Regresión múltiple y otras técnicas multivariadas

Proyecto Final

Rivera Torres Francisco de Jesús
Rodríguez Maya Jorge Daniel
Samayoa Donado Víctor Augusto
Trujillo Barrios Georgina

Junio 04, 2019

Introducción

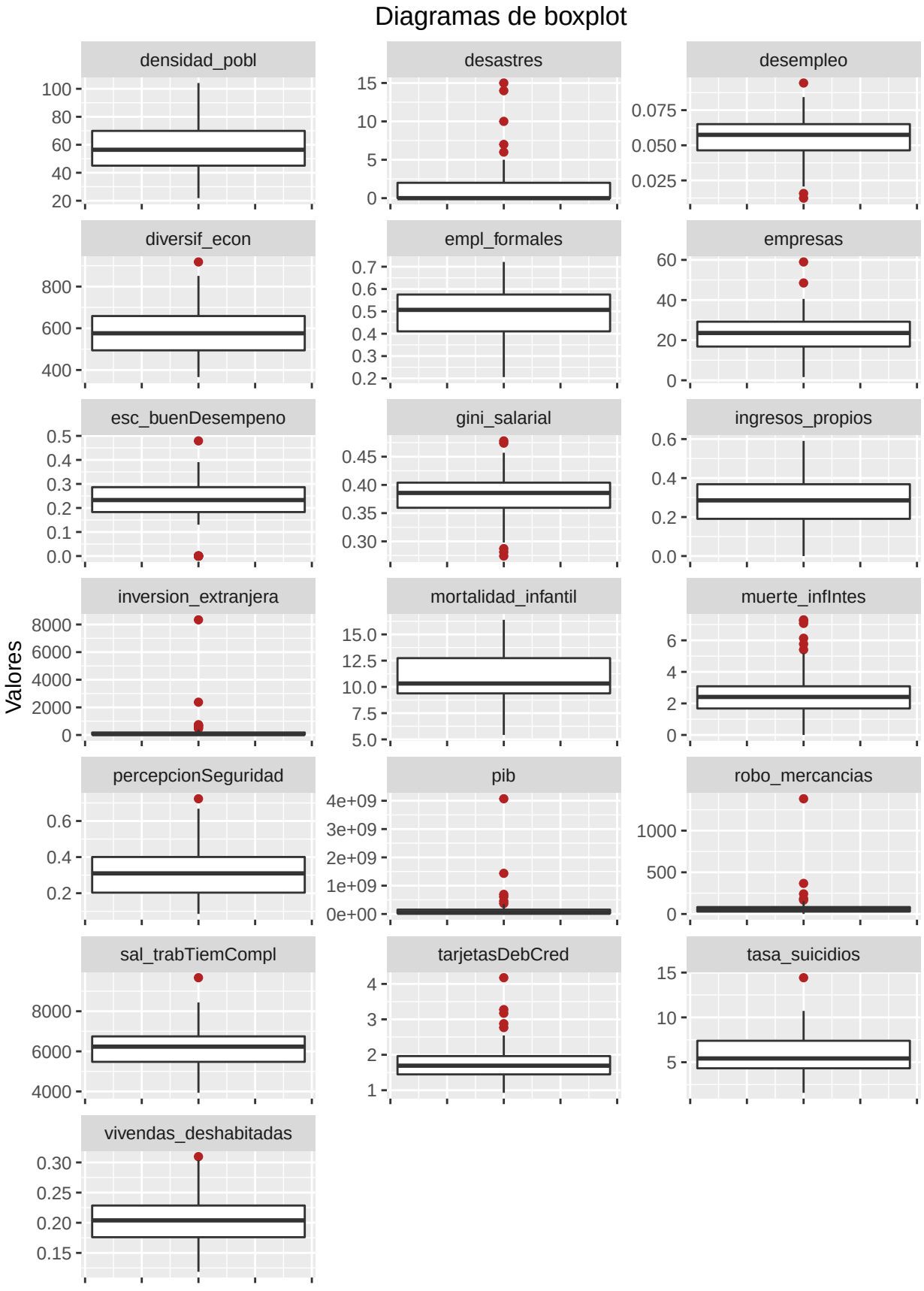
Aqui ponemos la motivación...

La variable objetivo es inversión extranjera

Análisis exploratorio

A continuación se realiza el análisis exploratorio de los datos para observar y comprender de mejor forma el comportamiento de estos, así como para poder visualizar datos a típicos o que sean influyentes. Para ello comenzaremos por realizar las gráficas box-plot las cuales nos permiten distinguir los rangos y las distribuciones de cada una de las variables.

Generamos las gráficas boxplot para comprender los rangos y distribuciones de cada una de las variables:


Diagramas de boxplot
Valores
densidad_pobl
desastres
desempleo
diversif_econ
empl_formales
empresas
esc_buenDesempeno
gini_salarial
ingresos_propios
inversion_extranjera
mortalidad_infantil
muerte_infIntes
percepcionSeguridad
pib
robo_mercancias
sal_trabTiemCompl
tarjetasDebCred
tasa_suicidios
vivendas_deshabitadas

Con estas gráficas podemos apreciar que ciertas variables parecen tener una distribución bastante simétrica, sin embargo en algunos casos se muestran variables que presentan un notable sesgo a la derecha y algunos outliers. Los casos más notables de estas son:

- Desastres
- Inversión extranjera
- Muertes de infantes
- PIB
- Robo de mercancias
- Tarjetas de débito y crédito

La variable desastres agrupa la mayor parte de los datos en un rango de cero a cinco y presenta algunos valores extremos que llegan al rango de 15.

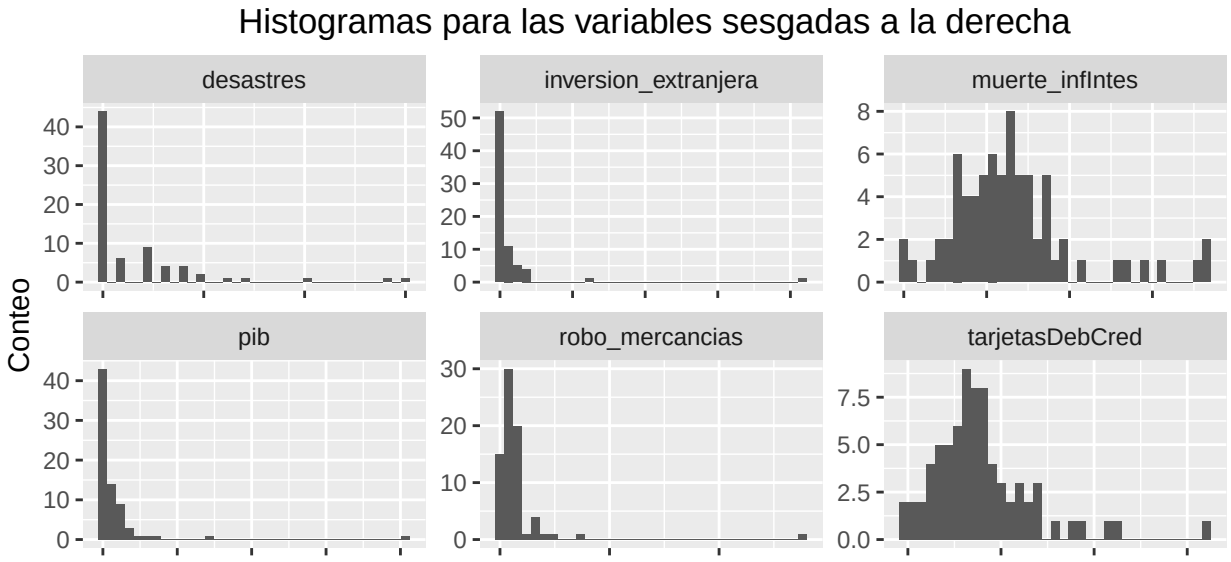
Inversión extranjera, es una de las variables que presentan un sesgo notorio, donde los datos están en su mayor parte concentrados alrededor del cero.

