PROYECTO FINAL DE REGRESIÓN MÚLTIPLE

Profesor: Javier Santibañez Cortez.
Alumnos: Gonzalo García Alarcón Estrada.
José Guerrero Rodríguez
Edgar Rodríguez Vázquez
Luis Fernando Torres Pineda.

8 DE JUNIO DE 2017

Regresión múltiple 17-2

Proyecto Final

Fecha de entrega: 08/06/2017

El objetivo es modelar el **PBIPP** de las entidades del país a partir de las variables sociodemográficas utilizadas en el índice de rezago social de CONEVAL. El conjunto de datos contiene las siguientes variables.

- **P15YM_NALF:** porcentaje de población de 15 años y más que no sabe leer y escribir.
- **P6A14_NASI:** porcentaje de población de 6 a 14 años que no asiste a la escuela.
- **P15YM_EBIN:** porcentaje de población de 15 años y más con educación básica incompleta.
- **VIV_PIS:** porcentaje de viviendas particulares habitadas con piso de tierra.
- **VIV_NEXC:** porcentaje de viviendas particulares habitadas que no disponen de excusado o sanitario.
- **VIV_NAGU:** porcentaje de viviendas particulares habitadas que no disponen de agua potable.
- **VIV_NDRE:** porcentaje de viviendas particulares habitadas que no disponen de drenaje.
- **VIV_NELE:** porcentaje de viviendas particulares habitadas que no disponen de electricidad.
- **VIV_NLAV:** porcentaje de viviendas particulares habitadas que no disponen de lavadora.
- **VIV_NREF:** porcentaje de viviendas particulares habitadas que no disponen de refrigerador.
- **PIBPP:** Producto interno bruto per capita en miles de pesos.

Los datos están en el archivo rezago.csv. Utilizar la información de las 32 entidades y los años 2005 y 2010 para hacer lo siguiente.

1. Hacer un análisis exploratorio de los datos.
2. Identificar observaciones atípicas, outliers.
3. Explorar si hay multicolinealidad en las variables explicativas.
4. Ajustar un modelo **RLM** tomando como respuesta a PIBPP y considerando las variables explicativas que sugieran las exploraciones anteriores (quizá sea necesario transformar a linealidad o eliminar por multicolinealidad). De igual manera, considerar las observaciones que sugieran los análisis previos (posiblemente sea necesario eliminar observaciones influyentes o atípicas).
5. Explorar no linealidad y heterocedasticidad. Confirmar los hallazgos con pruebas de falta de ajuste (no linealidad en las v. explicativas), no aditividad (no linealidad en la respuesta) y homocedasticidad.
6. En caso que las pruebas resulten positivas para no linealidad o heterocedasticidad, aplicar las medidas correctivas que se consideren necesarias y ajustar de nuevo el modelo con las variables transformadas.
7. Explorar la normalidad de los errores. Confirmar con alguna prueba de normaliadad.
8. Si el supuesto de normalidad es razonable, hacer la prueba de significancia del modelo y presentar la tabla ANOVA completa. Si no hay normalidad, hacer la prueba de significancia del modelo utilizando bootstrap (se debe aproximar la distribución del estadístico F).
9. Si el supuesto de normalidad es razonable, hacer pruebas de t simultáneas para las componentes del vector β. Si no hay normalidad, hacer pruebas individuales para los componentes del vector β utilizando bootstrap. Interpretar los resultados en el contexto del problema.
10. Calcular los coeficientes R^2 y R^2-ajustado. Interpretar los resultados.

Análisis exploratorio de datos.

En este primer punto se nos pide realizar un análisis exploratorio de los datos, para ello se procedió a verificar que los datos cumplieran con cada uno de los siguientes supuestos.

- Linealidad.
- Homocedasticidad.
- Independencia de las observaciones.
- Independencia lineal de las variables explicativas.
- Normalidad.

Para ello se ajustó un modelo de regresión múltiple de la variable PIBPP que contenga todas las variables explicativas, es decir.

$$\mathbf{PIBPP} = \mathbf{P15YM_{NALF}} + \mathbf{P6A14_{NASI}} + \mathbf{P15YM_{EBIN}} + \mathbf{VIV_{PIS}} + \mathbf{VIV_{NEXC}} + \mathbf{VIV_{NAGU}} + \mathbf{VIV_{NDRE}} + \mathbf{VIV_{NELE}} + \mathbf{VIV_{NLAV}} + \mathbf{VIV_{NREF}}$$

Una vez ajustado dicho modelo con R, se procedió a verificar cada uno de los supuestos mencionados con anterioridad obteniendo los siguientes resultados:

- El supuesto de linealidad no se cumple, ya que se realizaron las gráficas correspondientes entre la variable de respuesta (**PIBPP**) contra cada variable explicativa, así como también se realizaron las gráficas de los residuos contra cada variable explicativa, donde se observó que habían dos datos atípicos, y la relación que había entre las variables de respuesta con las variables explicativas era lineal, sin embargo, esta conclusión no era tan clara, ya que en algunas variables se puede percibir que los residuos presentaban un ligero patrón en su distribución a lo largo de la recta cero, por esta razón se decidió realizar las pruebas de ***falta de ajuste*** y de ***no aditividad de Tukey***, a fin de poder corroboran nuestra hipótesis de linealidad.

  En la prueba de falta de ajuste realizada para cada variable explicativa, se corroboró que en efecto la relación que hay entre cada variable con la variable de respuesta es lineal, debido a que los coeficientes cuadráticos de cada variable en cada regresión realizada fueron no significativos, lo que indica que hay evidencia para sostener que la relación entre la variable de respuesta y las variables explicativas es lineal.

  De la prueba de no aditividad de Tukey, se observó que no había aditividad en la variable de respuesta, lo que esto indica que se debe realizar un transformación de la variable de respuesta para corregir la falta de aditividad, en nuestro caso elegimos aplicar el logaritmo, y se ajustó nuevamente el RLM obteniendo las gráficas de los residuos contra cada variable, y se observó que el problema de no aditividad se había corregido.

- El supuesto de Homocedasticidad, de las gráficas no se cumple, de los gráficos de los residuos contra cada variable se pudo observar patrones en la distribución de los residuos alrededor de la recta y = 0.

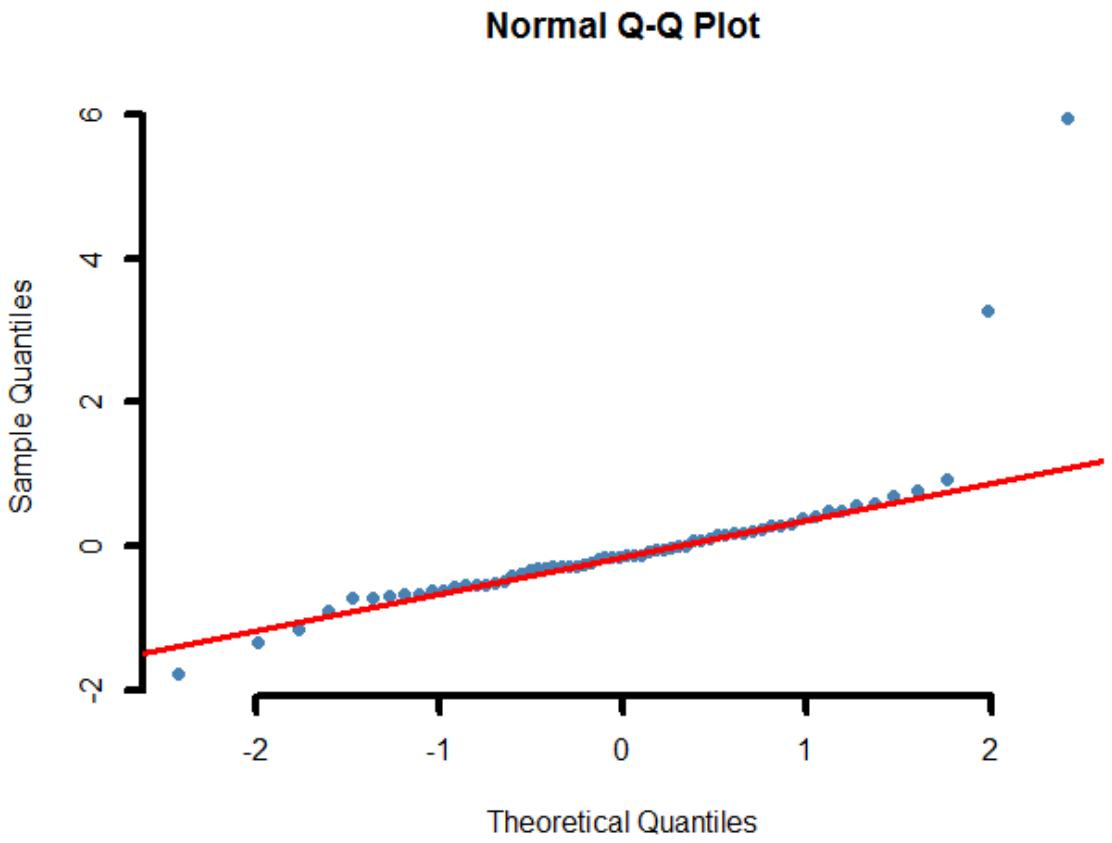
- El supuesto de independencia de las observaciones se verificó a través de la prueba de Durbin-Watson, la cual se muestra a continuación.

```
dwtest(modelo)
Durbin-Watson test
data:  modelo
DW = 1.8958, p-value = 0.3385
alternative hypothesis: true autocorrelation is greater than 0
```

