Problema 1 Proyecto Análisis Multivariado

Nombres Diego Aravena
Alonso Campos

Problema 1

Datos de Naciones Unidas, UN. Los datos en el archivo `UNData`, en Canvas contiene algunas variables, entre ellas `ppgdp`, el producto nacional bruto por persona de 2009 en dólares estadounidenses, `fertility`, la tasa de natalidad por cada 1000 mujeres en la población en el año 2009, `lifeExp`, esperanza de vida y `pctUrban` porcentaje de población urbana; además de dos variables cualitativas, `region` y `grupo`. Los datos son para 199 localidades, en su mayoría países miembros de la ONU, pero también otras áreas como Hong Kong que no son países independientes.

(a) Haga un análisis exploratorio de los datos. Escriba sus conclusiones.

Solución

En primer lugar se realiza un análisis descriptivo de los datos, es decir, se obtienen medidas de resumen de cada variable numérica.

Tabla 1: Resumen descriptivo de las variables `fertilty, ppdgp, lifeExpF` y `pctUrban`

	Media	Desviación Estándar	Mediana	Mínimo	Máximo
`fertility`	2.76	1.34	2.26	1.13	6.92
`ppgdp`	13011.95	18412.44	4684.50	114.80	105095.40
`lifeExpF`	72.29	10.12	75.89	48.11	87.12
`pctUrban`	57.93	23.43	59.00	11.00	100.00

Se puede observar en la Tabla 1 que las magnitudes de cada variable son diferentes, pues la variable asociada al Producto Nacional Bruto es significativamente mayor que las demás. Por el contrario la variable `fertility` tiene valores menores en comparaciones

con otras variables. Y las restantes, `lifeExpF` y `pctUrban`, poseen magnitudes similares, no obstante, `pctUrban` tiene un rango (diferencia entre valor máximo y mínimo) mayor.

Boxplot

En segundo lugar, para conocer la naturaleza de los datos se realizarán box plots para cada variable numérica. Con el objetivo de para comparar los valores de diferentes grupos y regiones

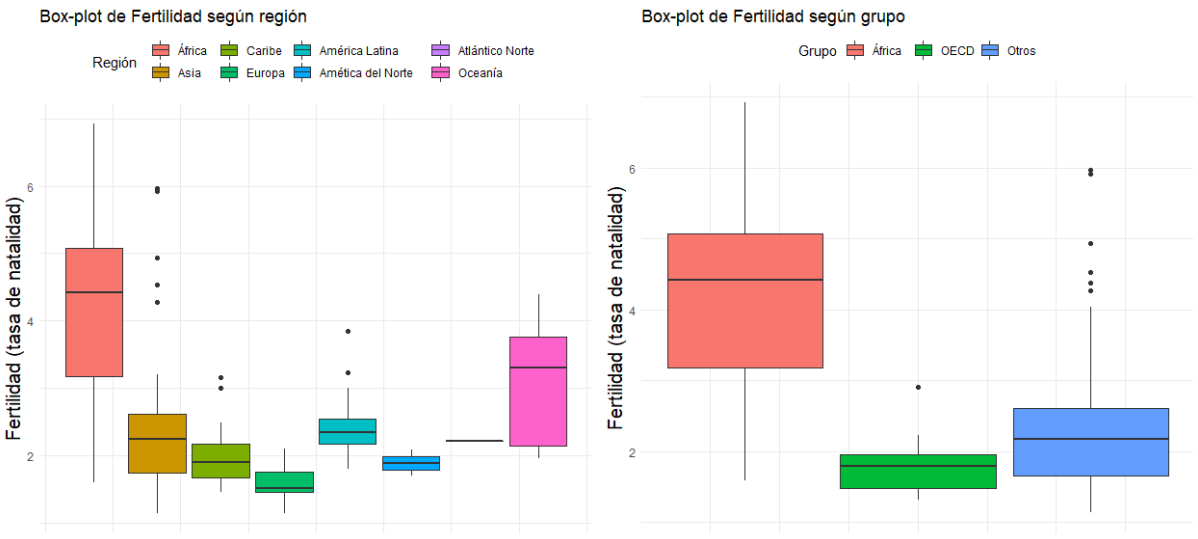

Figura 1: Box plot variable fertilidad según región y grupo

Se puede apreciar en la Figura 1 que existen diferencias entre grupos y regiones. En Asia y Caribe se observa un comportamiento similar en la variable fertilidad. Sin embargo, el comportamiento en general es diferente para categoría de los gráficos. También se realizará el mismo procedimiento para las demás variables.