Muerte de infantes, PIB, robo de mercancías y tarjetas de débito y crédito, son otras de las variables que presentan un sesgo y en todos los casos también presentan valores extremos.

Hay algunas otras variables que aparentemente tienen un comportamiento simétrico, pero que también presentan valores extremos que deben tomarse en cuenta, como son `diversif_econ` o `esc_buenDesempeno`.

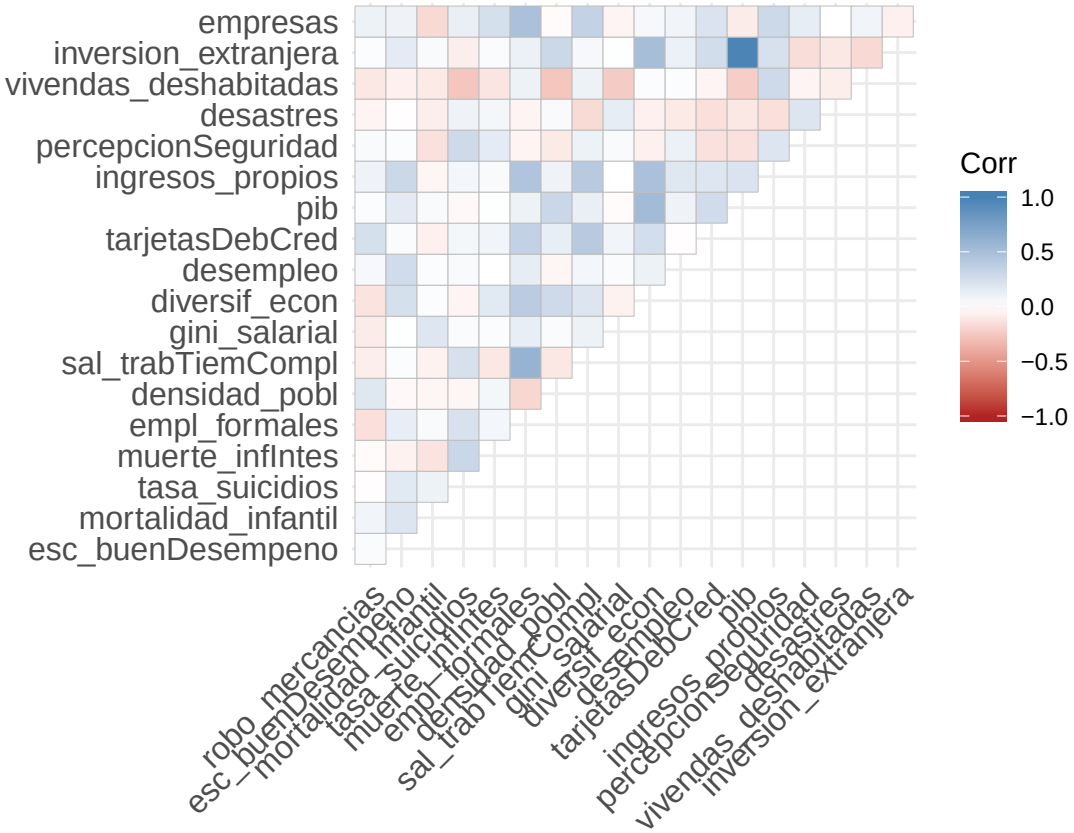
Como se está haciendo el análisis general de los datos para tratar de explicar la inversión extranjera en términos de las otras variables, el análisis exploratorio de las boxplot, nos muestra, como era de esperarse, que algunas variables tendrán un mayor efecto en el comportamiento de nuestra variable objetivo, pero también nos alerta de dar un tratamiento adecuado para evitar que las variables influyentes que se observan puedan sesgar los resultados, esto se tomará en cuenta más adelante en el tratamiento de los datos.

Para visibilizar mejor el comportamiento de las variables que presentan un sesgo elaboramos sus histogramas.



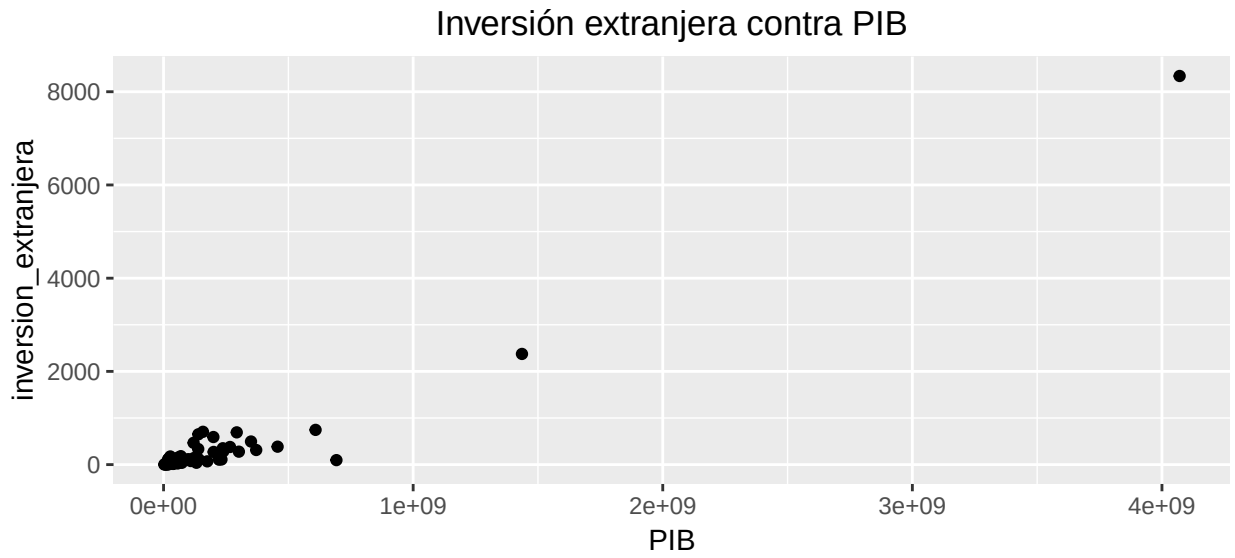
En efecto, en los histogramas anteriores, podemos apreciar que hay un sesgo positivo en estas variables.

