
267  0.14       -0.15       0.17       0.15        0.10       0.07
283  0.15       -0.33      -0.19       0.25       -0.04       0.11
    dfb.IND_05 dfb.IND_064 dfb.IND_048 dfb.IND_07 dfb.IND_116 dffit   cov.r
6    0.00      -0.01        0.00       -0.01      -0.01       -0.03    1.22_*
27   0.20      -0.11        0.05       -0.01      -0.05        0.41    1.24_*
29   0.02       0.03       -0.01        0.00       0.04        0.13    1.23_*
39   0.23       0.31        0.07        0.13      -0.08       -0.79    0.68_*
46   0.00      -0.02        0.00        0.00       0.00        0.02    1.97_*
61   0.05      -0.01        0.05       -0.04       0.06        0.43    1.22_*
92  -0.32      -0.09        0.00       -0.16      -0.18       -0.59    0.77_*
95  -0.08       0.00        0.03        0.01      -0.03       -0.11    1.25_*
96   0.07       0.04       -0.06        0.01       0.02        0.14    1.23_*
103  0.07      -0.02       -0.08       -0.30      -0.14       -0.56    0.75_*
```

```
108  0.12       0.06       -0.09       0.75       -0.11       0.86_*  1.39_*
193  0.00       0.00        0.00       0.01        0.01      -0.04    1.23_*
205  0.42       0.97        0.05       0.09        0.30      -1.23_*  1.42_*
259  0.09       0.02        0.01       0.08        0.37      -0.80_*  0.65_*
260 -0.08       0.08        0.18      -0.13        0.05       0.64    0.48_*
267  0.03      -0.04       -0.16      -0.21        0.06       0.51    0.69_*
283  0.11      -0.10        0.08      -0.10       -0.25       0.79    0.54_*
    cook.d hat
6     0.00   0.12
27    0.01   0.18
29    0.00   0.13
39    0.03   0.08
46    0.00   0.46_*
61    0.01   0.17
92    0.02   0.06
95    0.00   0.14
96    0.00   0.13
103   0.02   0.05
108   0.04   0.31_*
193   0.00   0.13
205   0.08   0.37_*
259   0.03   0.07
260   0.02   0.03
267   0.01   0.04
283   0.03   0.06
```



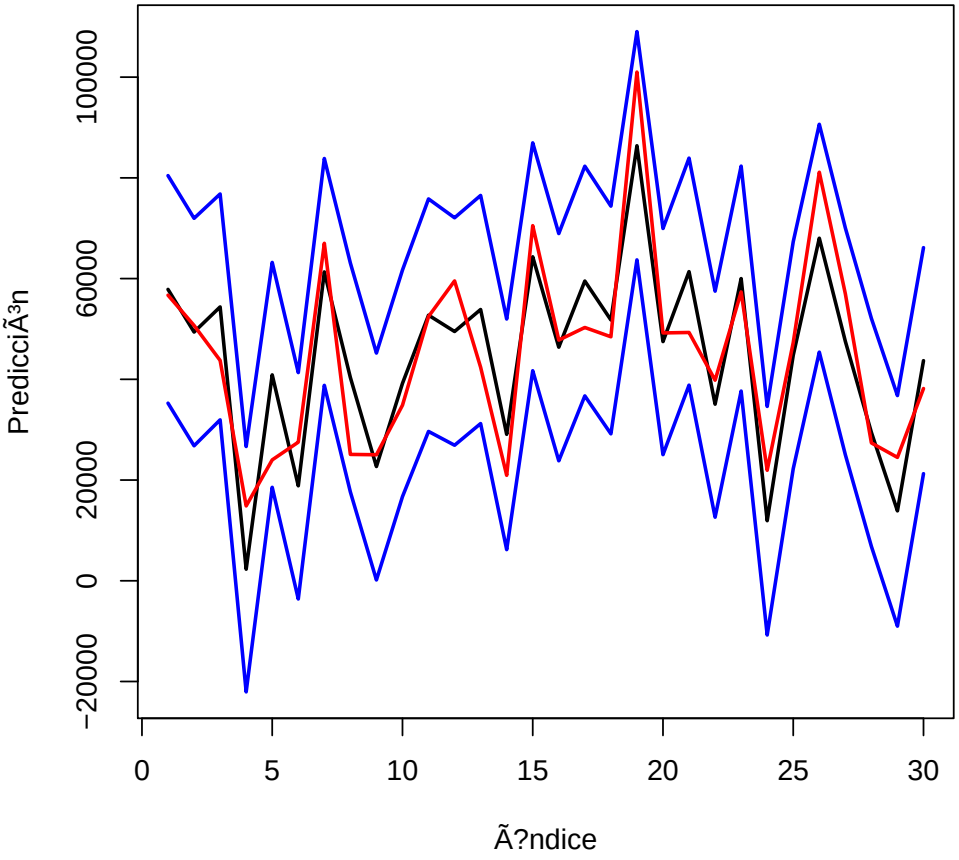
Podemos ver que aparentemente hay 4 puntos discrepantes $(33, 230, 231, 253)$ Podemos ver que ninguna distancia es > 0.5 por lo que no tenemos dato atipicos e infliyentes.

Ahora haremos el modelo bajo las variables predictorias

Ahora valuamos el modelo con INEGI_INE_PRED