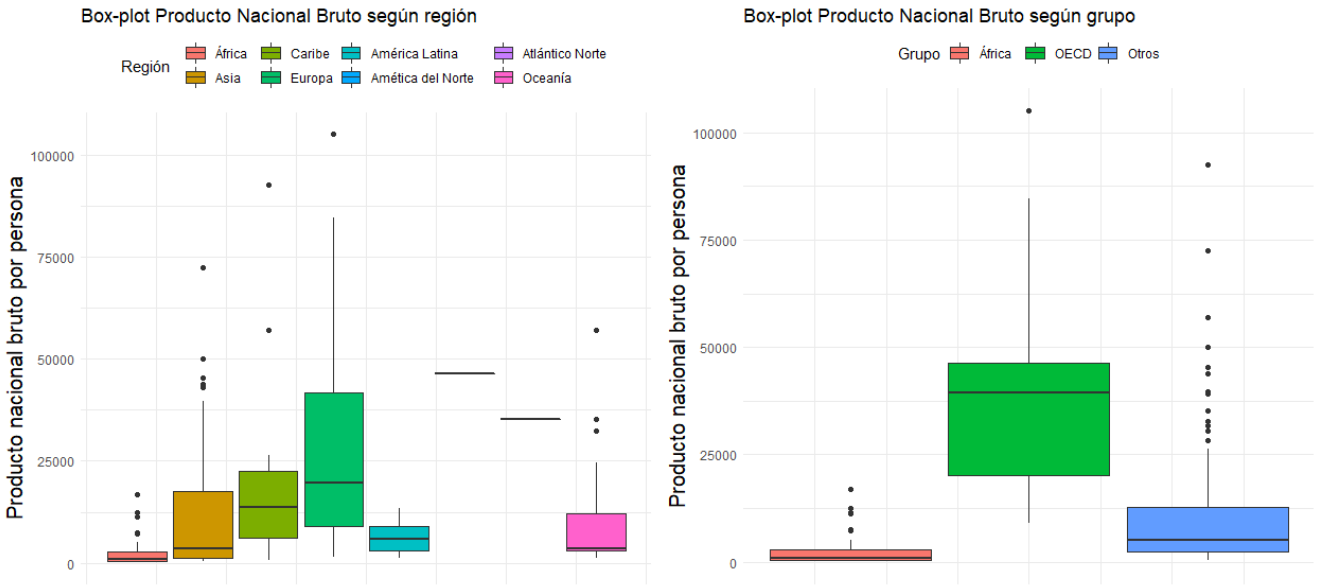

Figura 2: Box plot Producto nacional bruto según región y grupo

En el caso de los box plot de la Figura 2 asociados a la variable `ppgdp`. Se muestran comportamientos similares entre Asia, América del Norte y Oceanía, pues éstos tienden a acercarse a cero y poseen una baja variabilidad. En el caso de los grupos, África y otros países son también cercanos a cero los gráficos.