Ahora analizaremos las correlaciones entre los pares de variables.



En la gráfica anterior se aprecia que las variables con más alta correlación son el PIB con inversion_extranjera, por lo que resulta conveniente hacer un diagrama de dispersión para

estas variables para estudiar su comportamiento.



Podemos notar que la inversión extranjera se concentra en torno del origen, salvo para dos valores extremos, después de revisar los datos encontramos que estos valores corresponden a la Ciudad de México y a la ciudad de Monterrey, son valores extremos o palanca que influyen en el comportamiento de los datos, haciendo de esta manera que los datos estén mayormente correlacionados de forma positiva.

Selección de modelos

Una vez se tiene un panorama del comportamiento de los datos, procedemos a generar modelos que nos ayuden a explicar la **inversion extranjera**.

Modelo 1

Ajustamos un modelo utilizando todas las variables para tener una punto de referencia del nivel de significancia del modelo y cuál es la máxima varianza que se puede explicar.

```
##
## Call:
## lm(formula = inversion_extranjera ~ ., data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -544.06  -97.84   16.37  105.67  321.38
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.271e+00  3.913e+02   0.006 0.995390
## robo_mercancias         6.056e-03  1.646e-01   0.037 0.970788
## esc_buenDesempeno      -1.997e+02  2.614e+02  -0.764 0.448321
## mortalidad_infantil    -1.475e+00  1.122e+01  -0.131 0.895869
## tasa_suicidios         -1.488e+01  1.223e+01  -1.217 0.228888
## muerte_infIntes         1.886e+01  1.843e+01   1.023 0.310747
## empl_formales           6.592e+02  3.520e+02   1.873 0.066416 .
## densidad_pobl           3.277e-01  1.660e+00   0.197 0.844220
## sal_trabTiemCompl      -1.266e-01  3.274e-02  -3.867 0.000293 ***
## gini_salarial           9.593e+02  6.377e+02   1.504 0.138231
## diversif_econ          -5.591e-01  3.079e-01  -1.816 0.074841 .
## desempleo               1.102e+03  1.587e+03   0.694 0.490509
## tarjetasDebCred         1.329e+01  5.312e+01   0.250 0.803320
## pib                     2.018e-06  5.831e-08  34.601  < 2e-16 ***
## ingresos_propios        5.005e+02  2.885e+02   1.735 0.088382 .
## percepcionSeguridad    -8.258e+01  1.818e+02  -0.454 0.651512
## desastres              -2.035e+00  8.542e+00  -0.238 0.812639
## vivendas_deshabitadas   9.149e+02  7.033e+02   1.301 0.198768
## empresas                1.278e+00  3.187e+00   0.401 0.690026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 192.8 on 55 degrees of freedom
## Multiple R-squared:  0.9721, Adjusted R-squared:  0.9629
## F-statistic: 106.4 on 18 and 55 DF,  p-value: < 2.2e-16
```

Modelo 2

Debido a que son muchas variables, decidimos utilizar la selección en ambas direcciones para seleccionar un subconjunto de variables para el modelo.

```
##
## Call:
## lm(formula = inversion_extranjera ~ ., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1092.65   -45.87    20.47    88.63   469.92
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.538e+02  2.985e+02  -0.515    0.608
## pib                  1.963e-06  6.227e-08  31.524   <2e-16 ***
```

```
## empl_formales        3.736e+01  2.813e+02   0.133    0.895
## ingresos_propios     4.426e+02  2.725e+02   1.624    0.109
## tarjetasDebCred     -3.767e+01  5.219e+01  -0.722    0.473
## percepcionSeguridad -2.658e+02  1.832e+02  -1.451    0.152
## diversif_econ       -3.028e-01  2.964e-01  -1.021    0.311
## gini_salarial        6.997e+02  6.468e+02   1.082    0.283
## desastres            2.449e+00  9.311e+00   0.263    0.793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 218.4 on 65 degrees of freedom
## Multiple R-squared:  0.9577, Adjusted R-squared:  0.9524
## F-statistic: 183.8 on 8 and 65 DF,  p-value: < 2.2e-16
```

Validación de supuestos y acciones correctivas

En esta sección se procede a validar los supuestos correspondientes a los modelos de regresión lineal múltiple (RLM).