```
      fit             lwr              upr       
 Min.   : 2297   Min.   :-22063   Min.   : 26657  
 1st Qu.:36095   1st Qu.: 13641   1st Qu.: 58549  
 Median :47512   Median : 25028   Median : 69996  
 Mean   :44800   Mean   : 22147   Mean   : 67453  
 3rd Qu.:56986   3rd Qu.: 34434   3rd Qu.: 79538  
 Max.   :86379   Max.   : 63717   Max.   :109041  
```

	fit	lwr	upr	
17	57855.321	35261.3088	80449.33	56707
4	49382.848	26801.4026	71964.29	50691
132	54377.801	31952.2291	76803.37	43756
26	2296.832	-22062.9229	26656.59	14862
47	40884.403	18557.0625	63211.74	24005
181	18856.025	-3620.5643	41332.61	27519
19	61324.373	38803.6863	83845.06	67006
174	40357.039	17630.5252	63083.55	25090
184	22682.989	158.0556	45207.92	25019
59	39202.425	16708.3890	61696.46	34856
188	52735.007	29651.0679	75818.95	52468
16	49492.668	26908.6045	72076.73	59536
176	53860.607	31213.5509	76507.66	42394
178	29073.360	6178.8458	51967.87	20936
79	64328.990	41705.5963	86952.38	70508
78	46375.988	23817.7978	68934.18	47785
139	59512.472	36711.5770	82313.37	50293
281	51779.997	29176.9362	74383.06	48444
192	86379.051	63717.0729	109041.03	101000
218	47490.312	25038.7445	69941.88	49208
234	61381.561	38840.1958	83922.93	49300
223	35059.247	12618.2498	57500.24	39820
88	60003.388	37671.8310	82334.94	57345
163	11922.599	-10771.6939	34616.89	21935
121	44783.733	22247.5828	67319.88	47294
197	68018.889	45398.4460	90639.33	81117
5	47534.229	25017.9170	70050.54	56986
100	29457.430	6822.1389	52092.72	27369
54	13878.311	-9013.5821	36770.20	24497
210	43707.875	21266.8272	66148.92	38169



Como podemos observar, en la grafica el modelo que se ajusto es bastante bueno para predecir nuestros datos de prediccion, por lo que podemos concluir que es un buen modelo

INTERPRETACIÓN DEL MODELO Dadas las variables que quitamos, el modelo para JVM se ve explicad por un total de 17 variables, de las cuales varias son de gente que cuenta con servicios básicos en su vivienda, otra variable que pudimos notar es el de Porcentaje de la poblacion de 12 aÃ±*osymsdivorciadayPorcentajedelapoblacionde*12*aosymss* *aquerealizo*, *yaqueestasapoyabanamadressolterasytrabajadoras*.

$Finalmente, construiremoselmejormodeloquedescribaalosdatosdecandidadoAndrsManuelLpezObrador(AMLO).Paraell$