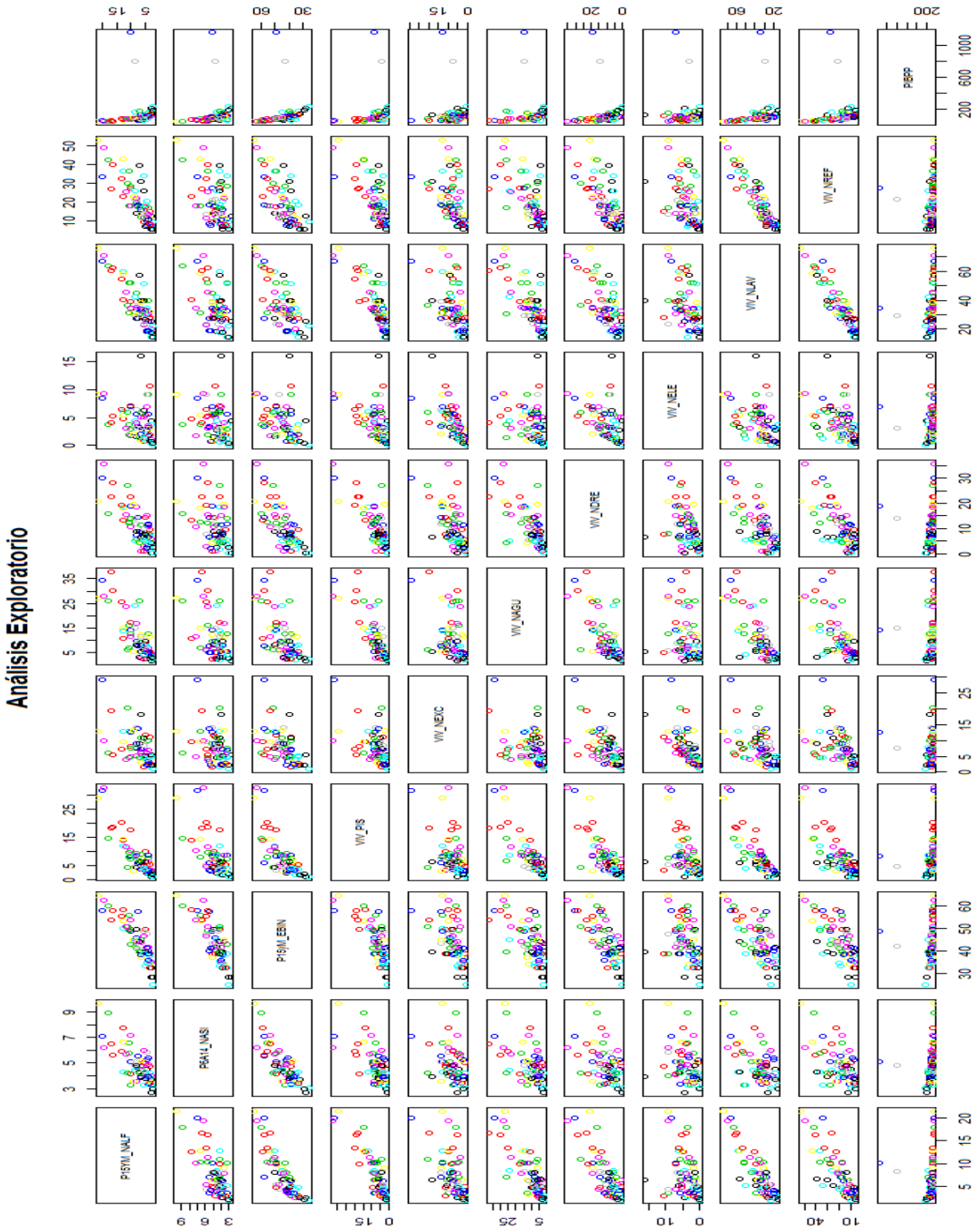
Con el resultado anterior se puede observar que hay independencia en las observaciones, ya que se tiene un P-value de 0.3385>0.05, el cual apoya a la hipótesis nula, que en este caso es la independencia de las observaciones.

- Para verificar la independencia de las variables explicativas, se ajustaron 10 modelo de RLM, en los cuales se tomó como variable de respuesta una variable explicativa y se generó el modelo el resto de las demás variables, con esto se pudo observar que algunas variables explicativas eran casi explicadas por el resto, algunas variables presentaban multicolinealidad.

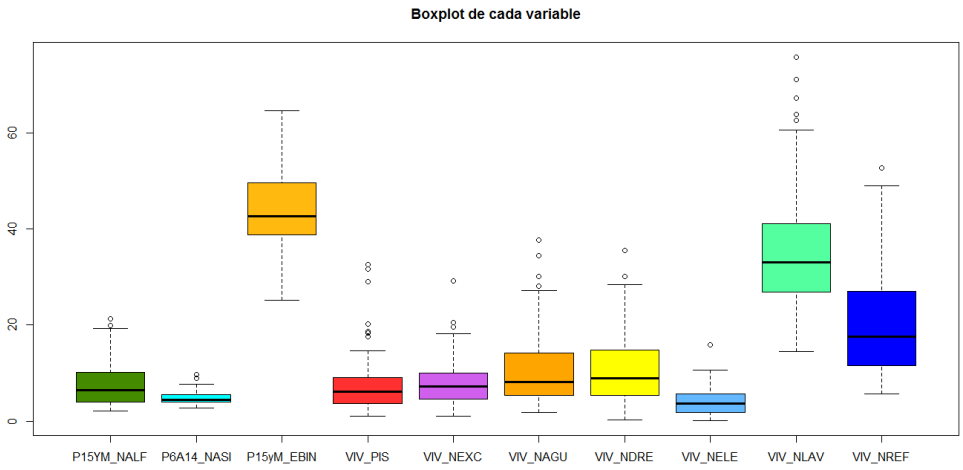
- Para verificar el supuesto de Normalidad, se ajustó un qqnorm, de los residuos del modelo y se observó que los datos presentaban normalidad pero tenían colas pasadas.



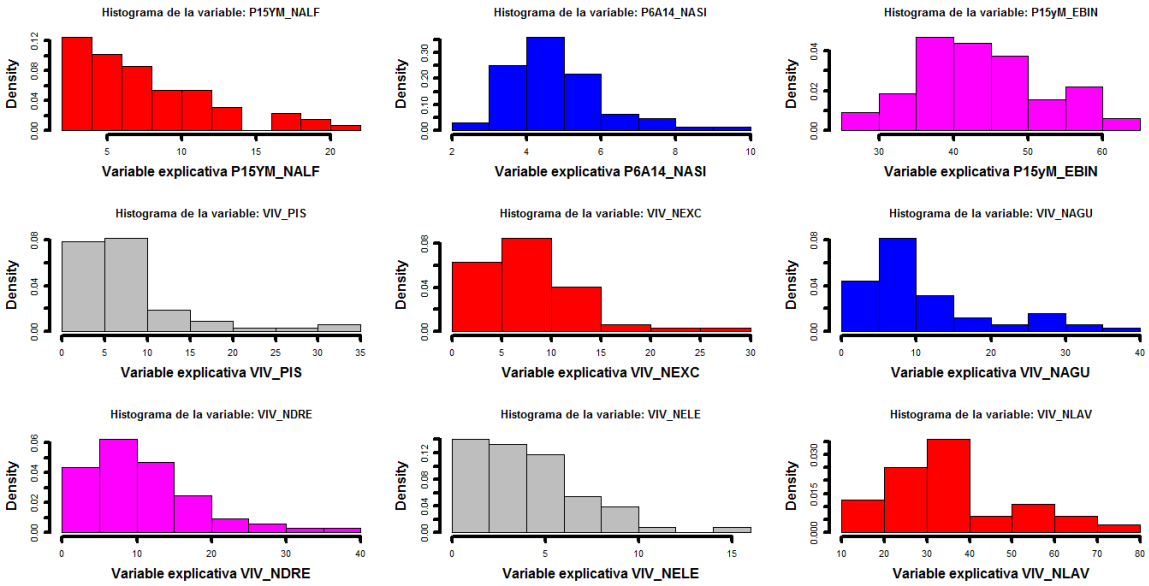
Más análisis exploratorio…

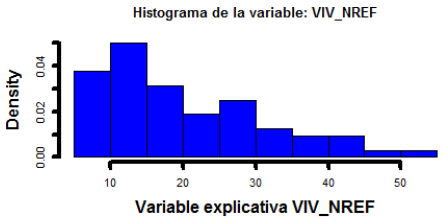


La gráfica de *pairs* nos muestra la relación que hay entre todas las variables, de ella se puede observar que la gran mayoria de las variables presentan una relación lineal entre ellas, sin embrago, este comportamiento no se presenta cuando las gráficas se realizan contra el PIBPP, estas gráficas se muestran en el último renglón y columna de la gráfica anterior.



