Universidad Autónoma de San Luis Potosí

Econometría

Tarea 2 13 de diciembre de 2021

Carlos Enrique Ponce Villagran

En el archivo Female GDPPPP.csv se encuentra, por país y a finales de 2018, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa (LFPRFemale), el Producto Interno Bruto a valores de Paridad de Poder Adquisitivo (GDPPPP, en cientos de dólares estadounidenses del 2017) y la población de esos países (Population).

1. Utilizando `Stata`:

   a) Encuentra los parámetros correspondientes para un modelo de regresión lineal múltiple que explique cómo afecta el valor del PIB-PPA en el porcentaje de mujeres mayores de años económicamente activas e interprétalos.

   Solución: Usando el comando `regress lfprfemale gdpppp` obtuvimos los siguientes parámetros de la Regresión

   $$\hat{\beta}_1 = 0.0037503 \quad \text{y} \quad \hat{\beta}_0 = 51.32837$$

   por lo tanto tenemos que por cada vez que el PIB-PPA aumenta tenemos que el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa aumenta en un 0.0037503.

   □

   b) ¿Cuál sería el porcentaje esperado de población femenina económicamente activa en un país con PIB en valores PPA de 122.5469?

   Solución: Evaluando esto en `Stata` con `mfx, at(122.5469)` obtenemos

   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 122.5469 = 51.787959$$

   □

   c) Realiza una gráfica de los datos junto con la curva de regresión para el modelo ajustado anteriormente.

   Solución: Usando `scatter lfprfemale gdpppp || line yh gdpppp` obtenemos la siguiente gráfica de los datos

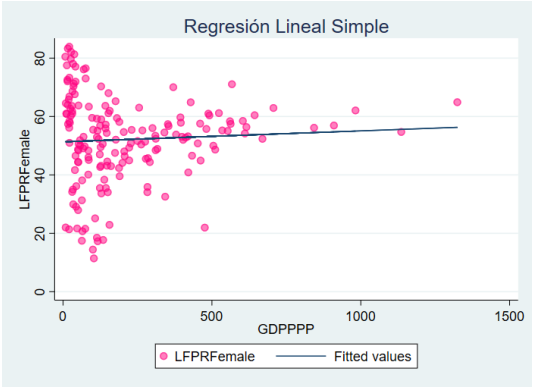

Figura 1: Gráfica de los datos con su recta ajustada

□

d) Identifica, mediante el nivel de influencia y la distancia de Cook, la existencia o no de datos atípicos para el modelo ajustado.

Solución: Usamos el comando `predict hrls, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate infrls=hrls>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos que las observaciones 7, 8, 11, 15, 23, 29, 44, 55, 56, 60, 69, 71, 76, 84, 93, 94, 112, 118, 130, 138, 142, 153, 154, 167, 168, 169.

Ahora, para la distancia de Cook utilizamos el comando `predict drls, cooksd` y utilizamos el comando `generate ckdrls=drls>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

e) Repetir los incisos a), b), c) y d) para un modelo sin intercepto.

a) *Solución:* Usando el comando `regress lfprfemale gdpppp, noconstant` obtuvimos el siguiente parámetro de la Regresión

$$\hat{\beta}_1 = 0.1174968$$

por lo tanto tenemos que por cada vez que el PIB-PPA aumenta tenemos que el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa aumenta en un 0.1174968.

□

b) *Solución:* Evaluando esto en `Stata` con`mfx, at(122.5469)` obtenemos

$$\hat{y} = \hat{\beta}_1 \cdot 122.5469 = 14.398867$$

□

c) *Solución:* Usando `scatter lfprfemale gdpppp || line yhrlsi gdpppp` obtenemos la siguiente gráfica de los datos

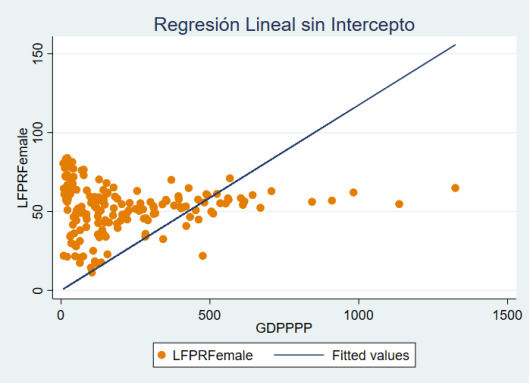

Figura 2: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `predict hrlsi, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate infrlsi=hrlsi>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos que las observaciones 7, 8, 11, 15, 23, 29, 44, 55, 56, 60, 69, 71, 76, 84, 93, 94, 112, 118, 130, 138, 142, 153, 154, 167, 168, 169.