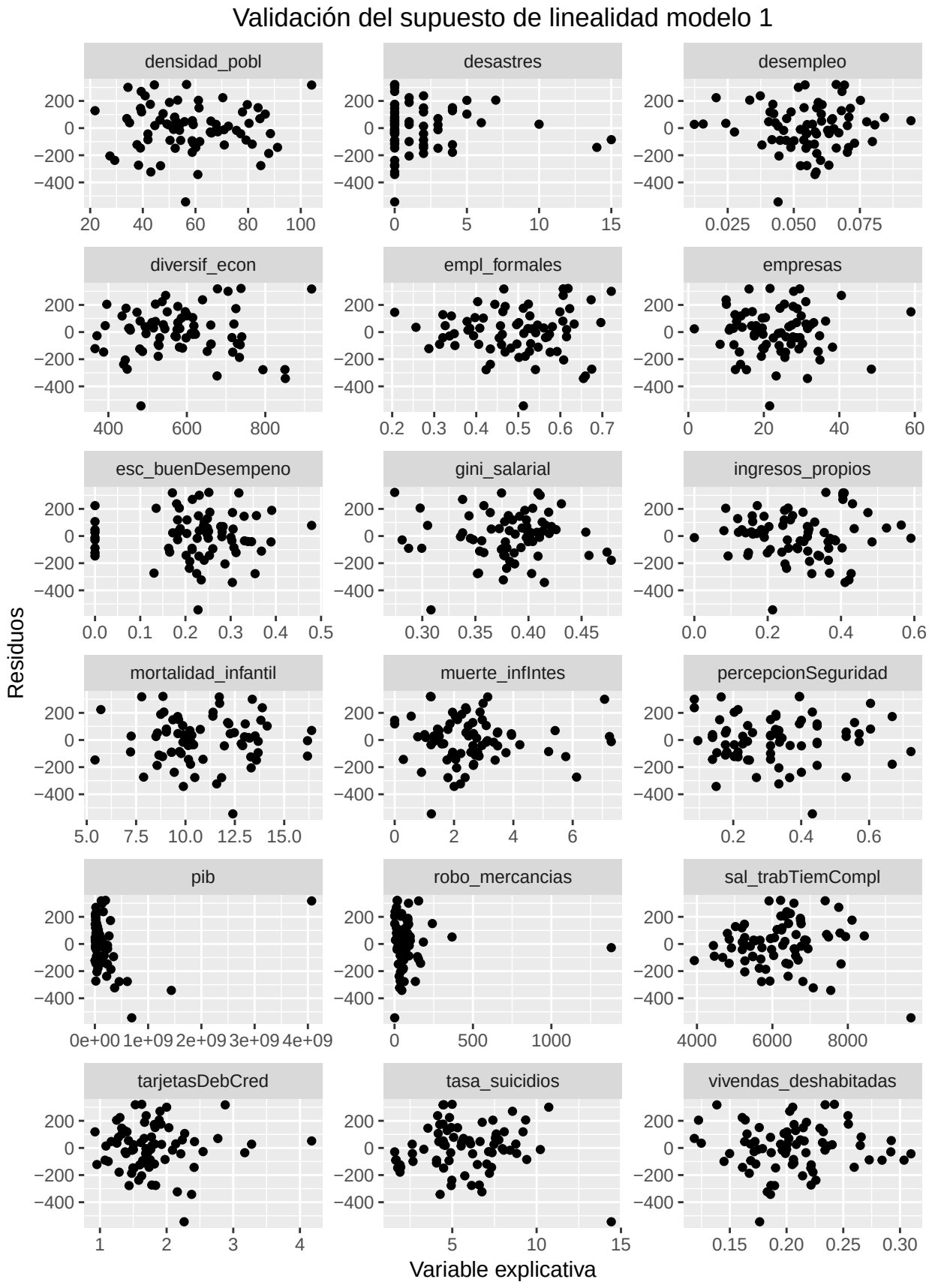
Linealidad

Sabemos que uno de los supuestos de los modelos de RLM es que la variable a explicar tenga una distribución normal alderedor del valor esperado por el modelo ajustado.

$$\mathrm{E}(Y_i|\mathrm{x}_i) = \mathrm{x}_i^T\beta, \qquad i = 1, \ldots, n$$

Modelos 1

Para revisar si se cumple el supuesto de linealidad, se procede a graficar los residuos del modelo respecto a cada una de las variables explicativas.


Validación del supuesto de linealidad modelo 1
densidad_pobl
desastres
desempleo
diversif_econ
empl_formales
empresas
esc_buenDesempeno
gini_salarial
ingresos_propios
mortalidad_infantil
muerte_infIntes
percepcionSeguridad
pib
robo_mercancias
sal_trabTiemCompl
tarjetasDebCred
tasa_suicidios
vivendas_deshabitadas
Residuos
Variable explicativa

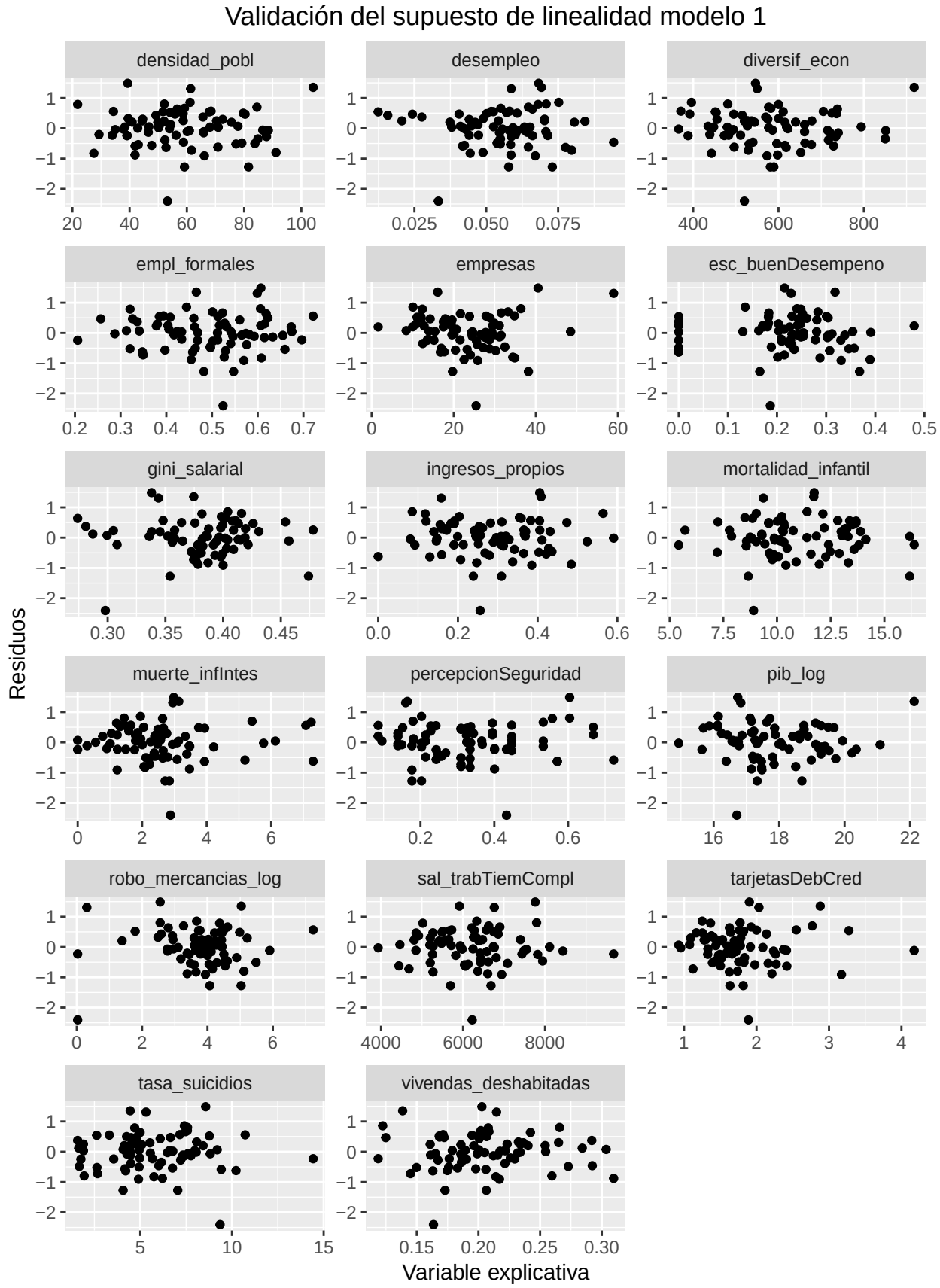
En la gráfica anterior se observa que las variables `pib` y `reobo_mrecancia` contienes datos atípicos, lo que no permite apreciar de forma adecuada si hay o no un patrón en las gráficas de residuos contra dichas variables.

Respecto a la variable desastres, se determina eliminarla. Esto debido a que todos son valores enteros y no preserva la linealidad.