Por otro lado, en los gráficos de *box-plot* correspondientes para cada variable explicativa se aprecia la gran mayoria de las variables presentan observaciones atípicas, ademas que la variable VIV_NLAV es la variable que presenta una mayor dispersión de sus observaciones, mientras que la variable P6A14_NASI es la variable que tiene una disperción muy pequeña, ya que su boxplot esta muy reducido, también se puede observarque la mayoria de las variables presenta un sesgo positivo, ya que la caja de las variables se encuentra muy cercano al bigote inferior, este comportamiento no se aprecia en las variables P15YM_EBIN y VIV_NLAV, las cuales presentan una gran dispersión, pero una simetría correcta.

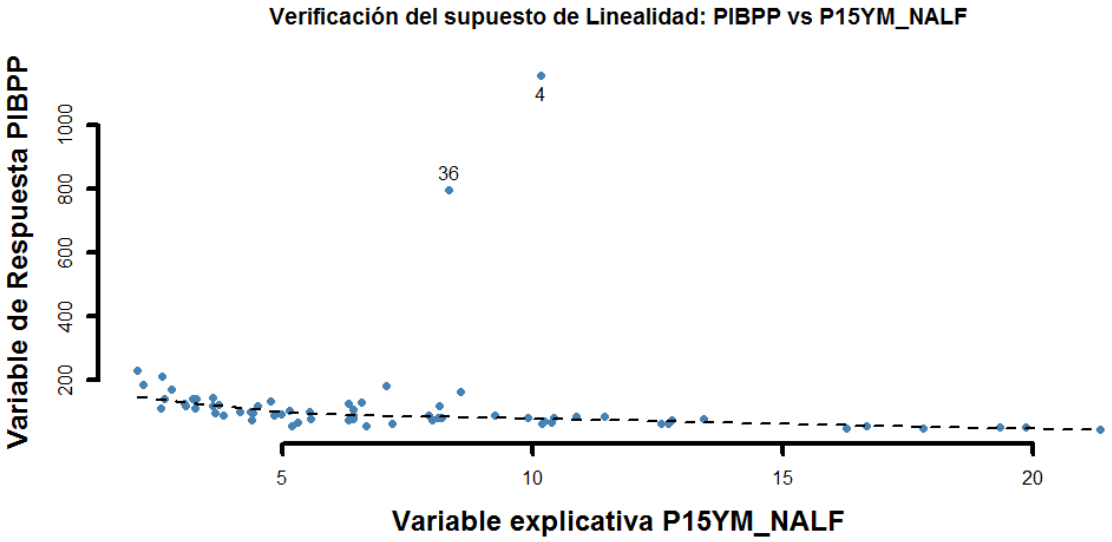




Para finalizar, de las gráficas de boxplot se observa que efectivamente las variables P15YM_EBIN y VIV_NLAV presentan un comportamiento simétrico en sus datos, además se puede observar que la gran mayoría de los datos presentan un sesgo a la derecha, teniendo una mayor cantidad de observaciones en los primeros valores del eje x.

Identificar Observaciones atípicas, *outliers.*

Para identificar si había observaciones atípicas en los datos, se realizaron las gráficas de la variable de respuesta contra cada variable explicativa, de ahí como se muestra en la siguiente imagen.



Se perciben dos observaciones que, intuímos, son atípicas. Dichas observaciones corresponden a la entidad de Campeche, para el año 2005 (punto 4) y 2010 (punto36), sin embargo, para identificar y eliminar alguna posible influencia de otros datos atípicos, que no se perciben en las gráficas realizadas, se utilizaron los residuos de *validación cruzada o jackknife,* los cuales se calcularon utilizando los residuos studentizados, es decir.

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}; \quad t_i = r_i\sqrt{\frac{n-p-1}{n-p-r^2{}_i}}$$

De esta manera, al calcular los residuos de *validación cruzada,* y bajo el supuesto de normalidad de los errores se tiene que $t_i \sim t_{n-p-1}$, entonces si ocupamos la corrección de Bonferroni y se

compara el cuantil $\alpha^* = \frac{\alpha}{2n}$ de una distribución t_{n-p-1} se podrá concluir que la i-ésima observación es atípica si ocurre.

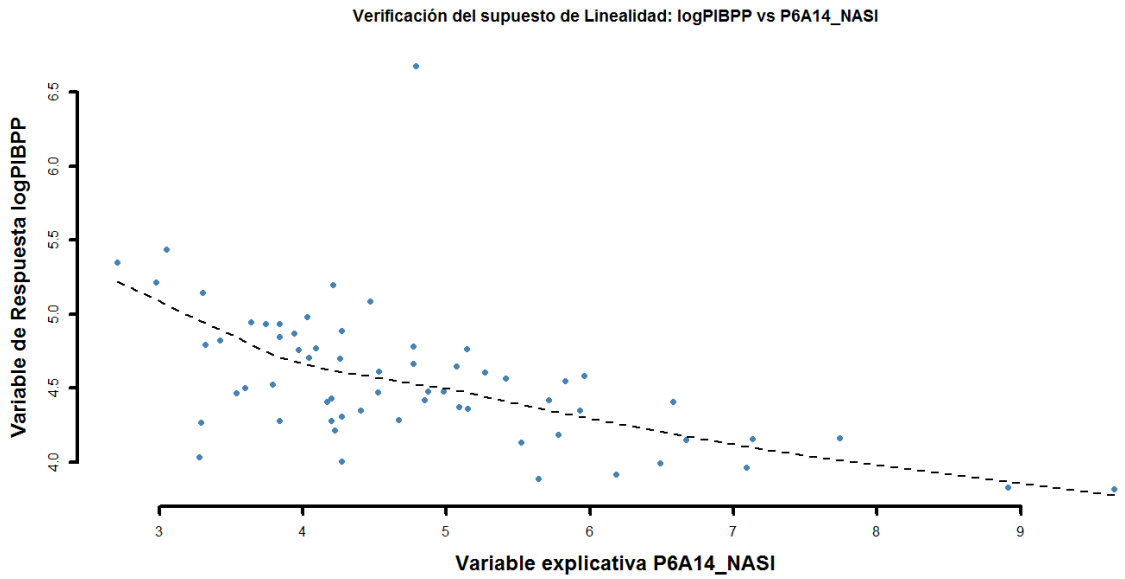
$|t_i| > t_{n-p-1}{}^{(1-\alpha^*)}$; $con\ \ \alpha^* = \frac{0.05}{2n}$

Por lo tanto se precedió a realizar la prueba para cada observación y se obtuvieron los siguientes resultados.

```
Residuos Estandarizados Residuos de Jackknife Datos positivos Tipo de Observacion
4                5.942125             8.076176        8.076176             Atipica
Tipo de Observacion Entidad Year
Atipica         Campeche 2005
```

De aquí se observa que los dos puntos que se señalaron como observaciones atípicas, solo Campeche para el año 2005 resulto ser atípica, además se sugeriría eliminar de igual forma la entidad de Campeche para el año 2010 pues presenta un valor grande, al momento de calcular el valor t para esta entidad en el año 2010, resultó ser de $t_{36} = 3.261937$, mientras que el cuantil de la prueba tiene un valor de $q = 3.564664$, donde la diferencia que hay entre ambos valores no es muy grade, sin embargo, no se decidió eliminar ya que este es un caso donde dos o más outliers juntos pueden ocultarse entre ellos, en nuestro caso la observación de Campeche del año 2005 oculto el efecto de la observación de Campeche para el año 2010.

Por lo tanto se concluye que la entidad de Campeche para el año 2005, es una observación atípica para nuestro estudio, por lo tanto se elimina de nuestra base.



Explorar si hay multicolinealidad en las variables explicativas.

Para explorar si hay multicolinealidad, se procedió a ajustar las regresiones lineales de cada variable explicativa contra las demás, con el objetivo de estudiar el coeficiente Adjusted R-squared (R^2) de cada regresión, ya que si dicho coeficiente es cercano a 1, esto indicaría que la variable explicativa X_i es *casi* una combinación lineal de las demás variables explicativas, por lo que eliminar esta variable no afectaría a nuestro estudio ya que esta es una combinación de las demás. Los coeficientes Adjusted R-squared correspondiente de cada modelo se muestran a continuación.

Adjusted

Modelo	Variable	Adjusted R-squared
1	P15YM_NALF	0.9401
2	P6A14_NASI	0.7816
3	P15yM_EBIN	0.8538
4	VIV_PIS	0.8770
5	VIV_NEXC	0.5941
6	VIV_NAGU	0.7279
7	VIV_NDRE	0.7908
8	VIV_NELE	0.5662
9	VIV_NLAV	0.9158
10	VIV_NREF	0.9029

En la tabla anterior se puede observar que la gran mayoría de las variables presentan un coeficiente de ajuste muy alto casi cercano a uno, esto indica que estas variables son casi linealmente dependientes, por esta razón se decidió considerar solo las variables con un Adjusted R-squared menor o igual a 0.80, quedando las siguientes variables para nuestro estudio.