Ahora, para la distancia de Cook utilizamos el comando `predict drlsi, cooksd` y utilizamos el comando `generate ckdrlsi=drlsi>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

f) Repetir los incisos a), b), c) y d) para un modelo linlog.

a) *Solución:* Usando aplicando los datos la transformación logaritmo con `generate lngdpppp=ln(gdpppp)` entonce aplicando la regresión con el comando `regress lfprfemale lngdpppp` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = -1.608956 \quad \text{y} \quad \hat{\beta}_0 = 59.81571$$

entonces la interpretación es que por unidad porcentual que el PIB-PPA aumenta, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa disminuye $0.01608956\,\%$

□

b) *Solución:* Encontraremos primero el valor de $\ln(122.5469)$ con el comando `display ln(122.5469)` esto nos muestra que $\ln(122.5469) = 4.8084938$ entonces con`mfx, at(4.8084938)` obtenemos

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 4.8084938 = 52.079057$$

□

c) *Solución:* Usando `scatter lfprfemale gdpppp || line yhlinlog gdpppp` obtenemos la siguiente gráfica de los datos

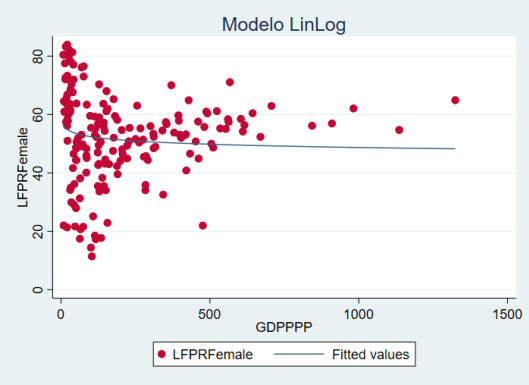

Figura 3: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `predict hlinlog, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate inflinlog=hlinlog>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos datos atipicos de la observación 1 a la 34 y de 141 a la 175.

Ahora, para la distancia de Cook utilizamos el comando `predict dlinlog, cooksd` y utilizamos el comando `generate ckdlinlog=dlinlog>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

g) Repetir los incisos a), b), c) y d) para un modelo loglin

a) *Solución:* Usando aplicando los datos la transformación logaritmo con `generate lnlfprfemale=ln(lfprfemale)` entonce aplicando la regresión con el comando `regress lnlfprfemale gdpppp` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = .0001764 \quad \text{y} \quad \hat{\beta}_0 = 3.862149$$

entonces la interpretación es que por cada vez quel PIB-PPA aumenta, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa aumenta en $0.01764\,\%$

□

b) *Solución:* Utilizando el comando `mfx, at(122.5469)` obtenemos $\ln(y)$ entonces al aplicar $\exp$ tenemos el valor

$$\hat{y} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 122.5469\right) = 48.606934$$

□

c) *Solución:* Usando `scatter lfprfemale gdpppp || line yhloglin gdpppp` obtenemos la siguiente gráfica de los datos

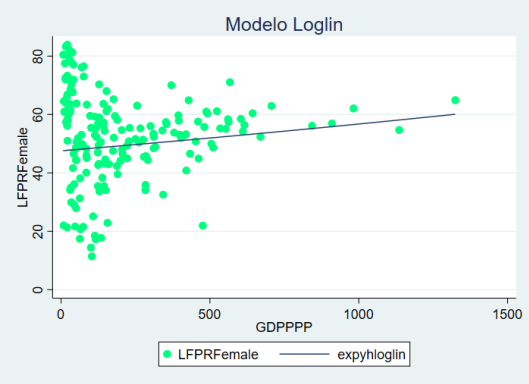

Figura 4: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `predict hloglin, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate infloglin=hloglin>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos datos atípicos de la observación 150 a la 175.