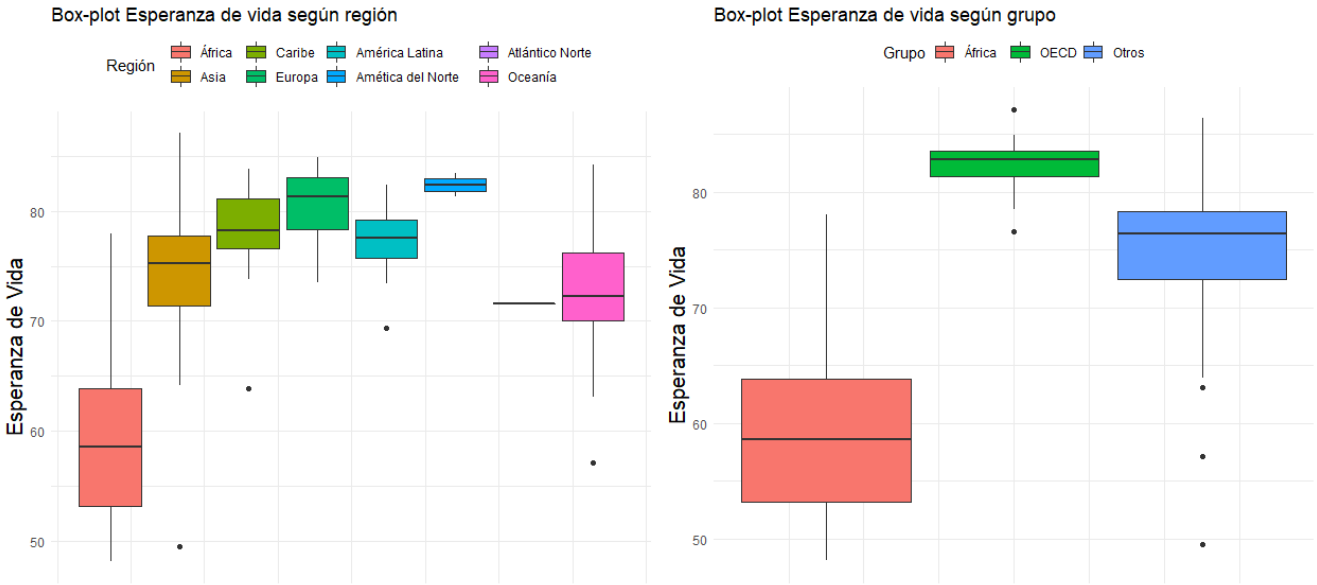

Figura 3: Box plot Esperanza de vida bruto según región y grupo

La Figura 3 es posible evidenciar que las regiones tienen comportamientos diferentes, en particular, África tiene datos que son menores al resto de las observación que hacen que el gráfico correspondiente esté ubicado más abajo. Se pueden obtener las mismas

conclusiones a partir de la variable grupo. Sumado al comentario anterior de la situación de África, se puede apreciar que los países pertenecientes a la OECD, poseen una mayor tasa de natalidad.

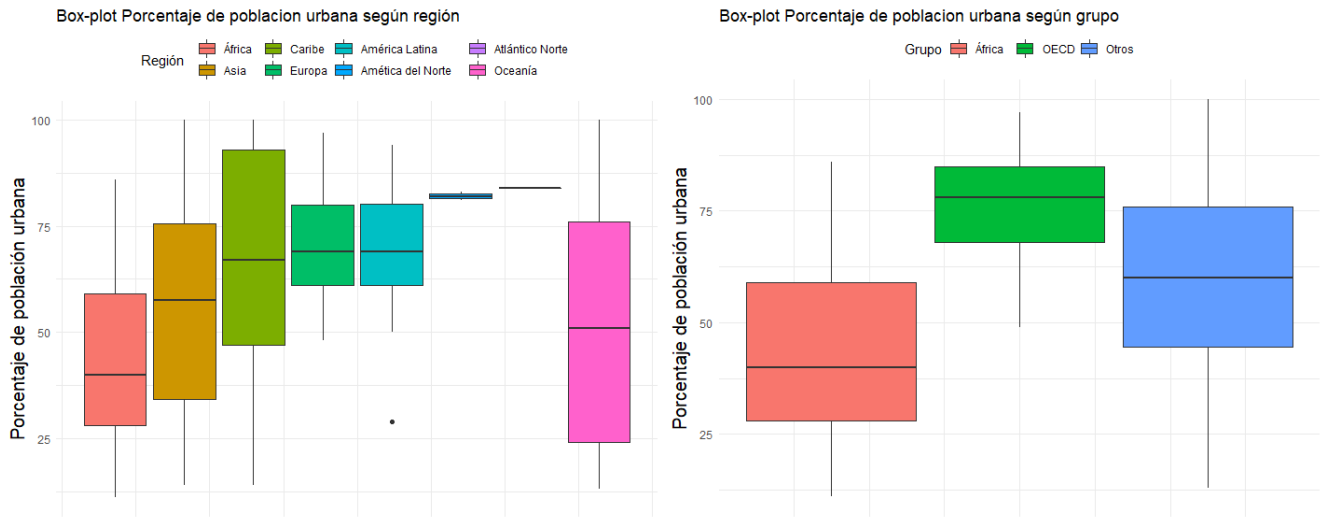

Figura 4: Box plot Porcentaje de población urbana según región y grupo

Finalmente, en la Figura 4 se realizan gráficos box plot para estudiar las posibles diferenciar entre regiones y grupos, así se aprecia que no hay diferencias entre América Latina y Europa en cuanto al porcentaje de población urbana. Nuevamente, África se encuentra más abajo en los valores de las variables, y es importante destacar la variabilidad presente en Oceanía. Si el análisis se centra en los grupos, es posible evidenciar que existen diferencias entre ellos, principalmente entre los grupos de países pertenecientes a África y OECD.

Matriz de correlación

Además, se estudia numéricamente la relación de las variables numéricas a través de una matriz de correlación. Es importante destacar que la correlación mide el grado de asociación entre las variables.