Modelo	Variable	Adjusted R-squared
2	P6A14_NASI	0.7816
5	VIV_NEXC	0.5941
6	VIV_NAGU	0.7279
7	VIV_NDRE	0.7908
8	VIV_NELE	0.5662

Con estas variables explicativas, se procedió a verificar nuevamente la multicolinealidad entre ellas, dando como resultado los siguientes Adjusted R-squared

Modelo	Variable	Adjusted R-squared
1	P6A14_NASI	0.2940
2	VIV_NEXC	0.5139
3	VIV_NAGU	0.5211
4	VIV_NDRE	0.5600
5	VIV_NELE	0.3731

De la tabla anterior se puede observar que los valores Adjusted R-squared, disminuyen para cada una de las variables, sin embargo, la variable que tuvo el decremento más llamativo fue **P6A14_NASI**, la cual paso de un valor de 0.7816 a 0.2940, la comparación de ambos valores se muestra a continuación.

Modelo	Variable	Adjusted R-squared Original	Adjusted R-squared Nuevo
1	P6A14_NASI	0.7816	0.2940
2	VIV_NEXC	0.5941	0.5139
3	VIV_NAGU	0.7279	0.5211
4	VIV_NDRE	0.7908	0.5600
5	VIV_NELE	0.5662	0.3731

Por lo que el nuevo modelo a ajustar con las variables explicativas es el siguiente.

$$\mathbf{log(PIBPP)} = \mathbf{P6A14_{NASI}} + \mathbf{VIV_{NEXC}} + \mathbf{VIV_{NAGU}} + \mathbf{VIV_{NDRE}} + \mathbf{VIV_{NELE}}$$

Ajustar modelo RLM.

En esta sección se procederá a ajustar el modelo, para ello serán considerados todas las sugerencias que se observaron en los análisis previos, para ello es importante recordar cuales son las observaciones que se encontraron durante el análisis que se realizó, dichas observaciones se muestran a continuación.

1. Al realizar la prueba de falta de ajuste para corroborar la linealidad de las observaciones, se pudo encontrar que todas las variables explicativas cumplían con esta prueba, sin embargo, al realizar la prueba de no aditividad de Tukey para ver detectar la falta de linealidad de la variable de respuesta, se observó que esta no se cumple, es decir, la prueba nos indica que la variable de respuesta no es lineal, por esta razón se realizar una transformación a la variable de respuesta, la cual es el logaritmo $\mathbf{log(PIBPP)}$.
2. De las pruebas de multicolinealidad que se realizaron a las variables explicativas, se concluyó que la gran mayoría de las variables explicativas son casi linealmente independientes, ya que la regresión que ajusto a cada una presentaron un valor R^2 muy cercano a uno, por lo que se decidió quedarse con aquellas variables que tuvieran un R^2 menor a 0.8, dando como resultado las siguientes variables explicativas $\mathbf{P6A14_{NASI}}$; $\mathbf{VIV_{NEXC}}$; $\mathbf{VIV_{NAGU}}$; $\mathbf{VIV_{NDRE}}$; $\mathbf{VIV_{NELE}}$.
3. Para finalizar se observaron datos atípicos, dichas observaciones corresponden a la entidad de Campeche para el año 2005, esta observación se corroboró al comparar los valores de los residuos de validación cruzada con el cualtil $|t_i| > t_{n-p-1}{}^{(1-\alpha^*)}$; por lo que se decidió eliminar esta entidad para continuar con nuestro estudio.

Una vez recapitulados las observaciones estas se aplicarán para la modelación de los datos sin las observaciones atípicas, y para ello se procederá a ajusta las observaciones con el siguiente modelo.

$$\log(\mathbf{PIBPP}) = \mathbf{P6A14_{NASI}} + \mathbf{VIV_{NEXC}} + \mathbf{VIV_{NAGU}} + \mathbf{VIV_{NDRE}} + \mathbf{VIV_{NELE}}$$

Obteniendo los siguientes resultados.

```
> rezago2 = rezago[-c(4),-c(1,2,3,4,6,7,12,13)]
> modelo.nuevo = lm(log(PIBPP) ~ P6A14_NASI+VIV_NEXC+VIV_NAGU+VIV_NDRE+VI
V_NELE, rezago2)
> summary(modelo.nuevo)

Call:
lm(formula = log(PIBPP) ~ P6A14_NASI + VIV_NEXC + VIV_NAGU +
    VIV_NDRE + VIV_NELE, data = rezago2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79416 -0.16383  0.00087  0.14672  2.18397

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.434785   0.185707  29.265  < 2e-16 ***
P6A14_NASI  -0.156160   0.043881  -3.559 0.000759 ***
VIV_NEXC    -0.006039   0.014121  -0.428 0.670496
VIV_NAGU     0.007685   0.008449   0.910 0.366898
VIV_NDRE    -0.022864   0.009893  -2.311 0.024469 *
VIV_NELE     0.019457   0.021459   0.907 0.368377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3846 on 57 degrees of freedom
Multiple R-squared:  0.3719,      Adjusted R-squared:  0.3168
F-statistic:  6.75 on 5 and 57 DF,  p-value: 5.283e-05
```

Se puede observar que este modelo es mucho mejor que el modelo que considera todas las variables explicativas, ya que se puede observar que el Adjusted R-squared: 0.3168, mientras que para el modelo con todas las variables explicativas se tiene un Adjusted R-squared: 0.1201, por lo tanto se puede concluir que este modelo ajusta mejor a los datos que el primer modelo.

Notemos que si ajustamos un modelo eliminando la entidad de Campeche para los años 2005 y 2010, se obtienen los siguientes resultados.

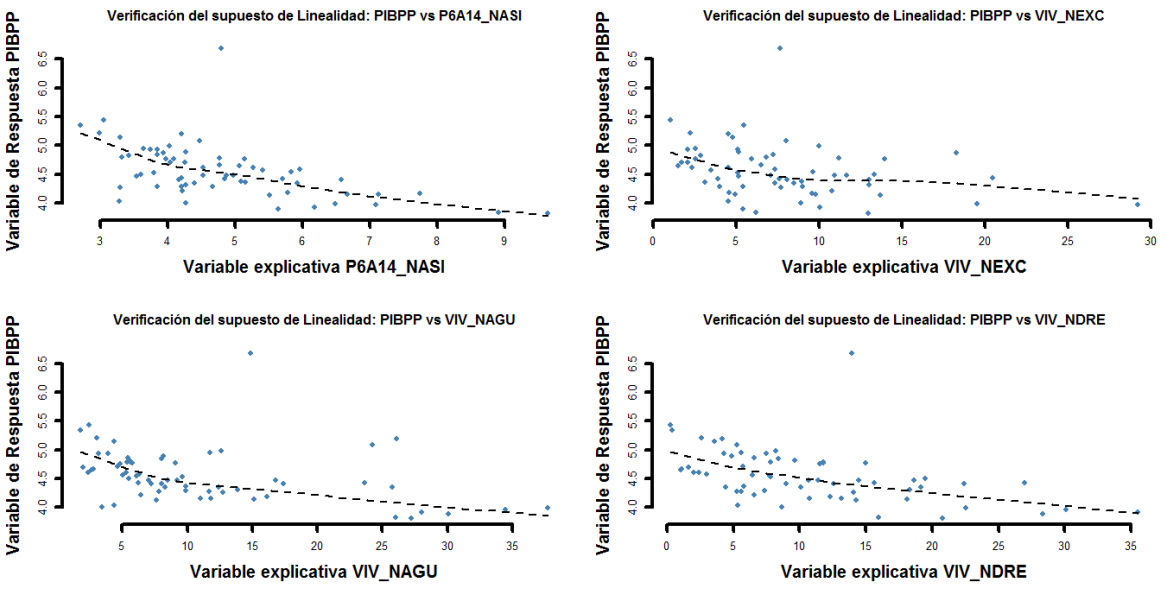
```
•       modelo.nuevo = lm(log(PIBPP) ~ P6A14_NASI+VIV_NEXC+VIV_NAGU+VIV_NDRE+VIV_
NELE, rezago2)
•       summary(modelo.nuevo)

Call:
lm(formula = log(PIBPP) ~ P6A14_NASI + VIV_NEXC + VIV_NAGU +
VIV_NDRE + VIV_NELE, data = rezago2)
Residuals:
Min       1Q   Median       3Q      Max
•       0.74951 -0.10194  0.02689  0.15077  0.51100

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.374525   0.121502  44.234  < 2e-16 ***
P6A14_NASI  -0.147729   0.028681  -5.151 3.48e-06 ***
VIV_NEXC    -0.006054   0.009224  -0.656 0.514323
```