de 0.799, lo que nos describe un buen modelo, sin embargo tambien podemos observar que muchas variables no son significativas, por lo que haremos un estudio más exhaustivo para ver si es necesario tener todas las variables, o podemos tener un modelo más simple con el cual se expliquen las observaciones.

El modelo se reduce, usando el AIC, es decir se busca que en cada modelo el AIC disminuya, esto lo hacemos con la función step de R, y después se analisan los Vif, y al ser mayores a 10 se eliminan, esto simpre y cuando la varianza predecida por eel modelo no disminuya abruptamente; finalmente se vuelve a utilizar la función step y con ella obtenemos el mejor modelo, que esta constituido por las variables:

IND_074——-Porcentaje de viviendas con separación de residuos
IND_115——-Porcentaje de la población de 12 años y más separada
IND_087——-Porcentaje de asistencia escolar de la población de 3 a 5 años
IND_047——-Densidad de población (hab/km2)
IND_815——-Porcentaje de la población ocupada que labora en el sector económico de comercio
IND_119——-Porcentaje de la población afiliada a servicios de salud
IND_082——-Porcentaje de población de 15 años y más con nivel de escolaridad media superior
IND_121——-Porcentaje de la población afiliada al IMSS
IND_804——-Estimador del total de población de 15 años y más (Mujeres)
IND_056——-Promedio de ocupantes por vivienda
IND_095——-Porcentaje de la población de 12 años y más económicamente activa (Total)
Complejidad—Grupo de complejidad electoral
IND_002——-Porcentaje estatal de la población
IND_088——-Porcentaje de asistencia escolar de la población de 6 a 11 años
IND_064——-Porcentaje de viviendas con disponibilidad de electricidad en la vivienda

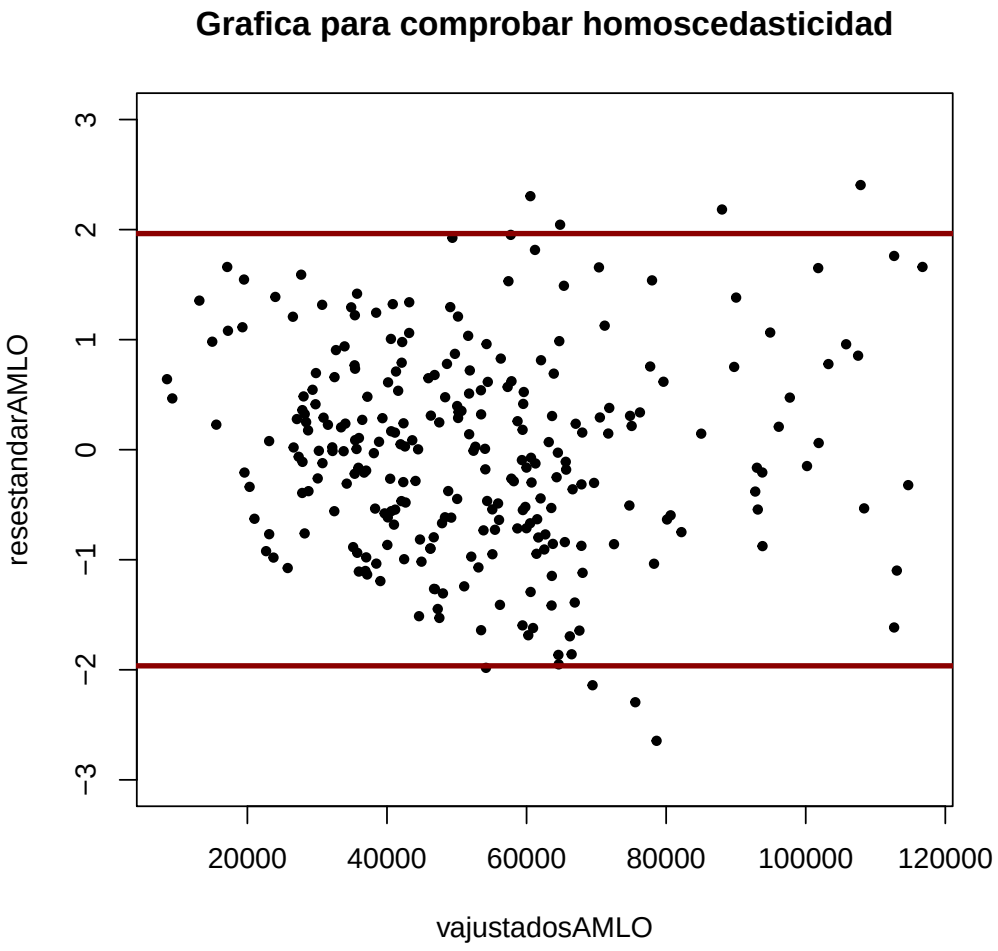
Dicho modelotiene una $R^2 = 0.7766$, que es menor a la del modelo completo pero por muy poco, lo que nos hace llegar a la conclusión de que sí es un mejor modelo, pues esta quitando muchas variables, pero aún así describe casi la misma cantidad de varianza. Ahora bien, veamos si este modelo cumple con los supuestos de la regresión lineal múltiple

Multicolinealidad Para ver si se cumple este supuesto debemos ver si no existe correlación entre ellos, más aún si el vif (variance inflation factor) es menor a 10