Se procede a aplicar la transformación logaritmo a las variables explicativa `pib` y `robo_mercancias` y se ajusta nuevamente el modelo con todas las variables.

```
##
## Call:
## lm(formula = log_inversion_extranjera ~ ., data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4057 -0.3304  0.0254  0.4141  1.4869
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.202e+01  2.155e+00  -5.575 7.38e-07 ***
## esc_buenDesempeno     -9.101e-01  9.770e-01  -0.932   0.3556
## mortalidad_infantil   -3.671e-02  4.065e-02  -0.903   0.3704
## tasa_suicidios        -6.687e-03  4.922e-02  -0.136   0.8924
## muerte_infIntes       -5.602e-02  6.992e-02  -0.801   0.4264
## empl_formales          3.146e+00  1.322e+00   2.381   0.0207 *
## densidad_pobl         -6.970e-04  6.153e-03  -0.113   0.9102
## sal_trabTiemCompl     -8.040e-05  1.299e-04  -0.619   0.5384
## gini_salarial          4.313e+00  2.310e+00   1.867   0.0671 .
## diversif_econ          2.083e-03  1.376e-03   1.514   0.1358
## desempleo              3.430e+00  5.907e+00   0.581   0.5638
## tarjetasDebCred       -3.965e-01  1.842e-01  -2.153   0.0357 *
## ingresos_propios       1.880e+00  1.011e+00   1.859   0.0683 .
## percepcionSeguridad   -9.310e-01  6.517e-01  -1.429   0.1587
## vivendas_deshabitadas  1.352e+00  2.573e+00   0.525   0.6013
## empresas               9.754e-04  1.250e-02   0.078   0.9381
## pib_log                7.367e-01  1.216e-01   6.058 1.22e-07 ***
## robo_mercancias_log   -8.897e-03  9.051e-02  -0.098   0.9221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7113 on 56 degrees of freedom
## Multiple R-squared:  0.8422, Adjusted R-squared:  0.7943
## F-statistic: 17.58 on 17 and 56 DF,  p-value: < 2.2e-16
```

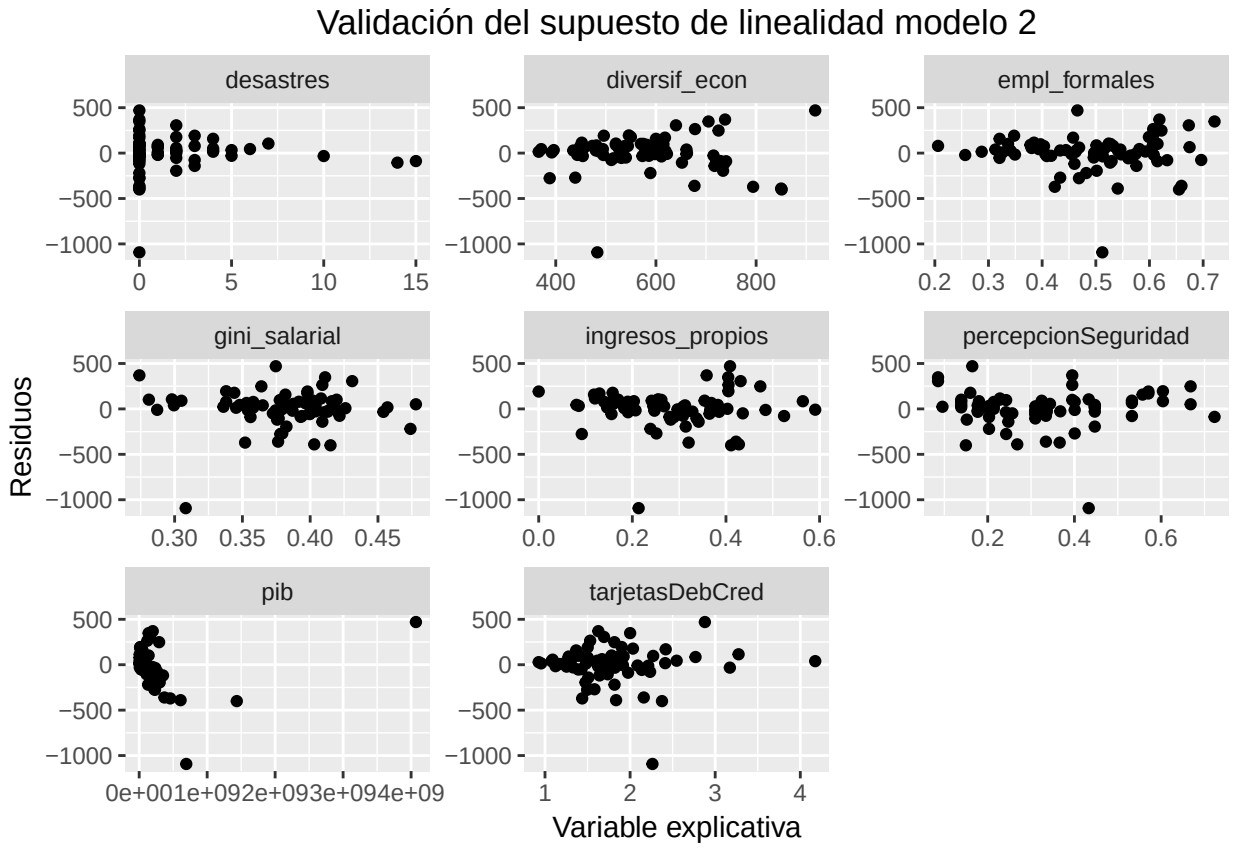
Se procede a analizar nuevamente el comportamiento del residuo para este nuevo ajuste respecto a cada una de las variables explicativas.


Validación del supuesto de linealidad modelo 1
densidad_pobl
desempleo
diversif_econ
empl_formales
empresas
esc_buenDesempeno
gini_salarial
ingresos_propios
mortalidad_infantil
muerte_infIntes
percepcionSeguridad
pib_log
robo_mercancias_log
sal_trabTiemCompl
tarjetasDebCred
tasa_suicidios
vivendas_deshabitadas
Residuos
Variable explicativa

Como puede apreciarse en el gráfico anterior los residuos no tienen patrones respecto a las variables explicativas. Por lo que concluimos que las transformaciones realizadas tanto en la varaible objetivo como en las variables explicativas permiten cumplir el supuesto de linealidad.

Modelo 2

Para revisar si se cumple el supuesto de linealidad, se procede a graficar los residuos del modelo respecto a cada una de las variables explicativas.