Ahora, para la distancia de Cook utilizamos el comando `predict dloglin, cooksd` y utilizamos el comando `generate ckdloglin=dloglin>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

h) Repetir los incisos a), b), c), y d) para un modelo loglog.

a) *Solución:* Usando aplicando la regresión con el comando `regress lnlfprfemale lngdpppp` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = -0.011403 \quad \text{y} \quad \hat{\beta}_0 = 3.954212$$

entonces la interpretación es que por unidad porcentual que el PIB-PPA aumenta, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa disminuye $0.00011403\,\%$.

□

b) *Solución:* Utilizando el comando `mfx, at(122.5469)` obtenemos

$$y = e^{\beta_0} e^{\beta_1 \ln(122.5469)} = 49.371893$$

□

c) *Solución:* Usando aplicando la función exponencial para `yhloglog` y graficar tenemos `scatter lfprfemale gdpppp || line eyhloglog gdpppp` obtenemos la siguiente gráfica de los datos

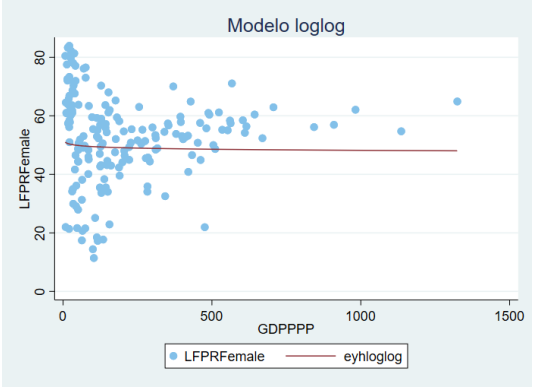

Figura 5: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `predict hloglog, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate infloglog=hloglog>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos datos atípicos de la observación 1 a la 34 y 142 a la 175 (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `predict dloglog, cooksd` y utilizamos el comando `generate ckdloglog=dloglog>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

i) Repetir los incisos a), b), c), y d) para un modelo reciproco.

a) *Solución:* Generamos los datos reciprocos de `GDPPPP` con `generate rcpgdpppp=1/gdpppp` aplicando la regresión con el comando `regress lfprfemale rcpgdpppp` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 166.8805 \quad \text{y} \quad \hat{\beta}_0 = 49.34352$$

Este modelo no tiene una interpretación clara de los parámetros.

□

b) *Solución:* Utilizando `display 1/122.5469` y el comando `mfx, at(0.00816014)` obtenemos

$$y = \beta_0 + \beta_1 \frac{1}{122.5469} = 50.70529$$

□

c) *Solución:* Usando `scatter lfprfemale gdpppp || line yhrcp gdpppp` obtenemos la siguiente gráfica de los datos

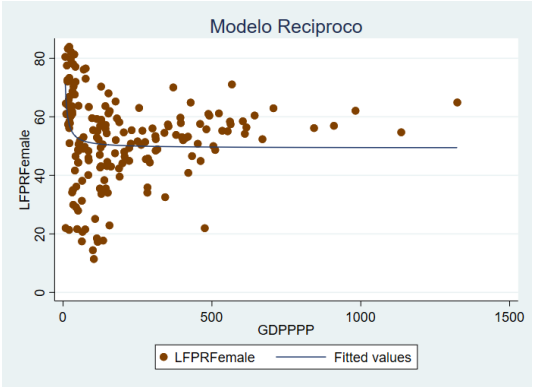

Figura 6: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `predict hrcp, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate infrcp=hrcp>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos datos atípicos de la observación 1 a la 21 (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `predict drcp, cooksd` y utilizamos el comando `generate ckdrcp=drcp>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

j) Repetir los incisos a), b), c), y d) para un modelo log-reciproco.

a) *Solución:* Aplicando la regresión con el comando `regress lnlfprfemale rcpgdpppp` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 2.574879 \quad \text{y} \quad \hat{\beta}_0 = 3.856772$$

Este modelo no tiene una interpretación clara de los parámetros.