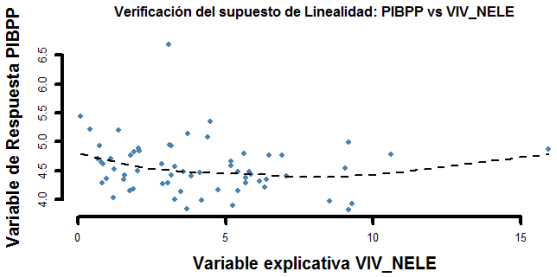
```
VIV_NAGU     0.005834   0.005523   1.056 0.295372
VIV_NDRE    -0.025207   0.006468  -3.897 0.000262 ***
VIV_NELE     0.026665   0.014042   1.899 0.062720 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.2512 on 56 degrees of freedom
Multiple R-squared:  0.6005,      Adjusted R-squared:  0.5648
F-statistic: 16.83 on 5 and 56 DF,  p-value: 3.988e-10
```

Lo primero que se puede observar es que este modelo es mucho mejor a los dos modelos anteriores, al modelo que considera todas las variables explicativas y que tiene todas las entidades y al modelo que se ajustó eliminando solo la entidad de Campeche para el año 2005, ya que éste considera la sugerencia de eliminar la entidad de Campeche para los años 2005 y 2010, obteniendo un Adjusted R-squared: 0.5648, mientras que para el modelo con todas las variables explicativas se tiene un Adjusted R-squared: 0.1201 y el modelo que sólo elimina la entidad de Campeche para el año del 2005 se tiene un Adjusted R-squared: 0.3168, por lo tanto se puede concluir que este modelo ajusta mejor a los datos que los dos anteriores, sin embargo, el modelo que considera el eliminar solo a Campeche para el año 2005, presenta también un buen ajuste y como se mencionó en la sección donde se identificaron los valores atípicos, se decidió sólo eliminar a Campeche para el año del 2005, realizando los siguientes análisis con la base que considera a Campeche para el año del 2010.

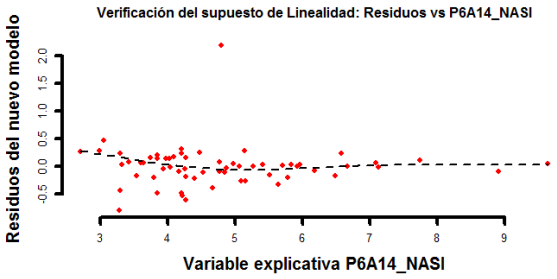
Pruebas de no linealidad y heterocedasticidad.

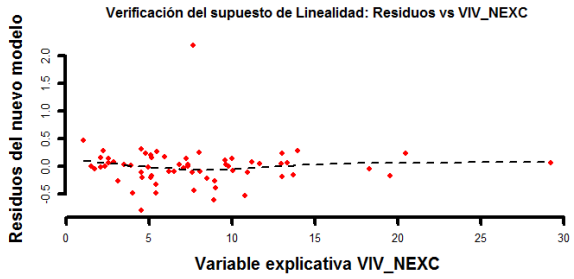
En esta sección se probará si se cumplen los criterios de linealidad y homocedasticidad, para ello se inicia por verificar la linealidad de las variables explicativas con la variable de respuesta, a través de las gráficas de las variables explicativas que se incluyeron en el modelo de la sección anterior con la variable de respuesta transformada, dichas gráficas se muestran a continuación.

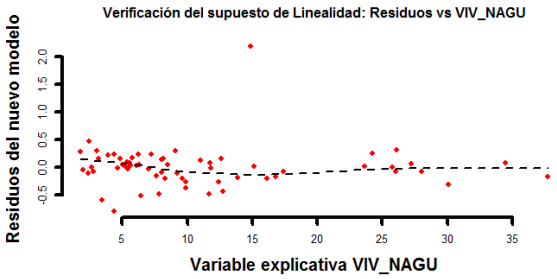


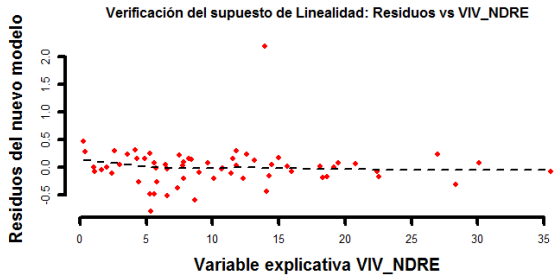


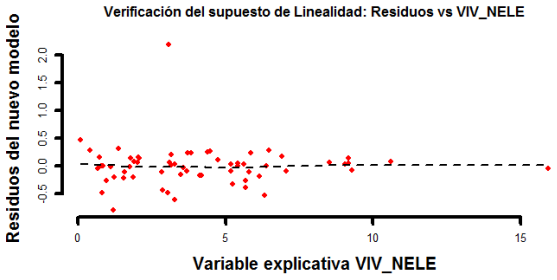
De las gráficas anteriores, se puede observar que la relación lineal entre la mayoría de las variables explicativas con la transformación de la variable de respuesta es lineal, a excepción de la variable de respuesta VIV_NELE, que se logra apreciar una pequeña curvatura en la dispersión de sus datos, sin embargo, como se sugiere en las notas del presente curso, una mejor comparación gráfica sería entre los residuos del modelo ajustado contra cada variable explicativa, ya que el grupo anterior de gráficas al ser la comparación marginal de cada variable explicativa contra la variable de respuesta transformada, puede llevar a concluir de manera errónea. Las graficas entre los residuos del modelo ajustado con cada variable explicativa se muestran a continuación.











De este grupo de gráficas se puede observar con mayor claridad que la relación que hay entre las variables explicativas que conforman el nuevo modelo es lineal con la transformación de la variable de respuesta, no obstante, se puede observar que hay un problema de homocedasticidad, mismo que se verifica más adelante. A continuación se procede a realizar las pruebas de *falta de ajuste* y de *no aditividad de Tukey*, a fin de corroborar la linealidad de las variables explicativas y la aditividad de la variable de respuesta.

Prueba de falta de ajuste

A continuación, se muestra los resultados de la prueba de falta de ajuste para cada variable explicativa.

Población de 6 a 14 años que no asiste a la escuela

```
•       lack.of.fit.v1 = lm(rezago2$logPIBPP ~ rezago2$P6A14_NASI+rezago2$VIV_NEX
C+rezago2$VIV_NAGU+rezago2$VIV_NDRE+
•       rezago2$VIV_NELE+I(rezago2$P6A14_NASI^2), rezago2)
•       summary(lack.of.fit.v1)

Call:
lm(formula = rezago2$logPIBPP ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC +
rezago2$VIV_NAGU + rezago2$VIV_NDRE + rezago2$VIV_NELE +
I(rezago2$P6A14_NASI^2), data = rezago2)
Residuals:
Min      1Q  Median      3Q     Max
•       0.8069 -0.1570  0.0105  0.1446  2.1911

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.553494   0.550739  10.084 3.38e-14 ***
rezago2$P6A14_NASI      -0.201945   0.204601  -0.987   0.3279
rezago2$VIV_NEXC        -0.005709   0.014313  -0.399   0.6915
rezago2$VIV_NAGU         0.007646   0.008522   0.897   0.3735
rezago2$VIV_NDRE        -0.022845   0.009977  -2.290   0.0258 *
rezago2$VIV_NELE         0.019272   0.021655   0.890   0.3773
I(rezago2$P6A14_NASI^2)  0.003999   0.017449   0.229   0.8195
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.3878 on 56 degrees of freedom
Multiple R-squared:  0.3725,      Adjusted R-squared:  0.3052
F-statistic:  5.54 on 6 and 56 DF,  p-value: 0.0001461
```

De los resultados anteriores, se observa que el coeficiente **I(rezago2$P6A14_NASI^2**), correspondiente a la población de 6 a 14 años que no asiste a la escuela es no significativo, esto indica que no hay suficiente evidencia para sospechar que la relación entre esta variable con el logaritmo del **PIBPP** no es lineal, por lo que no se rechaza la hipótesis nula **H0:η1=0.**

Porcentaje de viviendas que no tiene excusado o sanitario

```
•       lack.of.fit.v2 = lm(rezago2$logPIBPP ~ rezago2$P6A14_NASI+rezago2$VIV_NEX
C+rezago2$VIV_NAGU+rezago2$VIV_NDRE+
•       rezago2$VIV_NELE+I(rezago2$VIV_NEXC^2), rezago2)
•       summary(lack.of.fit.v2)

Call:
lm(formula = rezago2$logPIBPP ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC +
rezago2$VIV_NAGU + rezago2$VIV_NDRE + rezago2$VIV_NELE +
I(rezago2$VIV_NEXC^2), data = rezago2)
Residuals:
```

```
Min       1Q   Median       3Q      Max
•       0.79254 -0.13765 -0.02117  0.12764  2.20425

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.4794092  0.2060936  26.587  < 2e-16 ***
rezago2$P6A14_NASI    -0.1533042  0.0445152  -3.444  0.00109 **
rezago2$VIV_NEXC      -0.0218908  0.0339562  -0.645  0.52177
rezago2$VIV_NAGU       0.0067829  0.0086837   0.781  0.43803
rezago2$VIV_NDRE      -0.0222063  0.0100395  -2.212  0.03107 *
rezago2$VIV_NELE       0.0235294  0.0230060   1.023  0.31082
I(rezago2$VIV_NEXC^2)  0.0005913  0.0011504   0.514  0.60926
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3871 on 56 degrees of freedom
Multiple R-squared:  0.3748,      Adjusted R-squared:  0.3079
F-statistic: 5.596 on 6 and 56 DF,  p-value: 0.000133
```

Al observar estos resultados, se observar que el coeficiente cuadrático de esta variable **I(rezago2$VIV_NEXC^2)**, es no significativo, esto implica que no hay suficiente evidencia en contra de la hipótesis nula, para suponer que la relación que hay entre **VIV_NEXC** con el **log(PIBPP)** no sea lineal.

