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

La inversión extranjera suele impactar positivamente el crecimiento económico y, en general, tiene consecuencias deseables para una economía: fuente de generación de empleo o transferencias tecnológicas, por ejemplo. Los análisis sobre el crecimiento económico suelen comparar resultados estatales y en este sentido ignoran la heterogeneidad estatal. Por ello, con un objetivo exploratorio, proponemos un análisis centrado en ciudades que permita dar cuenta de cuáles son los factores que explican la inversión extranjera (eligiendo de entre aquellas variables que típicamente se utilizan en el análisis del desarrollo económico) y, aún más, de cómo estos inciden en ella.

Para este análisis usamos la base de datos de que el IMCO genera con el propósito de calcular un índice de competitividad de ciudades. Nuestro objetivo, más allá de realizar una clasificación, es encontrar los factores determinantes que explican la inversión extranjera y, por ello, consideramos conveniente centrarnos en las ciudades más importantes del país: en el entendido, de que son las ciudades las que típicamente impulsan el crecimiento de un estado y de que la inversión puede verse condicionada por características locales más que por estadísticas agregadas estatales (las cuales si bien pueden influir, intuimos que no son más fuertes que las condiciones estrictamente locales). Con este impulso, hicimos una preselección intuitiva de variables que pensamos podrían influir en la decisión de invertir:

- percepción de seguridad de las ciudades, medida como porcentaje de ciudadanos que reportan sentirse seguros,
- ingresos propios del gobierno (como una proxy de qué tan bueno es el gobierno),
- monto reportado en robo de mercancías en la ciudad (medida como pesos por cada millón de pesos de PIB),
- mortalidad infantil (decesos de menores de un año por cada mil nacidos vivos),
- tasa de suicidios (por cada 100 mil habitantes),
- viviendas deshabitadas (como porcentaje de las viviendas),

- escuelas de calidad (porcentaje de escuelas con desempeño bueno o excelente en prueba PLANEA),
- muertes por infecciones intestinales (por cada 100 mil habitantes),
- empleados en el sector formal (como porcentaje de la población ocupada),
- diversificación económica (número de sectores económicos presentes),
- salario mensual para trabajadores de tiempo completo (pesos corrientes),
- desigualdad salarial (coeficiente de Gini salarial),
- densidad poblacional (personas por hectárea),
- desempleo (porcentaje de la PEA),
- número de tarjetas de débito y crédito por cada adulto (proxy de servicios financieros)
- pib (Miles de pesos 2014)