□

b) *Solución:* Utilizando el comando `mfx, at(0.00816014)` obtenemos

$$y = \exp(\beta_0 + \beta_1 \frac{1}{122.5469}) = 48.316982$$

□

c) *Solución:* Usando `scatter lfprfemale gdpppp || line expyhlogrcp gdpppp` obtenemos la siguiente gráfica de los datos

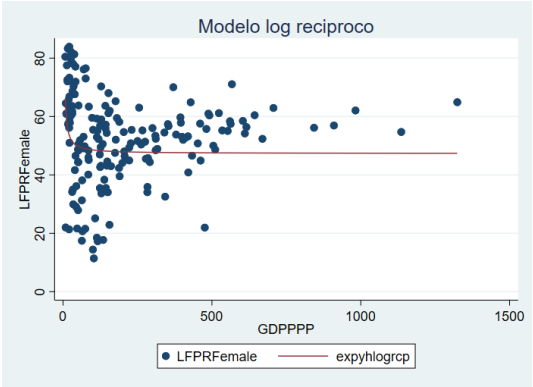

Figura 7: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `predict hlogrcp, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate inflogrcp=hlogrcp>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos datos atípicos de la observación 1 a la 21 (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `predict dlogrcp, cooksd` y utilizamos el comando `generate ckdlogrcp=drcp>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

k) Repetir los incisos a), b), c) y d) para un modelo polinomial de grado 3.

a) *Solución:* Generando los datos `generate gdpppp3=gdpppp`$\hat{3}$ y aplicando la regresión con el comando `regress lfprfemale gdpppp3` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 6.8 \times 10^{-9} \quad \text{y} \quad \hat{\beta}_0 = 51.69942$$

Este modelo no tiene una interpretación clara de los parámetros.

□

b) *Solución:* Utilizando el comando `mfx, at(122.5469)` obtenemos

$$y = \hat{\beta}_0 + \hat{\beta}_1 (122.5469)^3 = 51.69942$$

□

c) *Solución:* Usando `scatter lfprfemale gdpppp || line expyhlogrcp gdpppp` obtenemos la siguiente gráfica de los datos

□

d) *Solución:* Usamos el comando `predict h3, hat` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos el comando `generate inf3=h3>=(2/175)` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a $2/175$ y por lo tanto son datos atípicos entonces de estos datos generados tenemos datos atípicos de la observación 169 a la 175 (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `predict d3, cooksd` y utilizamos el comando `generate ckd3=d3>1` para generar un conjunto de datos con unos y ceros, donde los unos indican que la i-ésima observación es mayor a 1 y por lo tanto, según el criterio de la distancia de Cook, este modelo no presenta datos atípicos.

□

l) De acuerdo con el criterio del error cuadrático medio, ¿cuál de todos los modelos anteriores es el más adecuado para los datos?

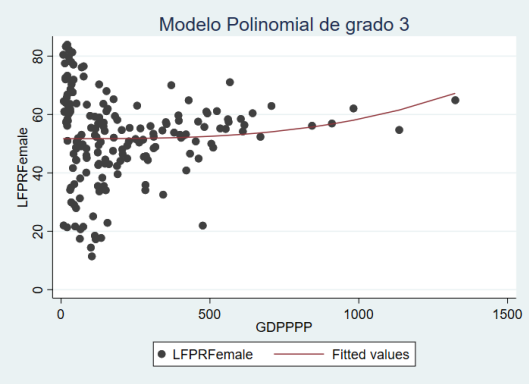

Figura 8: Gráfica de los datos con su recta ajustada

Solución: A continuación presentamos los errores cuadráticos medios de todos los modelos que se mostraron en las tablas al usar `regress`

Modelo	RS	Sin intercepto	linlog	loglin	loglog	Reciproco	log-reciproco	Polinomial 3
MSE	225.981292	1623.67798	223.10987	231.8277	229.4852	212.859993	218.215	224.251851

Por lo tanto, el modelo con el MSE más pequeño es el modelo Reciproco, por lo tanto por el criterio del MSE tenemos que el modelo reciproco es el modejo que mejor ajusta a los datos.