```
   IND_074     IND_115     IND_087     IND_047     IND_815     IND_119
1.937475    2.437532    1.527408    2.973433    2.927926    2.245809
   IND_082     IND_121     IND_804     IND_056     IND_095 Complejidad
4.353239    4.546753    1.462475    1.850290    4.767715    3.766790
   IND_002     IND_088     IND_064
1.463409    3.209927    4.347050
```

Homosedasticidad

Para ver si este supuesto se cumple, primero veremos que sucede al graficar los residuos estandarizados, contra los valores ajustado. Como podemos observar en la gráfica siguiente, puede decirse que si es una gráfica nula, por lo que tendriamos varianza constante, es decir, la homocedasticidad si se cumple.



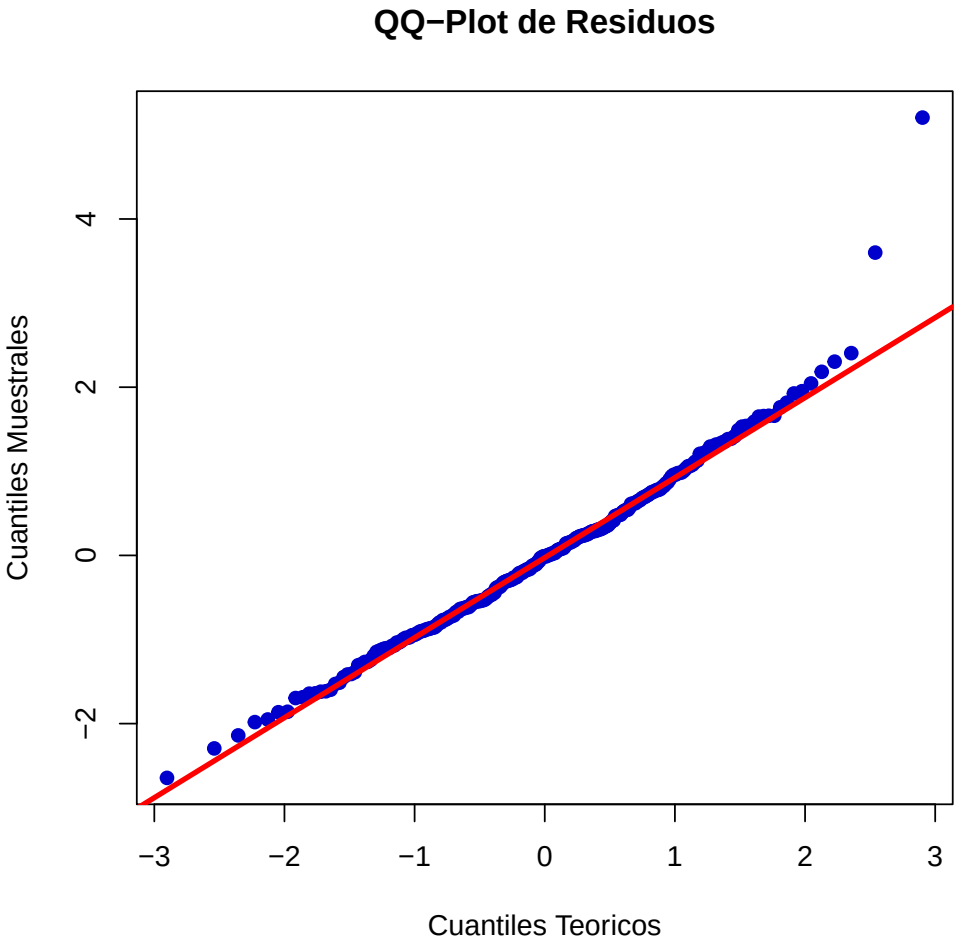
Hagamos la prueba formal para ver si la homesedasticidad se cumple:

```
        studentized Breusch-Pagan test

data:  AMLO8
BP = 33.017, df = 15, p-value = 0.004668
```

Como podemos observar, el pvalor es mucho más pequeño que el nivel de significancia 0.05, por lo que se puede afirmar que no se cumple el supuesto de varianza constante, más adelante se tratará de arreglar dicho supuesto con una transformación.

Normalidad

Al igual que para la prueba anterior, primero se hará una prueba gráfica



Este gráfico nos indica que los datos se ajustan bien en el centro de la distritribución y en la cola izquierda, sin embargo la cola derecha se ve que podria ocasionarnos problemas
Ahora bien, al hacer la prueba formal obtenemos:

```
        Anderson-Darling normality test

data:  resestandarAMLO
A = 0.58455, p-value = 0.1268
```

Es decir, tenemos un pvalor más grande que el nivel de significancia, por lo que aparentemente los residuos sí se distribuyen normales.
Independencia
Para ver este supuesto se hará la prueba.

```
        Durbin-Watson test

data:  AMLO8
DW = 1.3131, p-value = 1.052e-09
alternative hypothesis: true autocorrelation is greater than 0
```

Donde obtenemos un Dw que no es tan cercano a 2, entonces se rechaza que exista autocorrelación nula residual, por lo que no son independientes
En general podemos decir que no se cumple ni la homosedasticidad, ni la independencia en los residuos, por lo que debemos tratar de transformar algunas (o todas) las variables para que el modelo mejore
TRANSFORMACIONES
Haremos las transformaciones con la lambda correspondiente, y al hacer diversas pruebas, obtenemos el que es el mejor modelo, pues con menos transformaciones ayuda a mejorar el modelo anterior, después de esto nos queda el siguiente modelo

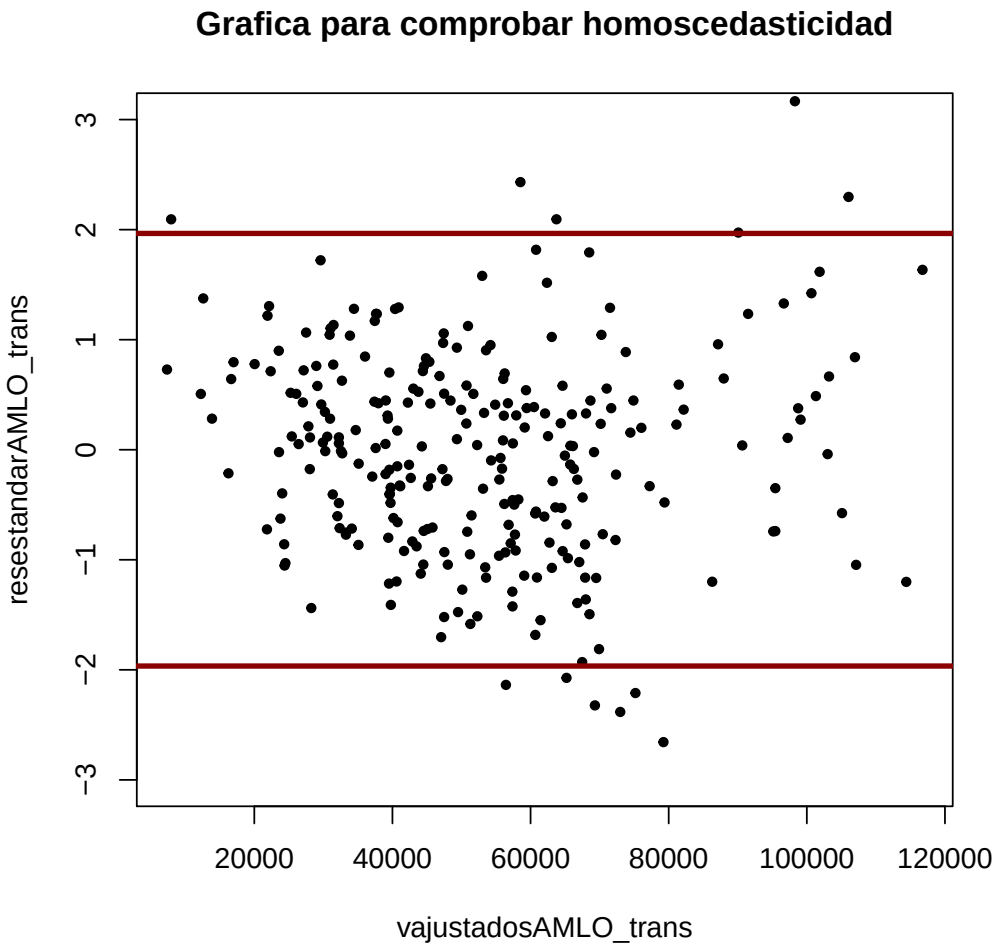
```
Call:
lm(formula = AMLO ~ IND_074 + IND_115_tra + IND_087_tra + IND_047_tra +
```

```
    IND_815 + IND_119 + IND_082_tra + IND_121_tra + IND_804 +
    IND_056 + IND_095 + Complejidad_tra + IND_002 + IND_088 +
    IND_064_tra, data = INEGI_INE_TRAIN)