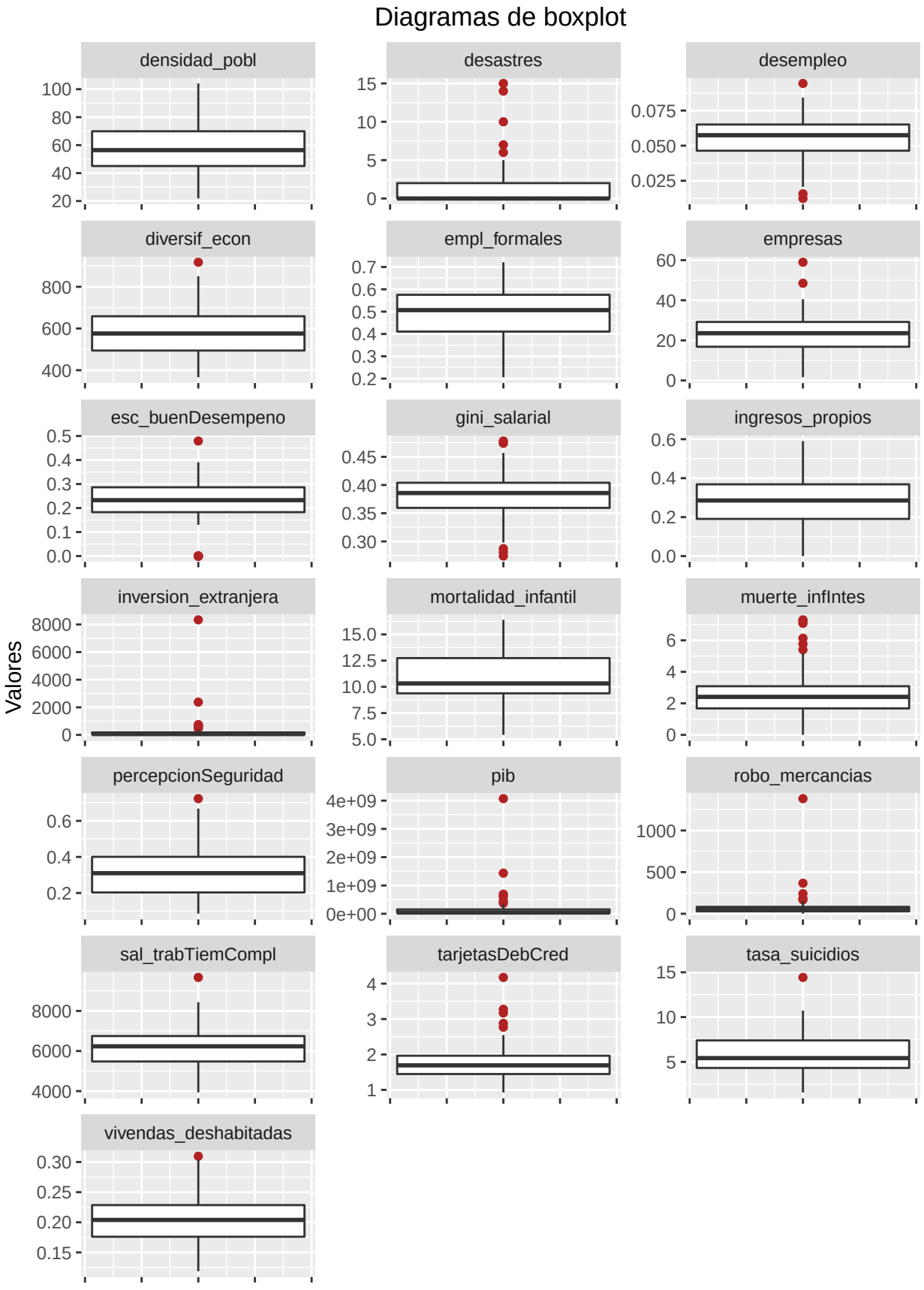
- Número de declaratorias de desastre en los últimos tres años

- inversion extranjera (dólares per cápita en promedio 3 años)
- número de empresas por cada mil de PEA

Análisis exploratorio

En esta sección se procederá a realizar el análisis exploratorio con el objetivo de tener una mejor entendimiento sobre el comportamiento de los datos.