□

m) De acuerdo con el criterio de Akaike, ¿cuál de todos los modelos anteriores es el más adecuado para los datos?

Solución: A continuación presentamos los estadísticos de Akaike utilizando el comando `estat ic`

Modelo	RS	Sin intercepto	linlog	loglin	loglog	Reciproco	log-reciproco	Polinomial 3
AIC	1447.196	1791.304	1444.958	138.6597	140.5884	1436.728	136.3195	1445.852

Por lo tanto, siguiendo el criterio de Akaike el modelo que se acerca mas al ajuste original es el modelo log-reciproco.

□

2. Repetir el inciso 1) utilizando `R`.

a) Encuentra los parámetros correspondientes para un modelo de regresión lineal múltiple que explique cómo afecta el valor del PIB-PPA en el porcentaje de mujeres mayores de años económicamente activas e interprétalos.

Solución: Usando el comando `rl<-lm(LFPRFemale GDPPPP,data = datos)` y `summary(rl)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 0.0037503 \quad \text{y} \quad \hat{\beta}_0 = 51.32837$$

por lo tanto tenemos que por cada vez que el PIB-PPA aumenta tenemos que el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa aumenta en un 0.0037503.

□

b) ¿Cuál sería el porcentaje esperado de población femenina económicamente activa en un país con PIB en valores PPA de 122.5469?

Solución: Evaluando esto con `predict(rl, newdata=data.frame(GDPPPP=122.5469))` obtenemos

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 122.5469 = 51.78796$$

□

c) Realiza una gráfica de los datos junto con la curva de regresión para el modelo ajustado anteriormente.

Solución: Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Regresión lineal simple"),pch=19,col="mediumpurple1")` obtenemos la siguiente gráfica de los datos y `abline(rl)` para graficar la recta

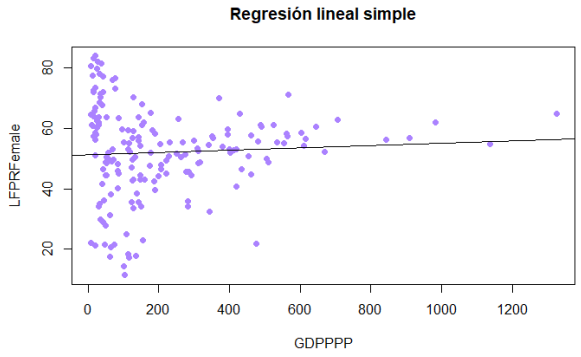

Figura 9: Gráfica de los datos con su recta ajustada

□

d) Identifica, mediante el nivel de influencia y la distancia de Cook, la existencia o no de datos atípicos para el modelo ajustado.

Solución: Usamos el comando `Hrl<-hatvalues(rl)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima e lnúmero de observación en caso de que `Hrl[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que 7, 8, 11, 15, 23, 29, 44, 55, 56, 60, 69, 71, 76, 84, 93, 94, 112, 118, 130, 138, 142, 153, 154, 167, 168, 169 son datos atípicos.

Ahora, para la distancia de Cook utilizamos el comando `drls<-cooks.distance(rl)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `drls[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

e) Repetir los incisos a), b), c) y d) para un modelo sin intercepto.

a) *Solución:* Usando el comando `rlint<-lm(LFPRFemale GDPPPP-1,data = datos)` y `summary(rlint)` obtuvimos el siguiente parámetro de la Regresión

$$\hat{\beta}_1 = 0.117497$$

por lo tanto tenemos que por cada vez que el PIB-PPA aumenta tenemos que el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activa aumenta en un 0.117497.