Por lo que la variable **VIV_NEXC** es lineal con el **log(PIBPP).**

Porcentaje de viviendas que no disponen de agua potable.

```
•       lack.of.fit.v3 = lm(rezago2$logPIBPP ~ rezago2$P6A14_NASI+rezago2$VIV_NEX
C+rezago2$VIV_NAGU+rezago2$VIV_NDRE+
•       rezago2$VIV_NELE+I(rezago2$VIV_NAGU^2), rezago2)
•       summary(lack.of.fit.v3)

Call:
lm(formula = rezago2$logPIBPP ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC +
rezago2$VIV_NAGU + rezago2$VIV_NDRE + rezago2$VIV_NELE +
I(rezago2$VIV_NAGU^2), data = rezago2)
Residuals:
Min       1Q   Median       3Q      Max
•       0.78721 -0.14558  0.00581  0.13495  2.16518

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.4047240  0.2165576  24.957  < 2e-16 ***
rezago2$P6A14_NASI    -0.1569290  0.0443284  -3.540 0.000813 ***
rezago2$VIV_NEXC      -0.0047595  0.0149715  -0.318 0.751740
rezago2$VIV_NAGU       0.0138941  0.0240371   0.578 0.565561
rezago2$VIV_NDRE      -0.0233302  0.0101163  -2.306 0.024824 *
rezago2$VIV_NELE       0.0183068  0.0220323   0.831 0.409555
I(rezago2$VIV_NAGU^2) -0.0001791  0.0006483  -0.276 0.783378
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3878 on 56 degrees of freedom
Multiple R-squared:  0.3727,      Adjusted R-squared:  0.3055
F-statistic: 5.546 on 6 and 56 DF,  p-value: 0.0001445
```

Como se observa, estos resultados indican que la hipótesis de linealidad entre la variable explicativa VIV_NAGU con el log(PIBPP) es lineal, ya que el coeficiente cuadrático la para variable de estudio I(rezago2$VIV_NAGU^2), es no significativo, esto indica que no hay suficiente evidencia para suponer que la relación que haya entre estas dos variables no sea lineal.

Porcentaje de viviendas que no disponen de electricidad.

```
•       lack.of.fit.v4 = lm(rezago2$logPIBPP ~ rezago2$P6A14_NASI+rezago2$VIV_NEX
C+rezago2$VIV_NAGU+rezago2$VIV_NDRE+
•       rezago2$VIV_NELE+I(rezago2$VIV_NDRE^2), rezago2)
•       summary(lack.of.fit.v4)

Call:
lm(formula = rezago2$logPIBPP ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC +
rezago2$VIV_NAGU + rezago2$VIV_NDRE + rezago2$VIV_NELE +
I(rezago2$VIV_NDRE^2), data = rezago2)
Residuals:
Min       1Q   Median       3Q      Max
•       0.79302 -0.16759  0.00592  0.14239  2.17612

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.4236319  0.2002711  27.081  < 2e-16 ***
rezago2$P6A14_NASI    -0.1571450  0.0447012  -3.515 0.000878 ***
rezago2$VIV_NEXC      -0.0061926  0.0142766  -0.434 0.666128
rezago2$VIV_NAGU       0.0078704  0.0086036   0.915 0.364231
rezago2$VIV_NDRE      -0.0199238  0.0211768  -0.941 0.350831
rezago2$VIV_NELE       0.0196832  0.0216926   0.907 0.368101
I(rezago2$VIV_NDRE^2) -0.0001006  0.0006392  -0.157 0.875496
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.3879 on 56 degrees of freedom
Multiple R-squared:  0.3722,      Adjusted R-squared:  0.3049
F-statistic: 5.533 on 6 and 56 DF,  p-value: 0.0001479
```

Los resultados, para esta variable indican que no se rechaza la hipótesis de linealidad entre la variable **VIV_NELE** con el **log(PIBPP),** ya que el coeficiente cuadrático para esta variable es no significativo, lo que implica que no hay evidencia para suponer que la relación entre estas dos variables no sea lineal.

Para finalizar se realiza esta última prueba para la variable **VIV_NREF,** donde los resultados se muestran a continuación.

Porcentaje de viviendas que no cuentan con refrigerador.

```
•       lack.of.fit.v5 = lm(rezago2$logPIBPP ~ rezago2$P6A14_NASI+rezago2$VIV_NEX
C+rezago2$VIV_NAGU+rezago2$VIV_NDRE+
•       rezago2$VIV_NELE+I(rezago2$VIV_NELE^2), rezago2)
•       summary(lack.of.fit.v5)

Call:
lm(formula = rezago2$logPIBPP ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC +
rezago2$VIV_NAGU + rezago2$VIV_NDRE + rezago2$VIV_NELE +
```

```
I(rezago2$VIV_NELE^2), data = rezago2)
Residuals:
Min      1Q  Median      3Q     Max
•       0.7940 -0.1638  0.0011  0.1467  2.1840

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.435e+00  1.935e-01  28.079  < 2e-16 ***
rezago2$P6A14_NASI    -1.562e-01  4.459e-02  -3.503 0.000912 ***
rezago2$VIV_NEXC      -6.038e-03  1.425e-02  -0.424 0.673368
rezago2$VIV_NAGU       7.686e-03  8.526e-03   0.901 0.371208
rezago2$VIV_NDRE      -2.288e-02  1.034e-02  -2.213 0.031017 *
rezago2$VIV_NELE       1.970e-02  5.132e-02   0.384 0.702570
I(rezago2$VIV_NELE^2) -1.874e-05  3.611e-03  -0.005 0.995878
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.388 on 56 degrees of freedom
Multiple R-squared:  0.3719,      Adjusted R-squared:  0.3046
F-statistic: 5.526 on 6 and 56 DF,  p-value: 0.0001495
```

De estos resultados se observar que la hipótesis de nula no se rechaza, ya que el término cuadrático para esta variable es no significativo, esto indica que no hay evidencia para suponer que la relación entre **VIV_NREF** y el **log(PIBPP)** no sea lineal.

De los análisis anteriores para la falta de ajuste, se pudo confirmar que la relación que hay entre cada variable con el log(PIBPP) es lineal, ya que el término cuadrático correspondiente a cada variable no resulto significativo, permitiendo no rechazar la hipótesis nula **H0:η1=0.**

La prueba de *no aditividad de Tukey* para detectar la no aditividad en la variable explicativa se muestra a continuación.

Prueba de no aditividad de Tukey

```
•       Z = fitted(modelo.nuevo)^2/(2*mean(rezago2$logPIBPP))
•       summary(lm(rezago2$logPIBPP ~ rezago2$P6A14_NASI+rezago2$VIV_NEXC+rezago2
$VIV_NAGU+rezago2$VIV_NDRE+rezago2$VIV_NELE+Z, rezago2))

Call:
lm(formula = rezago2$logPIBPP ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC +
rezago2$VIV_NAGU + rezago2$VIV_NDRE + rezago2$VIV_NELE +
Z, data = rezago2)
Residuals:
Min       1Q   Median       3Q      Max
•       0.82594 -0.14025  0.01386  0.13988  2.21439

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.11474   15.44886  -0.331    0.742
rezago2$P6A14_NASI  0.35462    0.74923   0.473    0.638
```

```
rezago2$VIV_NEXC    0.01336    0.03176   0.421    0.676
rezago2$VIV_NAGU   -0.01930    0.04041  -0.478    0.635
rezago2$VIV_NDRE    0.05417    0.11323   0.478    0.634
rezago2$VIV_NELE   -0.04678    0.09936  -0.471    0.640
Z                   3.36118    4.92180   0.683    0.497

Residual standard error: 0.3864 on 56 degrees of freedom
Multiple R-squared:  0.3771,      Adjusted R-squared:  0.3103
F-statistic:  5.65 on 6 and 56 DF,  p-value: 0.0001216
```

Es importante recordar que esta prueba consiste en ajustar una RLM con las variables explicativas, incluyendo un término **Z,** dicho parámetro es calculado como el cociente entre los valores ajustados al cuadrado y dos veces la media de la variable de respuesta, es decir.

$$Z = \frac{\hat{Y}}{\frac{2\sum_{i=1}^{63} log(PIBPP_i)}{63}}$$

Este parámetro permite capturar la no linealidad entre las variables explicativas con la variable de respuesta, contrastando las siguientes hipótesis **H_0 : Z= 0 vs. H_1 : Z≠ 0.**

De esta manera se puede observar que el parámetro Z es no significativo, lo que indica que si hay aditividad entre las variables explicativas y la variable de respuesta, es decir, la variable de respuesta log(PIBPP) es lineal.