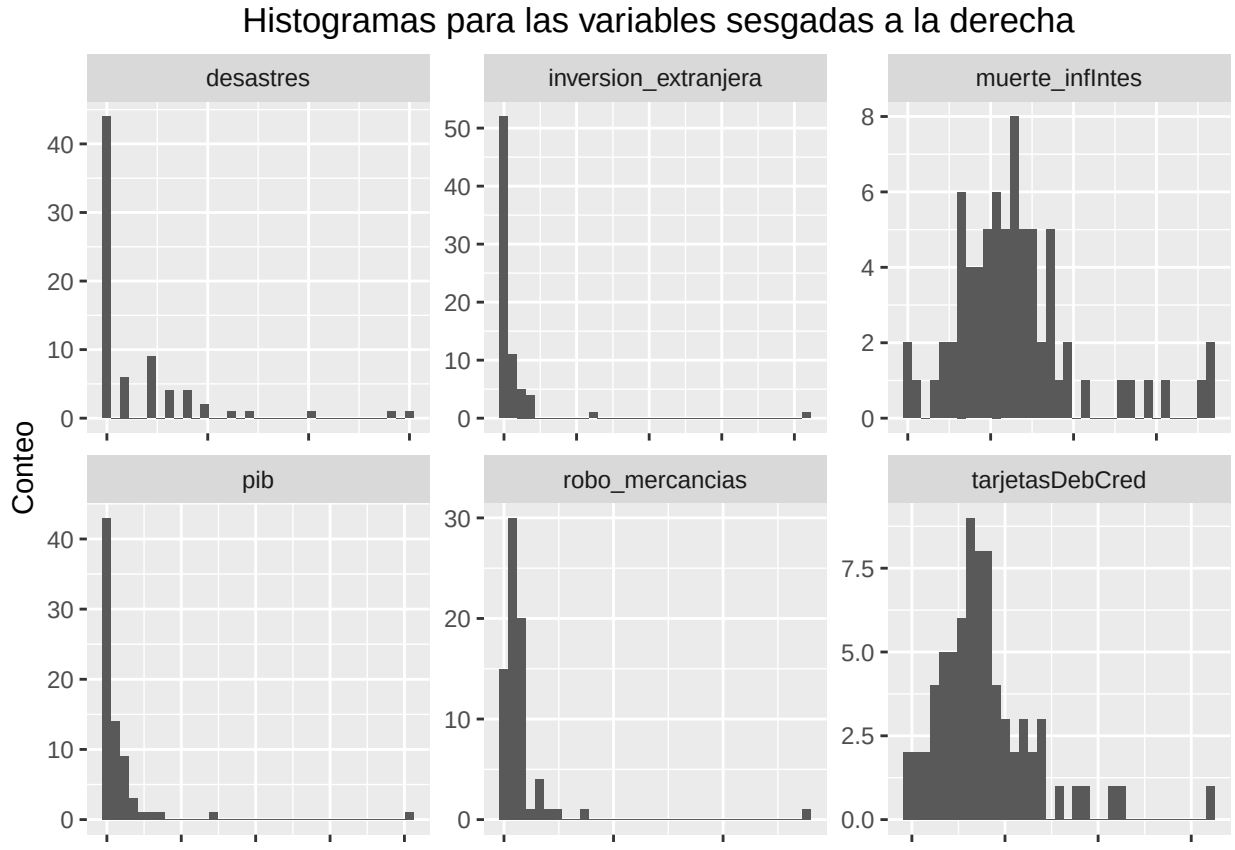
Generamos las gráficas boxplot para comprender los rangos y distribuciones de cada una de las variables:


Diagramas de boxplot
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
Valores