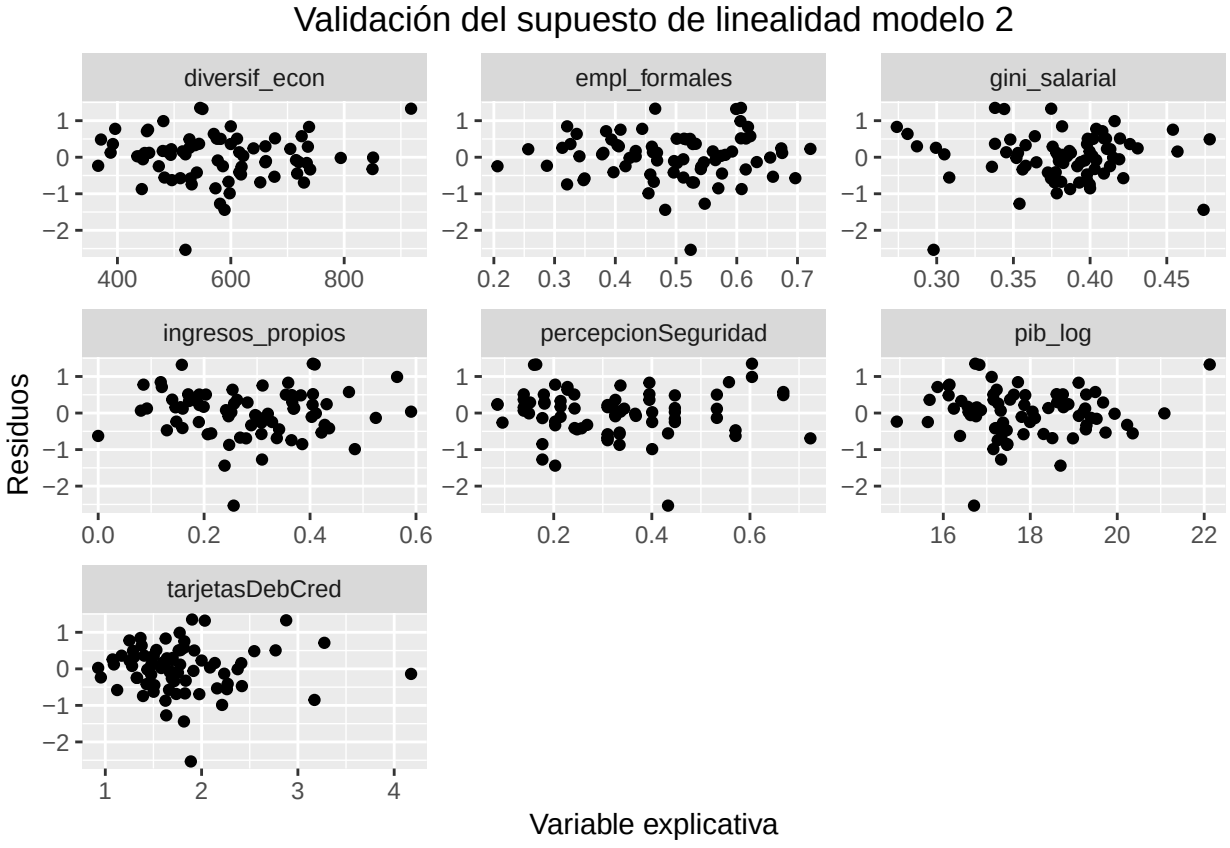


En este caso se procederá de manera análoga al modelo anterior eliminando la variable desastres y aplicando la transformación de logaritmo tanto a `inversión_extranjera` y a `pib`.

```
##
## Call:
## lm(formula = log_inversion_extranjera ~ ., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.53369 -0.39198  0.05138  0.36589  1.35029
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -11.743450   1.437769  -8.168 1.33e-11 ***
## empl_formales         3.014006   0.887685   3.395  0.00116 **
## ingresos_propios      1.852521   0.832966   2.224  0.02957 *
## tarjetasDebCred      -0.442560   0.158254  -2.797  0.00676 **
## percepcionSeguridad  -1.028288   0.552803  -1.860  0.06732 .
## diversif_econ         0.002022   0.001139   1.775  0.08059 .
## gini_salarial         3.415627   1.990182   1.716  0.09081 .
## pib_log               0.704582   0.094528   7.454 2.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6792 on 66 degrees of freedom
## Multiple R-squared:  0.8305, Adjusted R-squared:  0.8125
## F-statistic: 46.19 on 7 and 66 DF,  p-value: < 2.2e-16
```

Se procede a analizar nuevamente el comportamiento del residuo para este nuevo ajuste respecto a cada una de las variables explicativas.



Como puede apreciarse en el gráfico anterior los residuos no tienen patrones respecto a las variables explicativas. Por lo que concluimos que las transformaciones realizadas tanto en la varaible objetivo como en las variables explicativas permiten cumplir el supuesto de linealidad.

Homocedasticidad

Para este supuesto debemos revisar que la varianza es constante

$$\mathrm{V}(Y_i|\mathrm{x}_i) = \sigma^2, \qquad i = 1, \ldots, n$$

Modelo 1

Modelo 2

No correlación

Generamos el diagrama de dispersión

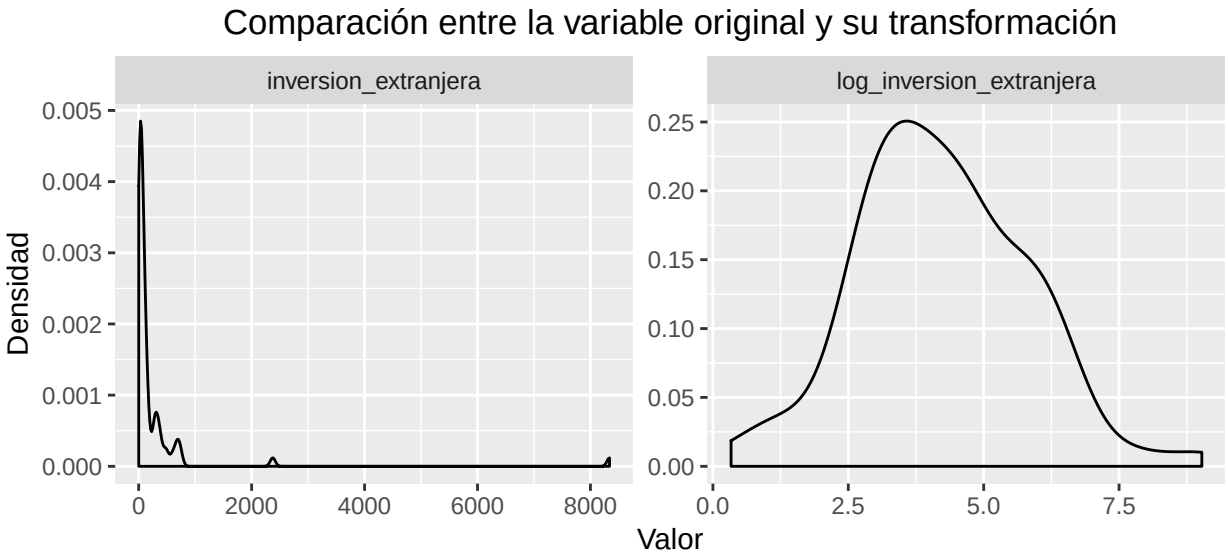
$$\mathrm{Cor}(Y_i, Y_j|\mathrm{x}_i, \mathrm{x}_j) = 0, \qquad i, j = 1, \ldots, n, \qquad i \neq j$$