Para finalizar esta sección se debe probar que el criterio de homocedasticidad se cumple, es decir, se tiene que verificar que se tiene varianza constante, sin embargo, se debe recordar que al realizar la gráficas entre el cada variable explicativa vs residuos del modelo ajustado, se observó que había linealidad pero posiblemente no **homocedasticidad**, ya que se lograba apreciar un ligero patrón en la distribución de los residuos a lo largo del eje x = 0 para la algunas variable, por lo tanto, para verificar estas sospechas se realiza la siguientes prueba para cada variable explicativa del modelo.

Dicha prueba consiste en ajustar un modelo de regresión a los residuos estandarizados al cuadrado, es decir.

$$e^{*2}{}_i = \gamma_0 + \gamma_0 X_{ki} + \eta$$

Donde $e^{*2}{}_i$ es el i-ésimo residuo estandarizado.

De esta manera procedió a calcular los residuos estandarizados y se ajustaron 5 regresiones, una para cada variable explicativa, y una última regresión tomando todas las variables explicativas, dichos resultados se muestran a continuación.

Porcentaje de la población de 6 a 14 años que no asiste a la escuela.

```
•       homocedasticidad.P6A16 = lm(u1 ~ rezago2$P6A14_NASI)
•       summary(homocedasticidad.P6A16)

Call:
lm(formula = u1 ~ rezago2$P6A14_NASI)
Residuals:
Min      1Q  Median      3Q     Max
•       1.1411 -0.8660 -0.7294 -0.3551 31.3364

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.6947     1.8999   0.892    0.376
rezago2$P6A14_NASI  -0.1643     0.3803  -0.432    0.667

Residual standard error: 4.101 on 61 degrees of freedom
Multiple R-squared:  0.003051,    Adjusted R-squared:  -0.01329
F-statistic: 0.1867 on 1 and 61 DF,  p-value: 0.6672
```

Porcentaje de viviendas que no cuentan con excusado o sanitario.

```
•       homocedasticidad.NEXC = lm(u1 ~ rezago2$VIV_NEXC)
•       summary(homocedasticidad.NEXC)

Call:
lm(formula = u1 ~ rezago2$VIV_NEXC)
Residuals:
Min      1Q  Median      3Q     Max
•       1.0142 -0.8757 -0.7565 -0.4428 31.3372

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.04077    0.94630   1.100    0.276
rezago2$VIV_NEXC -0.01745    0.10167  -0.172    0.864

Residual standard error: 4.107 on 61 degrees of freedom
Multiple R-squared:  0.0004828,   Adjusted R-squared:  -0.0159
F-statistic: 0.02947 on 1 and 61 DF,  p-value: 0.8643
```

Porcentaje de viviendas que no cuenta con agua potable.

```
•       homocedasticidad.NAGU = lm(u1 ~ rezago2$VIV_NAGU)
•       summary(homocedasticidad.NAGU)

Call:
lm(formula = u1 ~ rezago2$VIV_NAGU)
Residuals:
Min      1Q  Median      3Q     Max
•       1.2113 -0.8203 -0.7228 -0.4681 31.2832

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.74497    0.84321   0.883    0.380
rezago2$VIV_NAGU  0.01449    0.06037   0.240    0.811

Residual standard error: 4.106 on 61 degrees of freedom
Multiple R-squared:  0.0009431,   Adjusted R-squared:  -0.01543
F-statistic: 0.05758 on 1 and 61 DF,  p-value: 0.8112
```

Porcentaje de viviendas que no cuentan con drenaje.

```
•       homocedasticidad.NDRE = lm(u1 ~ rezago2$VIV_NDRE)
•       summary(homocedasticidad.NDRE)

Call:
lm(formula = u1 ~ rezago2$VIV_NDRE)
Residuals:
```

```
Min      1Q  Median      3Q     Max
•       1.1106 -0.8518 -0.7573 -0.4178 31.3071

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.798164   0.892955   0.894    0.375
rezago2$VIV_NDRE 0.009928   0.067779   0.146    0.884

Residual standard error: 4.107 on 61 degrees of freedom
Multiple R-squared:  0.0003516,   Adjusted R-squared:  -0.01604
F-statistic: 0.02145 on 1 and 61 DF,  p-value: 0.884
```

Porcentaje de viviendas que no cuentan con electricidad.

```
•       homocedasticidad.NELE = lm(u1 ~ rezago2$VIV_NELE)
•       summary(homocedasticidad.NELE)

Call:
lm(formula = u1 ~ rezago2$VIV_NELE)
Residuals:
Min      1Q  Median      3Q     Max
•       1.2858 -0.9677 -0.7113 -0.2966 31.2171

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.3743     0.8885   1.547    0.127
rezago2$VIV_NELE  -0.1135     0.1749  -0.649    0.519

Residual standard error: 4.094 on 61 degrees of freedom
Multiple R-squared:  0.006858,    Adjusted R-squared:  -0.009423
F-statistic: 0.4212 on 1 and 61 DF,  p-value: 0.5188
```

De los resultados anteriores se puede observar que nuestra percepción erra errónea acerca de la falta de homocedasticidad, que en cada prueba el modelo no es significativo, lo que nos dice que hay evidencia para sostener que la varianza es constante.

Por otro lado se re ajusta un modelo para los residuos estandarizados, tomando en cuenta todas las variables explicativas se obtienen los siguientes resultados.

```
•       homocedasticidad = lm(u1 ~ rezago2$P6A14_NASI+rezago2$VIV_NEXC+rezago2$VI
V_NAGU+rezago2$VIV_NDRE+rezago2$VIV_NELE)
•       summary(homocedasticidad)

Call:
lm(formula = u1 ~ rezago2$P6A14_NASI + rezago2$VIV_NEXC + rezago2$VIV_NAGU +
rezago2$VIV_NDRE + rezago2$VIV_NELE)
Residuals:
Min      1Q  Median      3Q     Max
•       1.7266 -0.9711 -0.7045 -0.0996 30.9743

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.09556    2.03448   1.030    0.307
rezago2$P6A14_NASI -0.28148    0.48073  -0.586    0.561
rezago2$VIV_NEXC    0.02633    0.15470   0.170    0.865
```

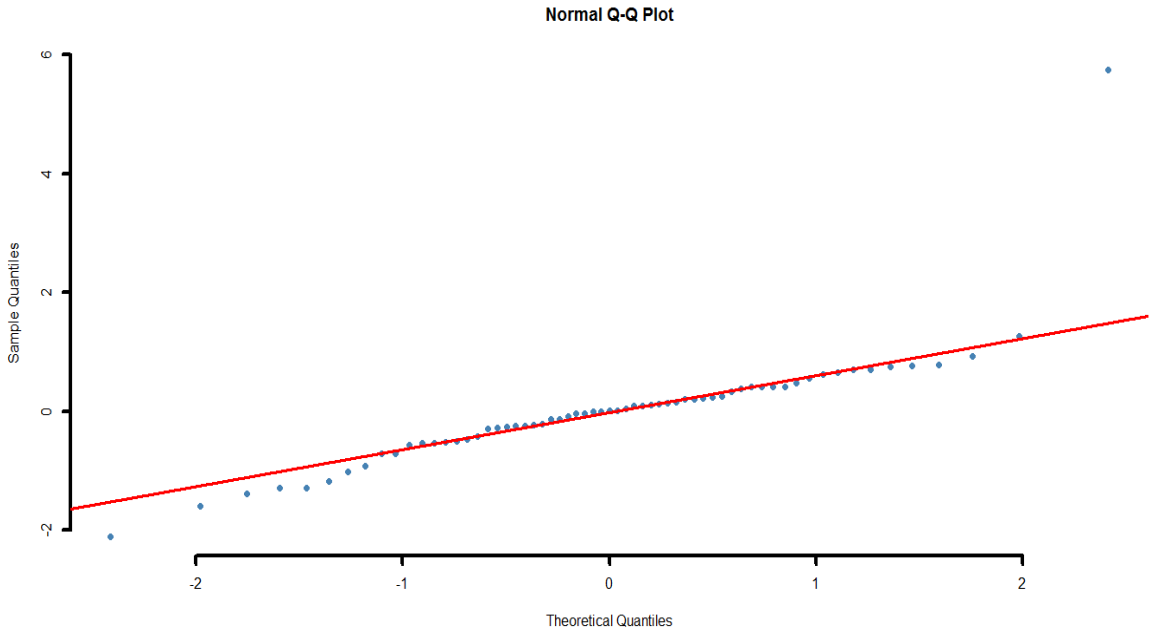
```
rezago2$VIV_NAGU    0.03687    0.09257   0.398    0.692
rezago2$VIV_NDRE    0.01717    0.10838   0.158    0.875
rezago2$VIV_NELE   -0.15321    0.23509  -0.652    0.517

Residual standard error: 4.214 on 57 degrees of freedom
Multiple R-squared:  0.01676,     Adjusted R-squared:  -0.06949
F-statistic: 0.1944 on 5 and 57 DF,  p-value: 0.9635
```

Nuevamente se puede observar, que el modelo no es significativo, por lo que hay evidencia para sostener que la varianza es constante, por lo que se puede concluir con esto, que en efecto, hay varianza constante.