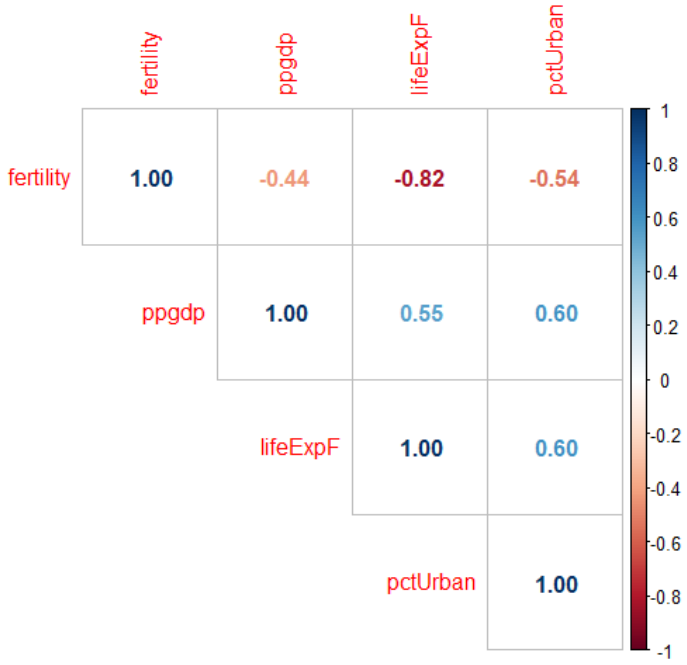

Figura 5: Gráficos de dispersión entre variables cuantitativas de la base de datos

Como se muestra en la Figura 5 existe asociación entre las variables numéricas. En particular, la correlación más alta es entre Esperanza de Vida y Fertilidad. En las demás variables, si bien la correlación podría no considerarse significativamente alta, tampoco se puede descartar que las correlaciones entre las variables sean bajas. Así se concluye que existe cierto grado de asociación entre las variables. Para precisar sobre las posibles relaciones de los datos, se procede a realizar gráficos de dispersión entre las variables.

Gráfico de Dispersión

A través del gráfico de dispersión es posible evidenciar patrones o relaciones entre las distintas variables.

Figura 6: Gráficos de dispersión entre variables cuantitativas de la base de datos

Como se muestra en la Figura 6, se puede apreciar que existe asociación entre las variables fertilidad y esperanza de vida, pues a mayor fertilidad tiende a aumentar también la otra variable. El gráfico de dispersión entre fertilidad y porcentaje de población urbana no tiende a mostrar un patrón muy claro, pero presenta una leve asociación inversamente proporcional. Además, las variables esperanza de vida y porcentaje de población urbana tienen a tener una asociación directamente proporcional y similar a una recta. El resto de gráficos no presentan asociaciones lineales. El gráfico de dispersión entre esperanza de vida y el producto nacional bruto (PBN) se observa un patrón exponencial creciente, mientras que en fertilidad y PBN también presentan la misma relación pero decreciente. Finalmente, las variables PBN y porcentaje de población urbana presentan un comportamiento logarítmico.

Histograma

También, es importante visualizar el comportamiento univariado de cada variable, para ello se realizarán histogramas para las variables continuas.