Modelo 1

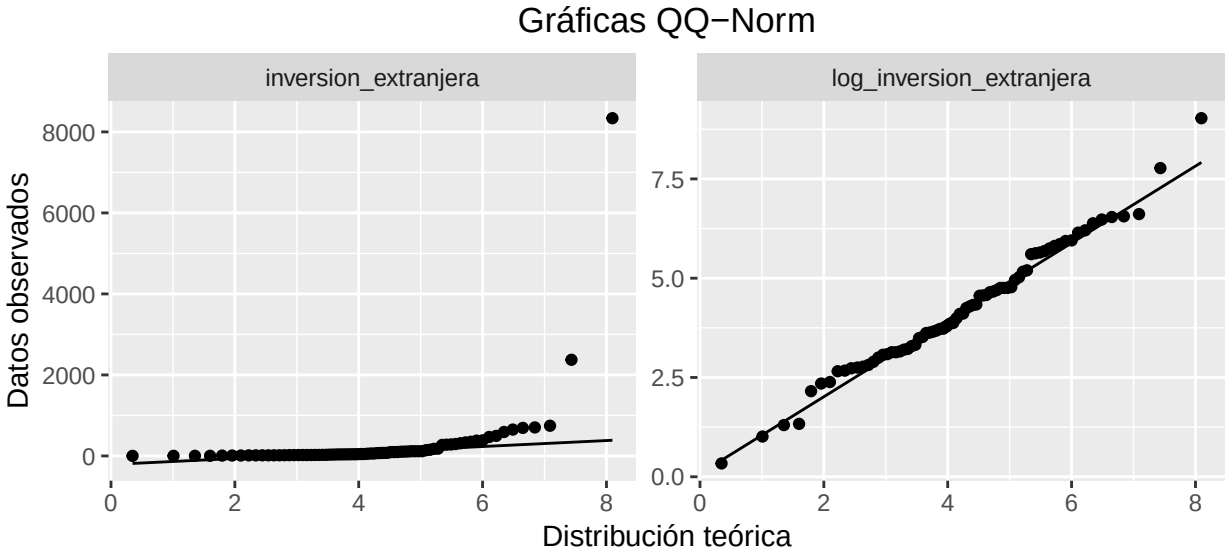
Modelo 2

Supuesto de normalidad

En ambos modelos, la variable a explicar es la inversión extranjera, por lo que se procede acomparar la distribución de dicha varaible y su transformación `log(inversion extranjera)`. Debido a que ya habíamos observado que dicha varaible tenía un sesgo hacia la derecha.



En la gráfica anteriro se puede apreciar que al aplicarle logaritmo a la variable se obtiene una forma más simétrica. Por lo que procedemos a realizar la comparación en la gráfica QQ-norm para determianr qué tato se aproxima a una distrbución normal.

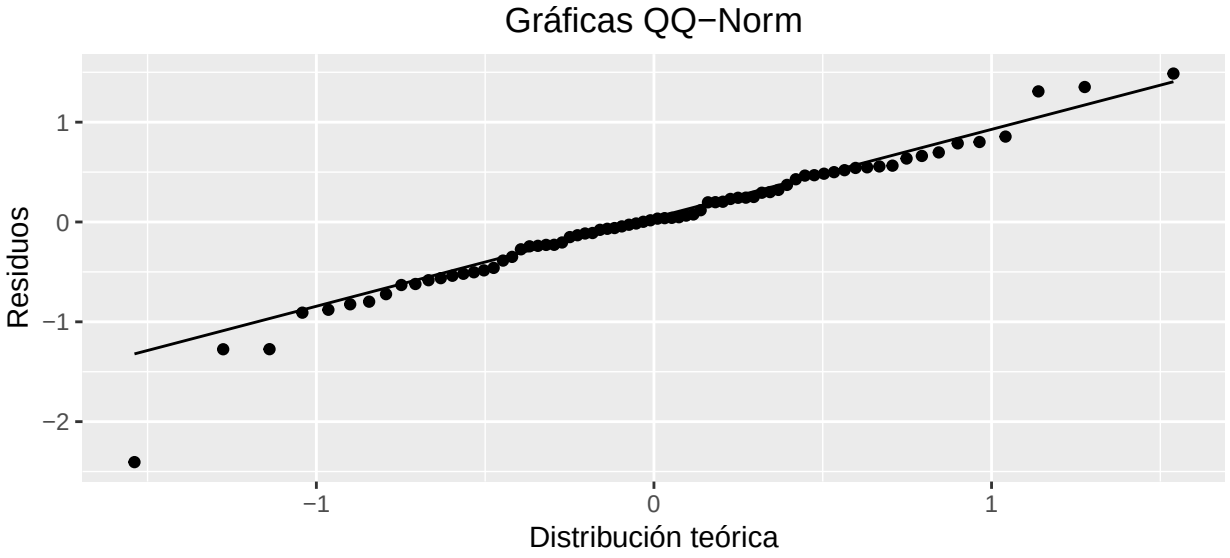


En este caso se logra apreciar que los datos transformados del logaritmo se ajustan mejor a una distribución normal con media $\mu = 4.222006$ y desviación estandar $\sigma = 1.5684134$.

Por tal motivo confirmamos que es necesario usar la transformación logaritmo de la varaible inversión extranjera.