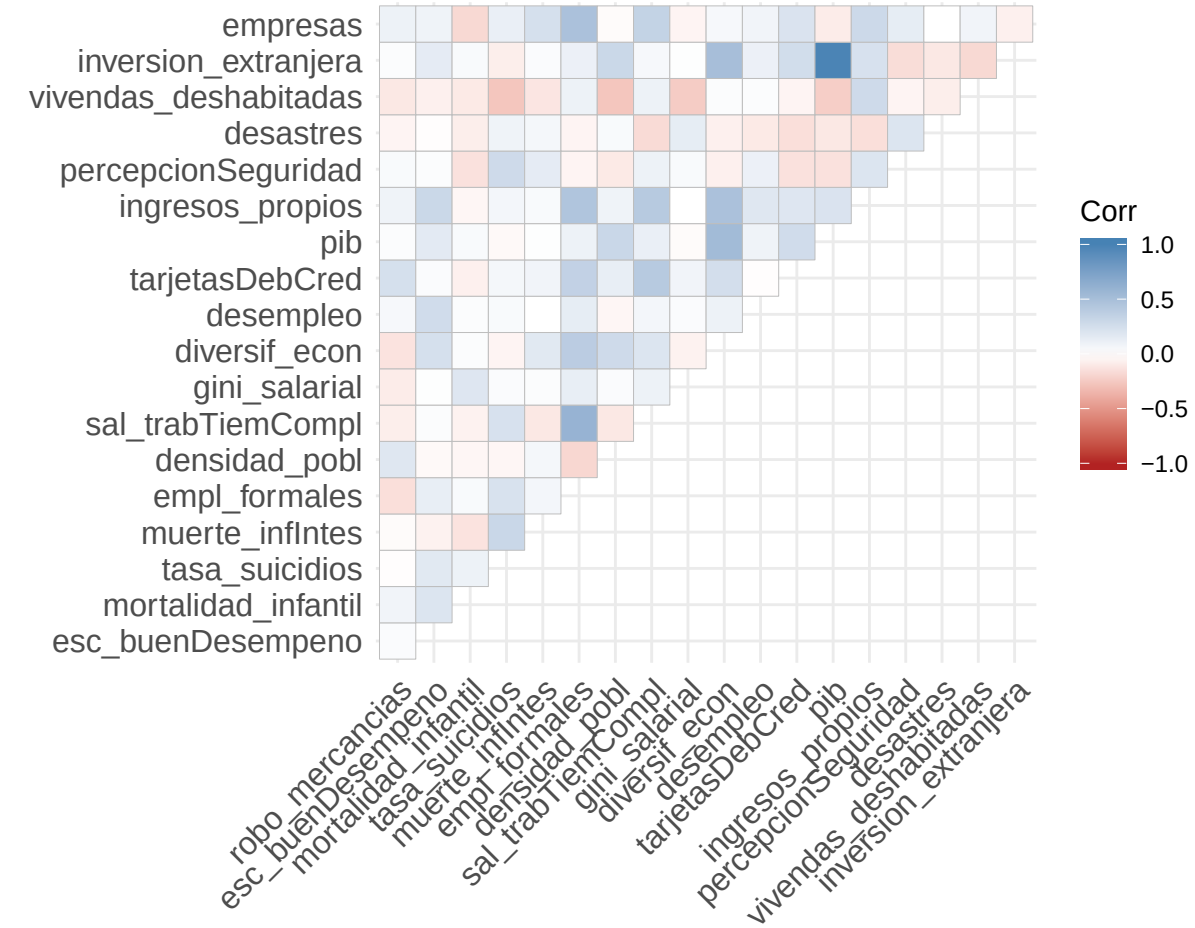
Con la gráfica anterior logramos apreciar que muchas variables parecen tener una distribución simétrica. Sin embargo, existen otras variables que tienen un sesgo a la derecha:

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

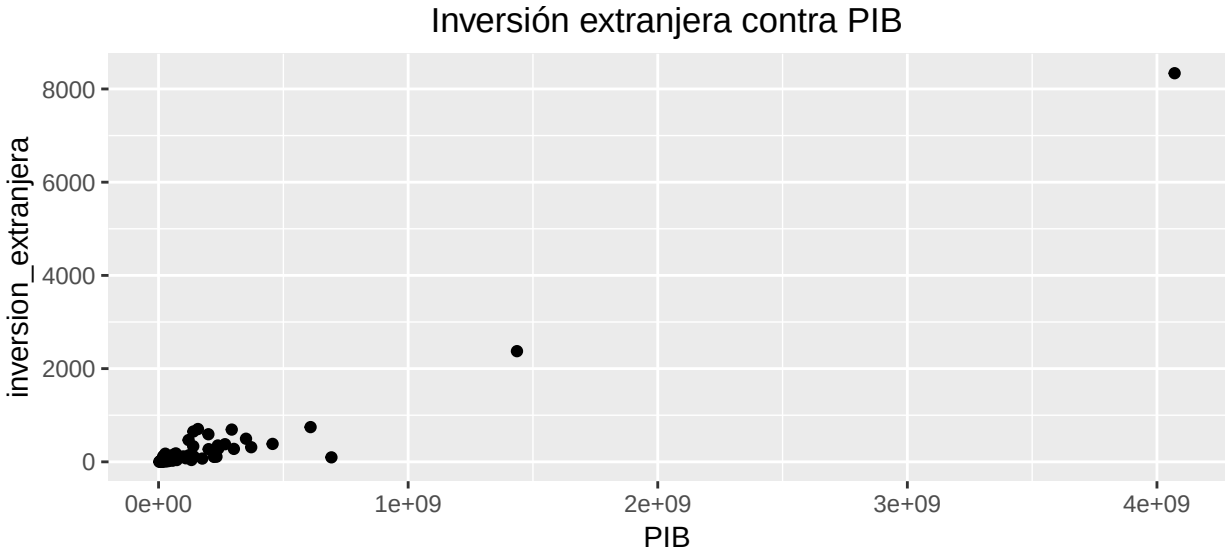
Para estas variables elaboraremos sus histográmas para entender en mejor manera su comprotamiento:



Procedemos a analizar las correlaciones entre los pares de variables:



En la gráfica anterior se puede apreciar que las variables que más alta correlación tienen son el `pib` con `inversion_extranjera`, por lo que procedemos a realizar un diagrama de dispersión para estas dos variables.



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 ~ pib + empl_formales + ingresos_propios +
##     tarjetasDebCred + percepcionSeguridad + diversif_econ + gini_salarial +
##     desastres, 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

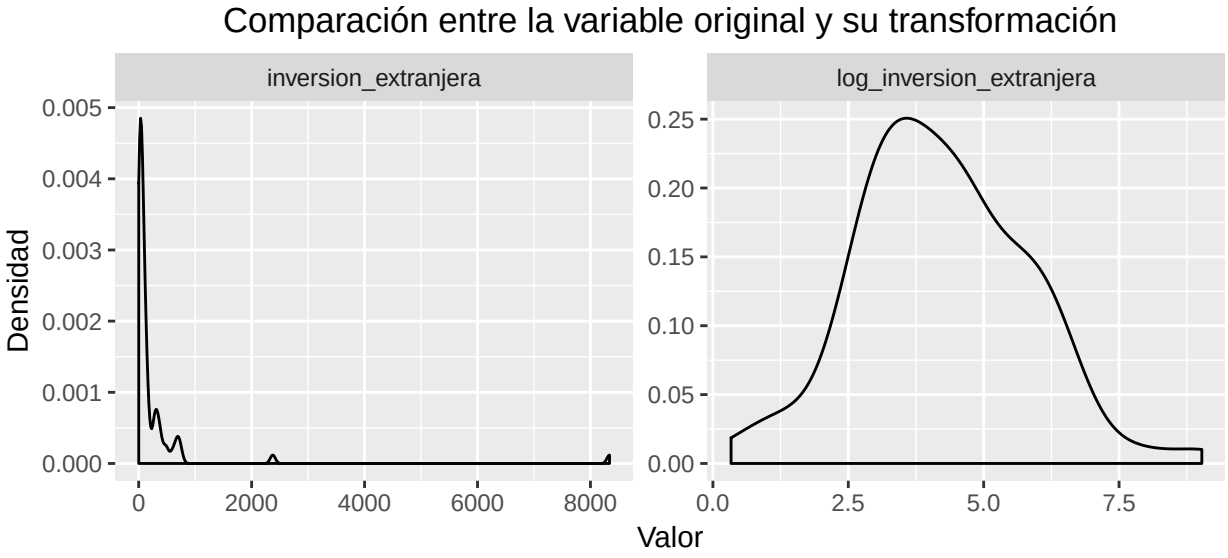
Sabemos que uno de los supuestos de los modelos de RLM es que la variable a explicar tenga una distribución normal alderedor del valor esperado por el modelo ajustado.