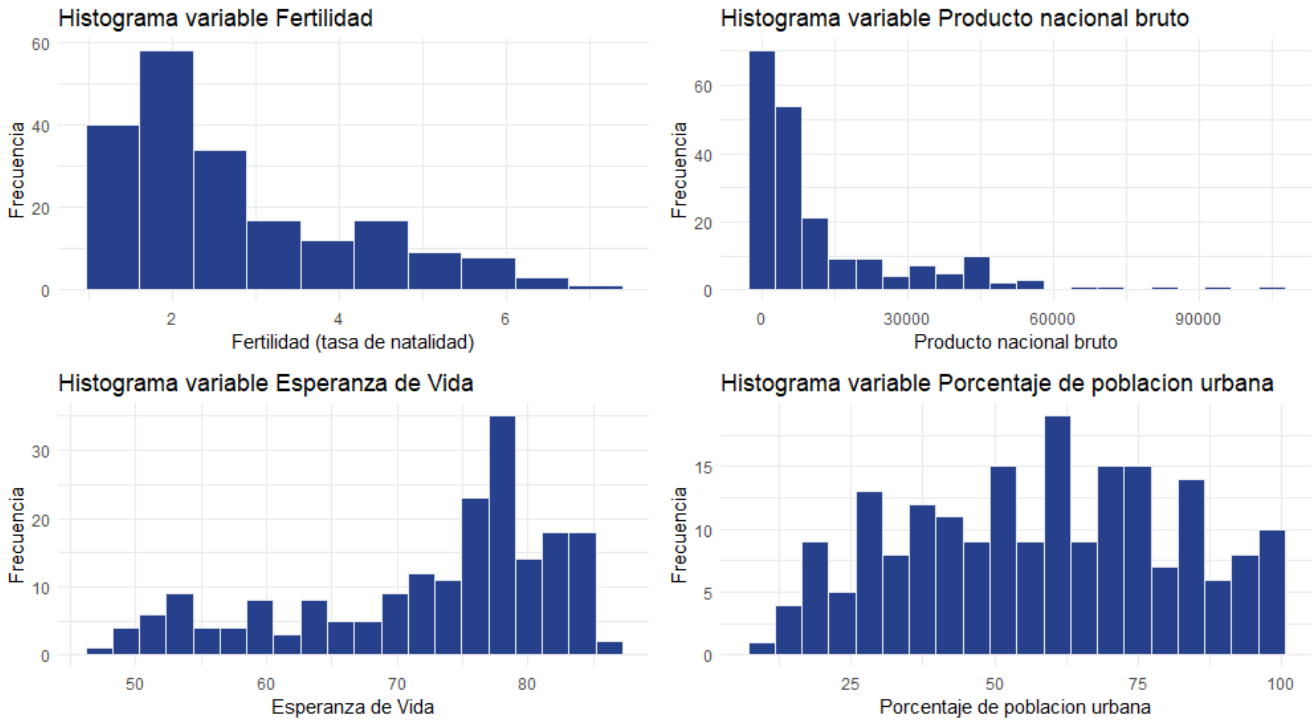

Figura 7: Histogramas por variable cuantitativa

Se puede observar en la Figura 7 que en caso de la variable fertilidad, se aprecia que ésta es asimétrica a la derecha, pues existen muchas observaciones con un baja tasa de natalidad, y pocas con valores altos. La variable `ppgdp` sigue el mismo comportamiento, sin embargo, éste es más pronunciado. Caso contrario, es la distribución de la variable `lifeExp` que es simétrica a la derecha. Finalmente, la variable `pctUrban` tiene un comportamiento simétrico.

Gráfico de barras

Finalmente se estudiará la distribución de países según región y grupo, es decir, las variables categóricas, para ello se utilizarán gráficos de barras.

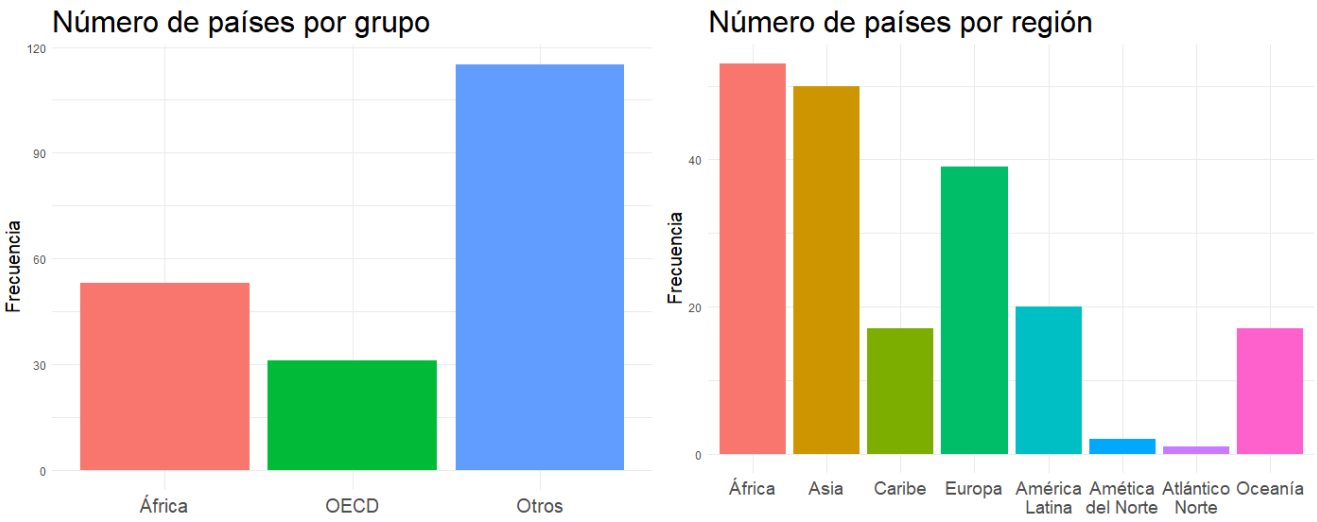

Figura 8: Histogramas por variable cuantitativa

En la Figura 8 se observa que el grupo con otro tipo de países tiene un número significativamente mayor que las demás categorías, siendo los países de la OECD que tienen menor cantidad de naciones. En cuanto a las regiones, África y Asia lideran la cantidad de países y le sigue Europa, además, las regiones de Caribe, América Latina y Oceanía tienen una cantidad menor. Finalmente, los países del Atlántico Norte solo considera un país mientras que América del Norte dos.

(b) Suponga que se desea analizar el efecto de las otras variables disponibles en $(fertility, lifeExp)$. Para esto, proponga un modelo estadístico y verifique si es adecuado. Usando algunas herramientas de inferencia estadística; responda si existe efecto de las otras variables sobre $(fertility, lifeExp)$. Escriba sus conclusiones

Solución

A continuación se quiere estudiar el efecto sobre las variables `(fertility, lifeExp)`.