Residuals:
   Min     1Q Median     3Q    Max
-31348  -8412    412   6793  56226

Coefficients:
                   Estimate  Std. Error t value Pr(>|t|)
(Intercept)       1.900e+05   5.299e+04   3.586 0.000403 ***
IND_074           4.396e+02   5.240e+01   8.389 3.43e-15 ***
IND_115_tra       3.180e+04   4.249e+03   7.486 1.17e-12 ***
IND_087_tra       6.436e+02   1.094e+02   5.881 1.28e-08 ***
IND_047_tra       4.992e+02   8.703e+02   0.574 0.566710
IND_815          -1.801e+03   3.260e+02  -5.524 8.18e-08 ***
IND_119          -6.064e+02   2.118e+02  -2.863 0.004545 **
IND_082_tra       1.709e+04   3.061e+03   5.584 6.02e-08 ***
IND_121_tra      -4.917e+03   9.551e+02  -5.148 5.28e-07 ***
IND_804           7.345e-02   5.971e-02   1.230 0.219780
IND_056          -1.336e+04   3.373e+03  -3.961 9.69e-05 ***
IND_095          -1.194e+03   2.500e+02  -4.777 3.01e-06 ***
Complejidad_tra  -3.465e+03   1.244e+03  -2.786 0.005742 **
IND_002           2.128e+02   9.855e+01   2.159 0.031797 *
IND_088          -1.410e+03   5.474e+02  -2.576 0.010572 *
IND_064_tra      7.480e-138  1.864e-138   4.012 7.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12040 on 254 degrees of freedom
Multiple R-squared:  0.7733,        Adjusted R-squared:  0.7599
F-statistic: 57.75 on 15 and 254 DF,  p-value: < 2.2e-16
```

Con el cual obtenemos un R^2 de 0.7599, es decir describe más varianza que incluso el modelo con todas las variables
Revisemos nuevamente todos los supuestos, para ver que el modelo realmente mejoró **Multicolinealidad** Para ver si se cumple este supuesto debemos ver si no existe correlación entre ellos, más aún si el vif (variance inflation factor) es menor a 10

```
       IND_074     IND_115_tra     IND_087_tra     IND_047_tra         IND_815
      1.762553        2.577077        1.629850        7.001221        3.005487
       IND_119     IND_082_tra     IND_121_tra         IND_804         IND_056
      2.310509        4.717035        5.590031        1.380166        2.118331
       IND_095 Complejidad_tra         IND_002         IND_088     IND_064_tra
      4.348969        5.034450        1.556172        1.405233        5.098825
```

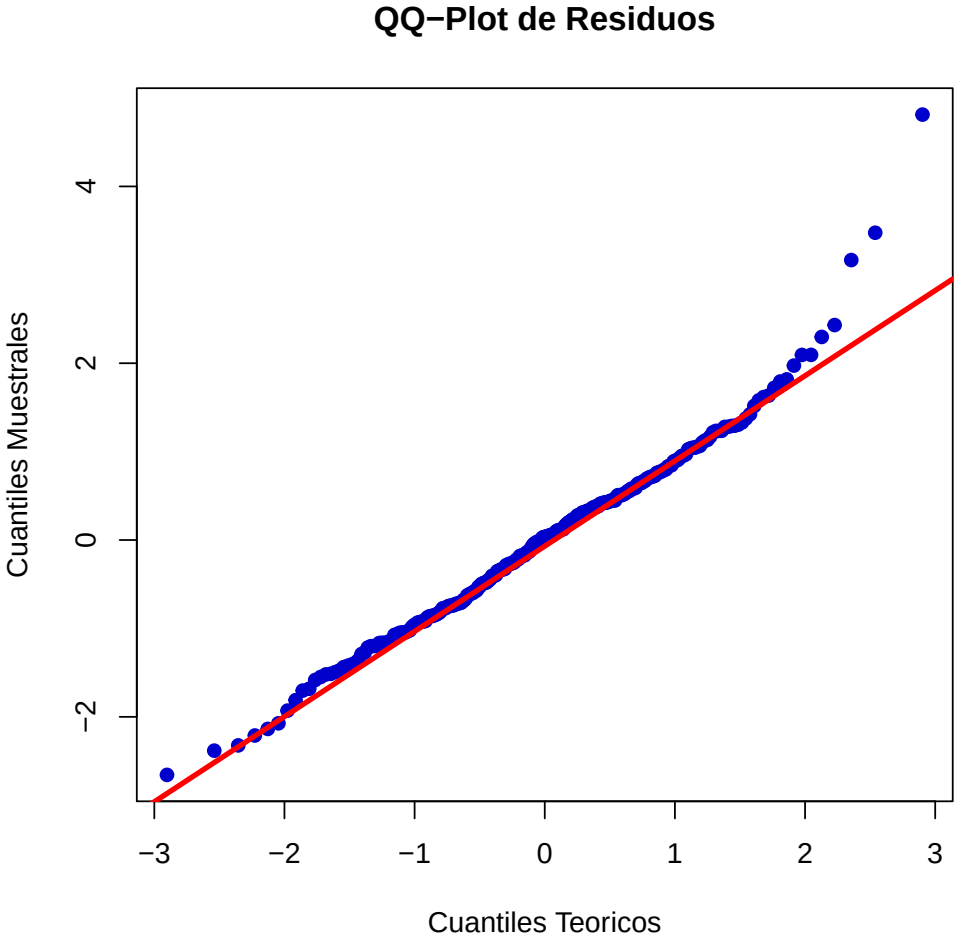
Los vif son menores a 10, de hecho son menores a 5, por lo que se puede descartar la multicolinealidad
Homosedasticidad
Para ver si este supuesto se cumple, primero veremos que sucede al graficar los residuos estandarizados, contra los valores ajustado. Como podemos observar en la gráfica siguiente, puede decirse que si es una g´rafica nula, por lo que tendriamos varianza constante, es decir, la homocedasticidad si se cumple.



Vemos que en la gŕafica se observa un comportamiento relativamente bueno, pues se ve como una gráfica nula, por lo que se puede decir que la homosedasticidad si se cumple

Normalidad

Al igual que para la prueba anterior, primero se hará una prueba gráfica



Este gráfico, al igual que el de antes de transformar, nos indica que los datos se ajustan bien en el centro y en la cola derecha de la distribución, sin mebargo puede haber probelmas en la cola derecha, para estar seguros se hara la siguiente prueba mucho más formal