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

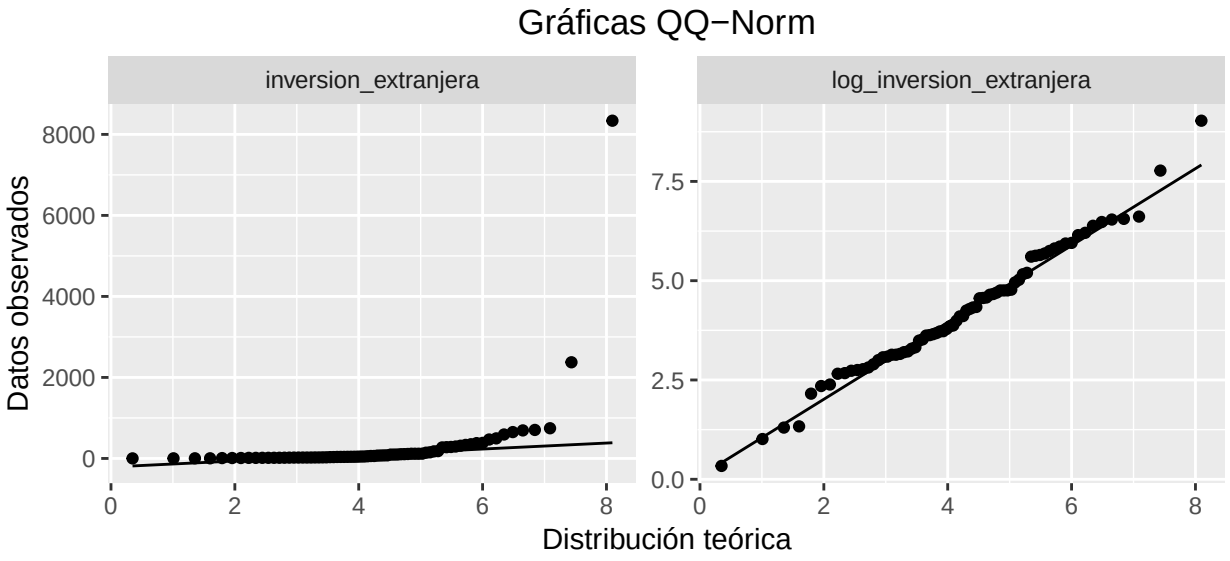
Modelos 1 y 2

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.

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```



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 $\sigma = 1.5684134$.

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

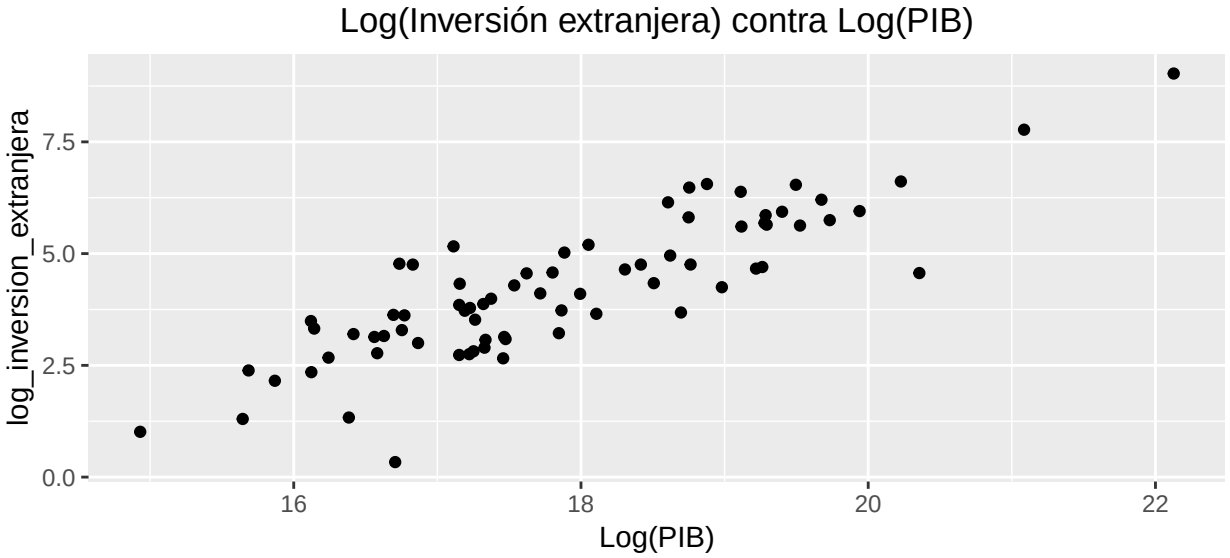
Modelo 2

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