Por lo tanto, la matriz de respuesta está dada por

$$\boldsymbol{Y} = (fertility, lifeExp)$$

y la matriz de diseño queda expresada por el resto de variables, además se debe notar la presencia de 2 variables cualitativas con más de 2 niveles, por lo que será necesario crear variables dummys para estudiar el efecto que tienen sobre la respuesta. A groso modo la matriz de diseño quedaría representada como

$$\boldsymbol{X} = (region, group, ppgdp, pctUrban)$$

donde $(region, group)$ se desglosan por sus respectivas variables dummys.

Luego, una primera hipótesis de interés es que la respuesta posee distribución normal. Para ello se quiere obtener la distancia de Mahalanobis para evaluar el supuesto y se denota por

$$\delta_i = (\boldsymbol{Y_i} - \boldsymbol{B}^T\boldsymbol{x_i})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y_i} - \boldsymbol{B}^T\boldsymbol{x_i}),\ i = 1, ..., n$$

y luego obtener una transformación con tal de realizar un QQ-Plot comparando los cuantiles de una distribución normal estándar.

$$z_i = \frac{(\hat{\delta}_i/p)^{1/3} - (1 - 2/9p)}{\sqrt{2/9p}} \tag{1}$$

La ecuación (1) es una transformación de la distancia de Mahalanobis de modo que su distribución corresponde a una normal estándar, además es necesario estimar δ, para lo cual se utilizan los estimadores máximo verosímiles de $\boldsymbol{B}$ y $\boldsymbol{\Sigma}$.

En primer lugar, la Figura 9 corresponde al QQ-plot considerando todas las covariables, se puede notar una fuerte curvatura en el centro, lo cual podría estar generando una pérdida de significancia del supuesto. Más aún, la Tabla 2 presenta los resultados después de realizar un test de Shapiro-Wilk, destacar que la hipótesis nula del test (denotada por H_0) es que los datos provienen de una distribución normal. Utilizando la función `shapiro.test` se rechaza la normalidad utilizando una significancia de 5%.

Estadístico W	Valor-p
0.9412	3.09×10^{-7}

Tabla 2: Test de Shapiro-Wilk modelo completo

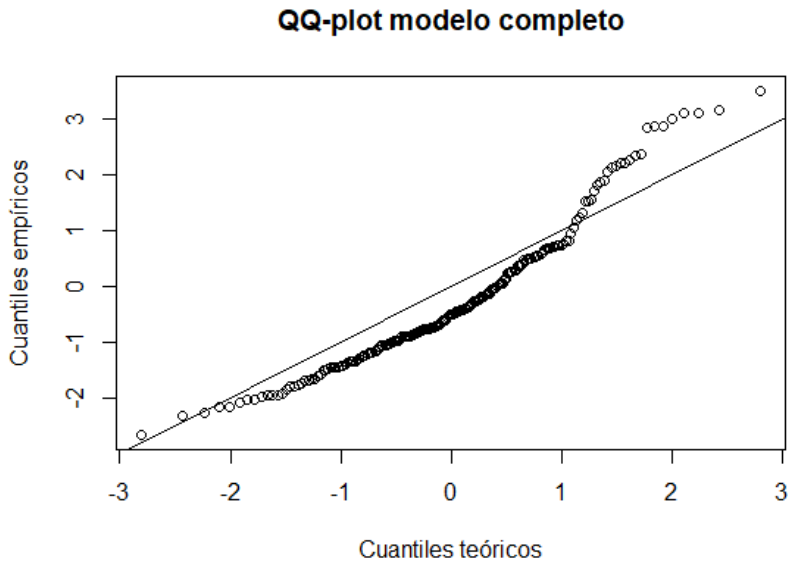

Figura 9: QQ-Plot modelo saturado

Luego se eliminó la variable `region` y en otro caso sólo se eliminó la variable `group`. La Figura 10 presenta los qq-plots respectivos a estos casos y la Tabla 3 presenta los

resultados del test de Shapiro-Wilk, donde se puede ver que en ambos casos nuevamente se rechaza normalidad.

Modelo	Estadístico W	Valor-p
Sin `group`	0.9729	6.95×10^{-4}
Sin `region`	0.9758	0.001647