```
        Anderson-Darling normality test

data:  resestandarAMLO_trans
A = 0.60023, p-value = 0.1181
```

Con esta prueba podemos decir que los residuos se distribuyen normal, pues el pvalor es mayor que el niven de significancia 0.05

Independencia

Para ver este supuesto se hará la prueba.

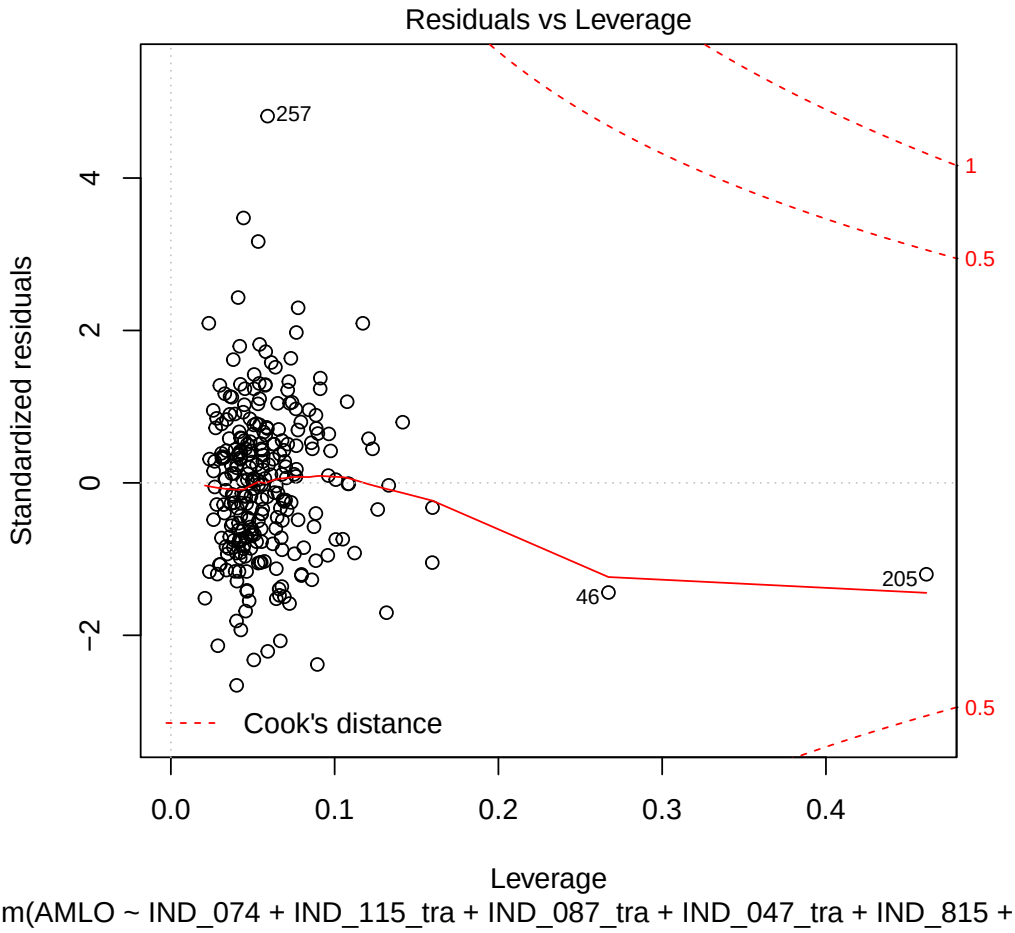
```
        Durbin-Watson test