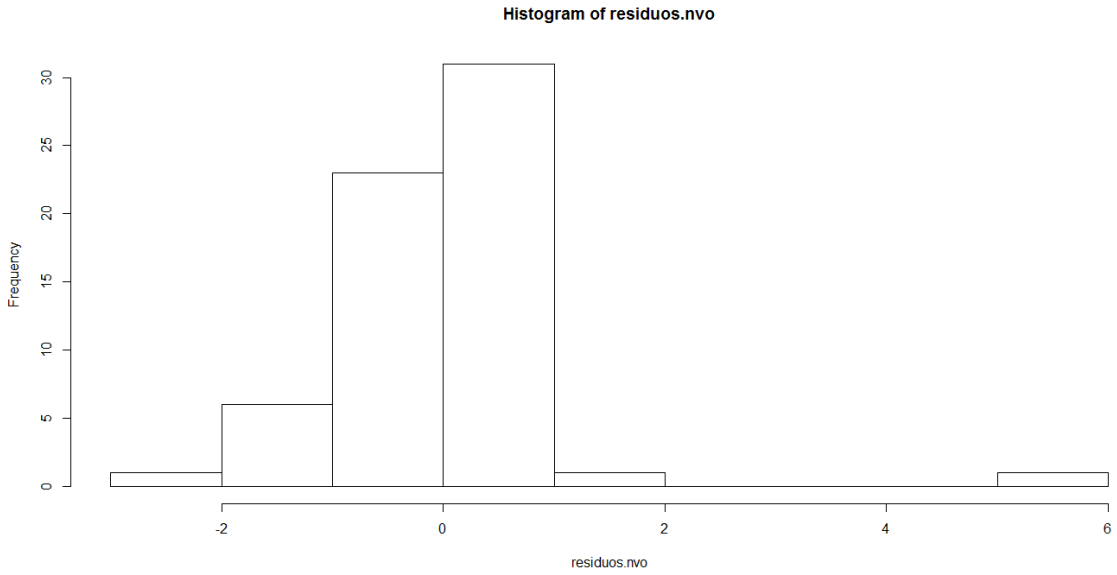
Ahora, dado que las pruebas resultaron negativas para no linealidad y heterocedasticidad, no es necesario aplicar alguna medida correctiva para modificar el modelo. Luego entonces, proseguimos a analizar los errores del mismo.

Análisis de normalidad de los errores.

En primer lugar mostramos el qqplot de los errores



Observemos que en general los cuantiles de los errores se asemejan bastante a la recta que representa los cuantiles de la distribución normal teórica. Sin embargo, se observa un punto atípico en la cola derecha.



Lo anterior se nota más claro en el histograma de los errores, donde se observa la carga en la cola derecha. Esto provoca además, que la distribución pierda simetría y no se asemeje a una distribución normal.

La prueba Anderson-Darling consiste en evaluar la bondad de ajuste de una distribución empírica con respecto a una distribución teórica. Esta prueba es sensible a las colas de la distribución, por lo que la usamos para evaluar la bondad de ajuste de los errores con respecto a la distribución normal. Recordemos que esto se hace sólo para fines prácticos, ya que no es del todo correcto utilizar una prueba de este estilo para evaluar la bondad de ajuste de los errores de una RLM.

```
> ad.test(residuos.nvo) #No se distribuyen normalmente!!!

        Anderson-Darling normality test

data:  residuos.nvo
A = 2.8167, p-value = 3.603e-07
```

Como era de esperarse, la prueba resulta significativa, por lo que se rechaza la hipótesis nula de que los errores se distribuyan normalmente.

Prueba de significancia del modelo.

Buscamos contrastar las hipótesis

$$H_0: \beta_i = 0 \ \forall i \quad vs \quad H_0: \beta_i \neq 0 \ \ para\ alguna\ i$$

$$i = 0, \dots, 5$$

Dado que los errores del modelo no se distribuyen de forma normal, no sabemos cuál es la distribución de los componentes de la tabla ANOVA. Por lo que en principio resulta imposible contrastar el estadístico F calculado con su respectiva distribución. Por ello, haremos uso de Bootstrap para, a través de 1000 simulaciones, aproximar dicha distribución. El código utilizado y la salida se muestran a continuación

```
> ######### Bootstrap para probar significancia ########
> set.seed(5)
> n <- nrow(rezago2)
> b0h <- mean(y1)
> res.esc <- (y1 - b0h)/sqrt(1-1/n)
>
> f.bs <- c()
> for(i in 1:1000){
+   ys <- b0h + sample(res.esc, n, T)
```

```
+   f.bs[i] <- anova(lm(ys ~ P6A14_NASI+VIV_NEXC+VIV_NAGU+VIV_NDRE+VIV_NELE, reza
go2))[1,'F value']
+ }
> quantile(f.bs, 0.95)
     95%
4.148935
> anova(modelo.nuevo)[1, 'F value']
[1] 25.7765
```

El cuantil 0.95 de la distribución simulada con bootstrap es 4.1489, mientras que el estadístico F obtenido de la tabla ANOVA (y que teóricamente se distribuye igual) es 25.77. Por lo tanto se rechaza H_0, lo que implica que alguna componente de $\boldsymbol{\beta}$ es distinta de 0.

Prueba de significancia para los componentes de β.

Una vez más, dado que los errores no se distribuyen de forma normal, calcularemos intervalos de confianza para cada componente de β simulando 1000 iteraciones con Bootstrap. El código utilizado, así como la salida del mismo, se muestran a continuación.

```
> ######### Bootstrap para IC ##########
> p <- 5
> k <- 1000
> b <- matrix(0,ncol=p+1,nrow=k)
>
> for (i in 1:k){
+   for(j in 1:(p+1)){
+     s <- sample(1:n, n, replace = T)
+     ys <- y1[s]
+     fit <- lm(ys ~ P6A14_NASI[s]+VIV_NEXC[s]+VIV_NAGU[s]+VIV_NDRE[s]+VIV_NELE[s
], rezago2)
+     b[i,j] <- coef(fit)[j]
+   }
+ }
>
> apply(b, 2, quantile, probs = c(0.05, 0.95))
        [,1]       [,2]         [,3]         [,4]        [,5]         [,6]
5%  5.197875 -0.2016321 -0.022880383 -0.003851301 -0.03276694 -0.002464072
95% 5.666209 -0.1129571  0.004922479  0.017121760 -0.01009170  0.044501956
```

Los intervalos de confianza para cada componente de β son

$$\beta_0: (5.1979\ , 5.6662)$$

$$\beta_1: (-0.2016\ , -0.1129)$$

$$\beta_2: (-0.0229\ , 0.0049)$$

$$\beta_3: (-0.0038\ , 0.0171)$$

$$\beta_4: (-0.0328\ , -0.0101)$$

$$\beta_5: (-0.0025\ , 0.0445)$$

La β_i cuyo intervalo de confianza contenga al 0, resulta ser no significativa para el modelo. Entonces, para β_2, β_3 y β_5 no hay evidencia suficiente para aseverar que son distintas de 0. A un nivel de confianza del 95%, el porcentaje de viviendas que no cuentan con excusado, el porcentaje de viviendas que no cuentan con agua potable y el porcentaje de viviendas que no cuentan con electricidad, no influyen en los resultados de **PBIPP**.

Cálculo de R^2 y R^2-ajustado.

Se muestra el código utilizado en R y la salida correspondiente

```
> ############ R y R2 #############
> I <- diag(n)
> J <- matrix(1, ncol=n, nrow=n)
> SC.reg <- t(y) %*% (H1 - (1/n)*J) %*% y
> SC.error <- t(y) %*% (I - H1) %*% y
> SC.TC <- t(y) %*% (I - (1/n)*J) %*% y
> CM.reg <- SC.reg/5
> CM.error <- SC.error/(n-6)
> CM.TC <- SC.TC/(n-1)
>
> R2 <- SC.reg/SC.TC; R2
          [,1]
[1,] 0.3718896
> R2.adj <- 1 - CM.error/CM.TC; R2.adj
          [,1]
[1,] 0.3167922
```

Observemos que tanto R^2 como R^2-ajustado son muy bajos; entre 0.3 y 0.4. Es decir, el modelo sólo explica entre el 30% y 40% de la variabilidad total de los datos, lo que resulta escaso.