Tabla 3: Test de Shapiro-Wilk modelo sin `group` y modelo sin `region`

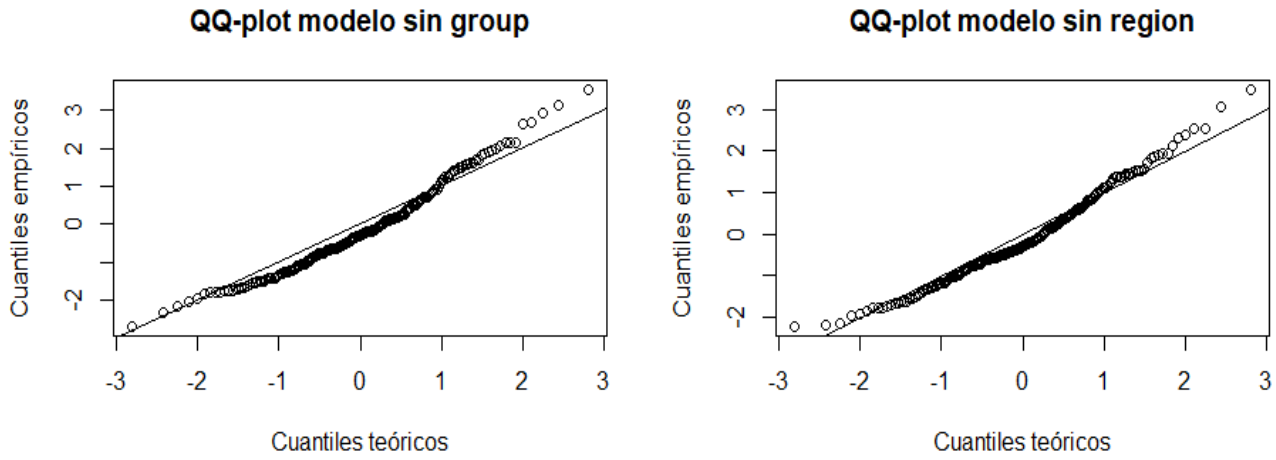

Figura 10: QQ-Plot modelo sin `group` y modelo sin `region`

Finalmente, el modelo propuesto considerará las covariables `ppgdp` y `pctUrban`, ya que, luego de analizar el QQ-plot ubicado en la Figura 11 y los resultados del test de Shapiro-Wilk en la Tabla 4, se observa que no se rechaza la hipótesis de normalidad. Por lo tanto, el modelo propuesto queda dado por

$$(fertility, lifeExp) = \boldsymbol{\beta_1} ppgdp + \boldsymbol{\beta_2} pctUrban + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N_2(0, \boldsymbol{\Sigma}) \quad (2)$$

donde $\boldsymbol{\beta_i}$ es un vector de dimensión 2.

Modelo	Estadístico W	Valor-p
sin `group`	0.9923	0.3782

Tabla 4: Test de Shapiro-Wilk ajustando el modelo con `ppgdp` y `pctUrban`

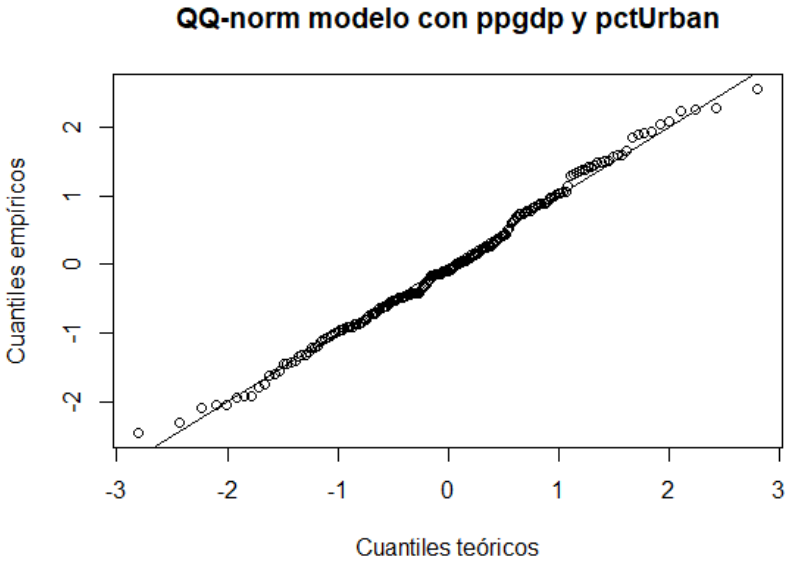

Figura 11: QQ-Plot modelo con `ppgdp` y `pctUrban`

Una vez que se comprobó el supuesto de normalidad se quiere estudiar la significancia de los predictores, pero sin antes revisar posibles problemas de colinealidad. La Tabla 5 presenta los factores de la Inflación de la Varianza (VIF) de cada predictor y el número de condicionamiento de la matriz de correlación de X. Con estos resultados se puede concluir que no existen problemas de colinealidad y que la matriz de datos queda bien condicionada.

Variable	VIF
ppgdp	1.558
pctUrban	1.558
N° condición	1.99

Tabla 5: Factores de inflación de la varianza y número de condición

Luego una primera hipótesis es $\boldsymbol{\beta_2} = \mathbf{0}$, es decir, los predictores de `pctUrban` no tienen influencia en la media de $\boldsymbol{Y}$. De este modo el modelo reducido es

$$(fertility, lifeExp) = \boldsymbol{\beta_1} ppgdp + \boldsymbol{\epsilon}$$

Mediante un test de razón de verosimilitud se obtiene que

$$-2ln\Lambda = 373,41$$

lo cual es mayor a $\chi^2_{1-\alpha}(2) = 5.99$ al considerar un nivel de significancia del 5%, por lo que se rechaza la hipótesis nula.