data:  AMLO_trans
DW = 1.5979, p-value = 0.0001403
alternative hypothesis: true autocorrelation is greater than 0
```

Con esta prueba obtenemos un DW relativamente cercano a dos, con lo que se puede decir que los residuos son independientes

En general, se puede afirmar que el modelo transformado es superior al no transformado, pues mejora en todas las pruebas de los supuestos y además la R^2 aumenta

VALORES INFLUYENTES

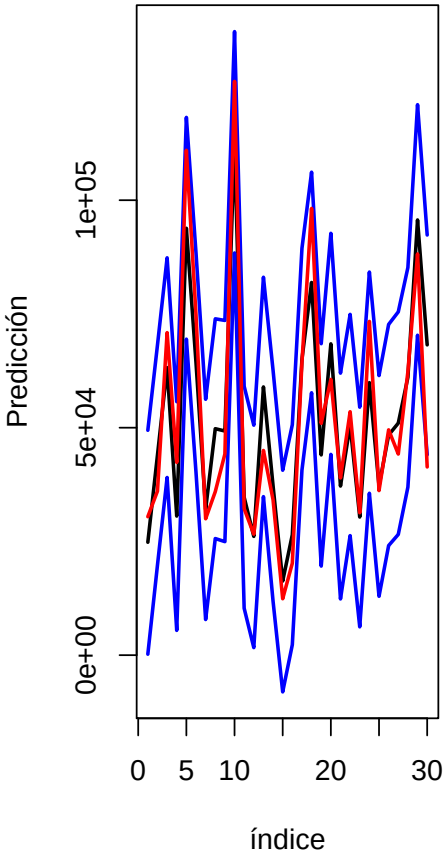
Para saber si existen valores influyentes que puedan estar afectando a nuestro modelo, se hace la siguiente gráfica



Se observa que no existe observaciones influyentes.
Entonces podemos decir que nuestro modelo es bueno, pues no existen observaciones que cambien su comportamiento.

PREDICCIÓN

Finalmente, y para poder validar nuestro modelo, haremos la predicción



Se observa en la gráfica que el modelo que se ajustó es bastante bueno para predecir nuestros datos de predicción, por lo que se puede concluir que sí es un buen modelo

INTERPRETACIÓN DE LOS PARÁMETROS

Como podremos observar, los parametros son sumamente pequeños, pues todas las variables son porcentajes Nuestras variables yafueron mencionadas, ahora bien, digamos cuales de ellas afectanpositivamente y cuales lo hacen negativamente.

Positivamente son.

1. Estimador del total de población de 15 años y más (Mujeres)
2. Porcentaje de asistencia escolar de la población de 3 a 5 años
3. Densidad de población (hab/km2)
4. Porcentaje de viviendas con separación de residuos
5. Porcentaje estatal de la población
6. Porcentaje de la población de 12 años y más separada
7. Porcentaje de población de 15 años y más con nivel de escolaridad media superior
8. Porcentaje de viviendas con disponibilidad de electricidad en la vivienda

Es decir, si estas variables aumentan, entonces los votos por AMLO también lo hacen
Ahora veamos que pasa con las negativas

1. Porcentaje de la población afiliada a servicios de salud

2. Porcentaje de la población de 12 años y más económicamente activa (Total)

3. Porcentaje de asistencia escolar de la población de 6 a 11 años

4. Porcentaje de la población ocupada que labora en el sector económico de comercio

5. Grupo de complejidad electoral

6. Porcentaje de la población afiliada al IMSS

7. Promedio de ocupantes por vivienda

Lo que dice que, si las variables antes mencionadas disminuyen, los votos por AMLO aumentan

CONCLUSIÓN Puesto que las variables significativas para cada modelo no son iguales estos no son comparables, sin embargo si podemos destacar aquellas que coinciden pues influyen, si bien no con la misma importancia si de manera general, tal es el caso de Porcentaje de población de 12 años y más separada que influye en las tres modelos puesto que mientras que para Enrique Peña Nieto y Josefina Vázquez Mota influyo de forma perjudicial a Andres Manuel López Obrador lo favoreció.

Puesto que las variables significativas para cada modelo no son iguales estos no son comparables, sin embargo si podemos destacar aquellas que coinciden pues influyen, si bien no con la misma importancia si de manera general, tal es el caso de Porcentaje de población de 12 años y más separada que influye en las tres modelos puesto que mientras que para Enrique Peña Nieto y Josefina Vázquez Mota influyo de forma perjudicial a Andres Manuel Lopez Obrador lo favorece.