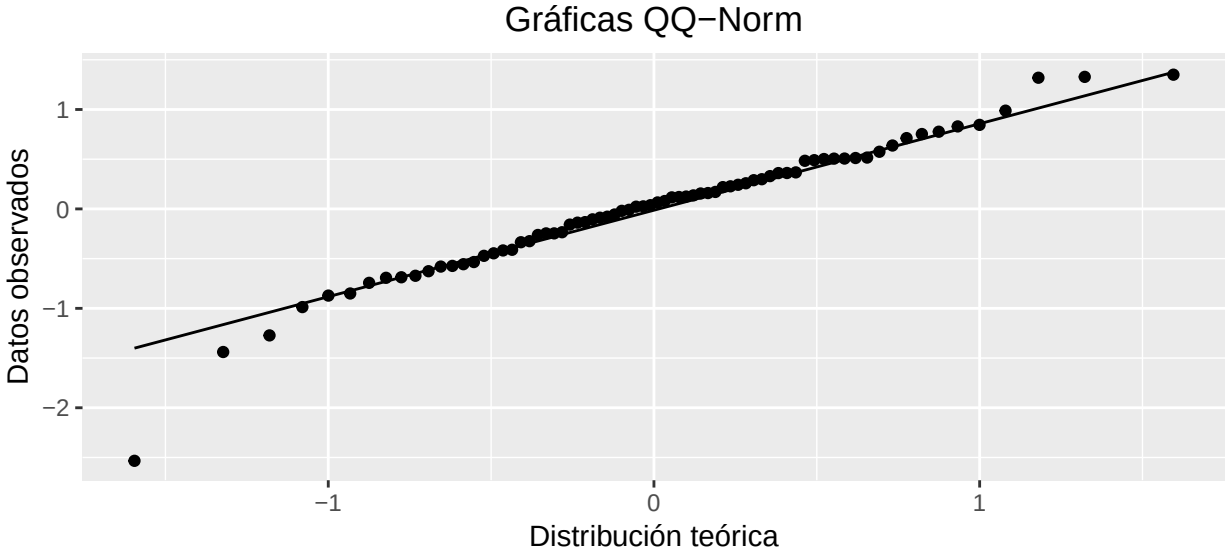
Modelo 1

Ahora procedemos a revisar la normalidad de los residuos para el primer modelo.



Modelo 2

Ahora procedemos a revisar la normalidad de los residuos para el primer modelo.



Análisis de multicolinealidad

Con base en el análisis exploratorio de los datos, sabemos que las correlaciones por pares entre las variables explicativas es baja. Sin embargo, es util obtener el índice de condición κ para tener un criterio cuantitativo sobre el nivel de colinealidad que existe entre las variables explicativas.

Modelo 1

En este caso se procede a calcular la κ para la matríz de variables explicativas:

- esc_buenDesempeno.
- mortalidad_infantil.
- tasa_suicidios.
- muerte_infIntes.
- empl_formales.
- densidad_pobl.
- sal_trabTiemCompl.
- gini_salarial.
- diversif_econ.
- desempleo.
- tarjetasDebCred.
- ingresos_propios.
- percepcionSeguridad.

- vivendas_deshabitadas.
- empresas.
- log_pib.
- log_robo_mercancias.

obteniendo así un valor de $\kappa_1 = 3.639e + 05$.

Modelo 2

En este caso se procede a calcular la κ para la matríz de variables explicativas:

- empl_formales.
- ingresos_propios.
- tarjetasDebCred.
- percepcionSeguridad.
- diversif_econ.
- gini_salarial.
- log_inversion_extranjera.
- pib_log

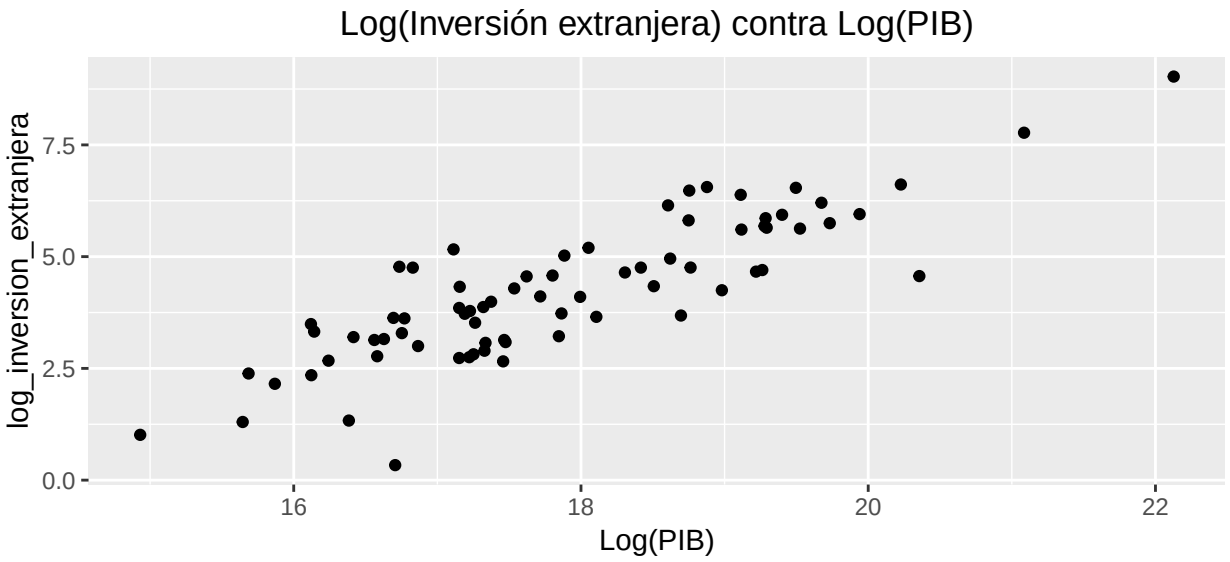
obteniendo así un valor de $\kappa_2 = 9,842$.

Análisis de observaciones atípicas y observaciones influyentes

A continuación procedemos a analizar los datos atípicos y las observaciones influyentes que observamos en los datos tanto en el análsisi exploratorio como en el análisis de validación de supuestos.

Modelo 1

En este modelo notamos que variables como el PIB tenían datos influyentes, principalemte el datos más extremo asociado al valle de México. Por lo que se decidió aplicar una transformación logaritmo para minizar dicho efecto y no tener que eliminarlos del análisis.



Modelo 2

En este modelo también se tiene la variable PIB con explicativa, por lo que se procede a realizar la misma tranformación al PIB y se vuelve a ajustar el modelo.

Análisis del modelo final

De acuerdo con el análisis previo, se decidió elegir el modelo 2. A partir del cual podemos realizar las siguientes interpretaciones.

El interceptó es $\beta_0 = -11.7434501$ quiere decir que si las demás variables tomarán un valor de cero la inversión extranjera sería negativa lo cual no es interpetable pues la inversión sólo puede ser mayor o igual a cero.

Por parte de los coeficientes

- $\beta_1 = 3.0140057$ la inversión extranjera crece 0.000001963 unidades por cada unidad que aumenta el pib.
- $\beta_2 = 1.8525206$ la inversión extranjera crece 37.36 unidades por cada unidad que aumentan los empleos formales.
- $\beta_3 = -0.4425601$ la inversión extranjera crece 442.6 unidades por cada unidad que aumentan los ingresos propios.
- $\beta_4 = -1.0282884$ la inversión extranjera decrece 37.67 unidades por cada unidad que aumentan las tarjetas de débito credito.
- $\beta_5 = 0.0020216$ la inversión extranjera decrece 265.8 unidades por cada unidad que aumentan la percepción de la seguridad.
- $\beta_6 = 3.4156272$ la inversión extranjera decrece .3028 unidades por cada unidad que aumentan la diversificación economica.

- $\beta_7 = 0.7045821$ la inversión extranjera crece 699.7 unidades por cada unidad que aumentan los gini salarial.
- $\beta_8 = NA$ la inversión extranjera crece 2.449 unidades por cada unidad que aumentan los desastres.