Por otro lado, una segunda hipótesis sería $\boldsymbol{\beta_1} = \mathbf{0}$, es decir, los predictores de `ppgdp` no tienen influencia en la respuesta. El modelo reducido en este caso es

$$(fertility, lifeExp) = \boldsymbol{\beta_2} pctUrban + \boldsymbol{\epsilon}$$

Y el test de razón de verosimilitud corresponde a

$$-2ln\Lambda = 29,86$$

lo cual es mayor a $\chi^2_{1-\alpha}(2) = 5.99$ al considerar un nivel de significancia del 5%, por lo que se estaría rechazando la hipótesis nula.

A modo de conclusión es posible obtener un modelo con 2 predictores significativos que logran explicar el conjunto de la variable respuesta, la ecuación de este modelo corresponde a la ecuación (2). Sus coeficientes estimados mediante máxima verosimilitud (MV) son los siguientes

$$\hat{\boldsymbol{B}} = (\beta_1^T, \beta_2^T) = \begin{pmatrix} -4.861 \times 10^{-5} & -0.00032 \\ 0.04924 & 1.19307 \end{pmatrix}$$

Además la matriz de covarianza de la respuesta estimada mediante MV está dada por

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 3.5906 & 25.9921 \\ & 498.3188 \end{pmatrix}$$

Una primera observación es que la variable `ppgdp` tiene menor influencia en la estimación para la respuesta que la variable `pctUrban`, es decir, para explicar la tasa de natalidad y la esperanza de vida la variable `pctUrban` tiene un mayor peso en la estimación, puesto que $\beta_2 > \beta_1$. Luego, es posible notar que la estimación para la variable `ppgdp`, β_1, es de signo negativo, lo que estaría diciendo que un mayor producto nacional bruto por persona reduciría la tasa de natalidad y la esperanza de vida en la media en el orden de 10^{-4}. Por otro lado, la estimación de la variable `pctUrban`, $\boldsymbol{\beta_2}$, es de signo positivo, lo cual tiene un efecto contrario que la variable `ppgdp`.

Otro intervalo de interés son los test coeficientes individuales, es decir, evaluar la significancia para los coeficiente β por cada predictor y respuesta. Para ello se realiza el siguiente intervalo de confianza

$$\hat{\beta}_{jk} \pm t_{1-\alpha/2}(n-q) S_k \sqrt{c_{jj}}$$

Donde $\hat{\beta}_{jk}$ es el estimador puntual de β_{jk}, S_k^2 es el k-ésimo elemento de la diagonal de $\frac{1}{n-q}\boldsymbol{E}^T\boldsymbol{E} = \boldsymbol{S}$ y c_{jj} es el j-ésimo elemento de $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

Repuesta	Covariable	Estimación	Límite Inferior	Límite Superior
`fertility`	`ppgdp`	-5×10^{-5}	-7×10^{-5}	-3×10^{-5}
`fertility`	`pctUrban`	0.04924	0.04311	0.05537

`lifeExp`	`ppgdp`	-0.00032	-0.00052	-0.00012
`lifeExp`	`pctUrban`	1.19308	1.12086	1.26529

Tabla 6: Intervalos de confianza para los test β individuales

En la Tabla 6 se puede apreciar las estimaciones para cada combinación de repuesta y predictor. Además, se pueden observar las estimaciones de cada parámetro y los intervalos de confianza. Se puede concluir que todos los parámetros son significativos pues no contienen el valor 0, ya que tanto los límites superiores como inferioes poseen el mismo signo.

En síntesis, primero se ajusto un modelo incluyendo las variables categóricas `group` y `region`, pero se comprueba que no cumple con los supuestos de normalidad multivariada a través del gráfico QQ-plot y el test Shapiro-Wilks de las distancias de Mahalanobis transformadas. Luego se ajusta el modelo solamente con la variables numéricas y se comprueba que, de esta manera, sí cumple con el supuesto de normailidad y se evalúa la significancia, tanto para cada coeficiente $\boldsymbol{\beta_i}$, es decir, para el predictor en ambas respuestas en conjunto, donde se concluye que las variables son significativas, por lo que cada covariable influye o explica las variable Fetilidad y Expectativa de Vida. Posteriormente, se realiza un análisis más específico para evaluar la significancia de los coeficientes β individualmente, para ello se realizan intervalos de confianza, donde se obtiene ninguno contiene al cero, por lo que se concluye que los cada predictor es significativo para cada respuesta, así cada covariable influye o explica las variables respuesta.