□

b) *Solución:* Evaluando esto con `predict(rlint, newdata=data.frame(GDPPPP=122.5469))` obtenemos

$$\hat{y} = \hat{\beta}_1 \cdot 122.5469 = 14.39887$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Regresión lineal simple"),pch=19,col="burlywood1")` obtenemos la siguiente gráfica de los datos y `abline(rlint)` para graficar la recta

□

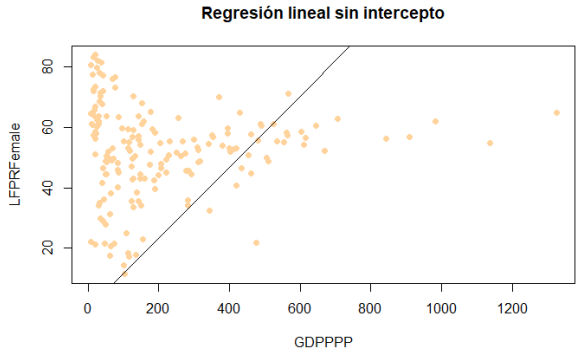

Figura 10: Gráfica de los datos con su recta ajustada

d) *Solución:* Usamos el comando `Hrlint<-hatvalues(rlint)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hrlint[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que 7, 8, 11, 15, 23, 29, 44, 55, 56, 60, 69, 71, 76, 84, 93, 94, 112, 118, 130, 138, 142, 153, 154, 167, 168, 169 son datos atípicos.

Ahora, para la distancia de Cook utilizamos el comando `drlsint<-cooks.distance(rlint)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `drlsint[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

f) Repetir los incisos a), b), c) y d) para un modelo linlog.

a) *Solución:* Usando el comando `rllinlog<-lm(LFPRFemale log(GDPPPP),data = datos)` y `summary(rllinlog)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = -1.6090 \quad \text{y} \quad \hat{\beta}_0 = 59.8157$$

La interpretación es que por cada unidad porcentual que el PIB-PPA aumenta, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activada disminuye en $1.6090\,\%$.

□

b) *Solución:* Evaluando esto con `predict(rllinlog, newdata=data.frame(GDPPPP=122.5469))` obtenemos

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \ln(122.5469) = 52.07906$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Regresión lineal simple"),pch=19,col="firebrick2")` obtenemos la siguiente gráfica de los datos y `lines(sort(predict(rllinlog, newdata=data.frame(GDPPPP=1:1500))))` el ajuste del modelo

□

d) *Solución:* Usamos el comando `Hlinlog<-hatvalues(rllinlog)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hlinlog[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que los datos 107 al 175 son atípicos (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `dlinlog<-cooks.distance(rllinlog)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `dlinlog[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

g) Repetir los incisos a), b), c) y d) para un modelo loglin.

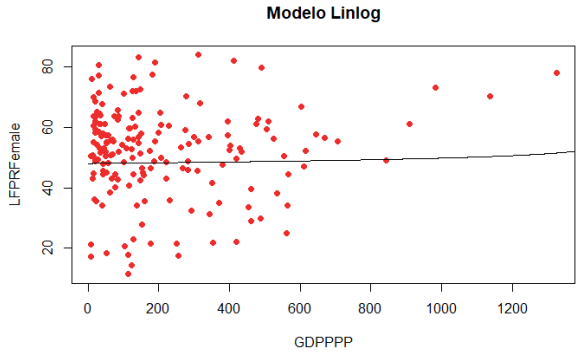

Figura 11: Gráfica de los datos con su recta ajustada

a) *Solución:* Usando el comando `rlloglin<-lm(log(LFPRFemale) GDPPPP,data = datos)` y `summary(rlloglin)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 0.0001764 \quad \text{y} \quad \hat{\beta}_0 = 3.8621492$$

entonces por cada unidad que el PIB-PPA aumenta, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activas aumenta en $0.01764\,\%$.

□

b) *Solución:* Evaluando esto con `exp(predict(rlloglin, newdata=data.frame(GDPPPP=122.5469)))` ya que sabemos que $Y = \exp(\beta_0 + \beta_1 x)$, entonces obtenemos el valor

$$\hat{y} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 122.5469\right) = 48.60693$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Regresión lineal simple"),pch=19,col="firebrick2")` obtenemos la siguiente gráfica de los datos y `lines(sort(predict(rllinlog, newdata=data.frame(GDPPPP=1:1500))))` el ajuste del modelo

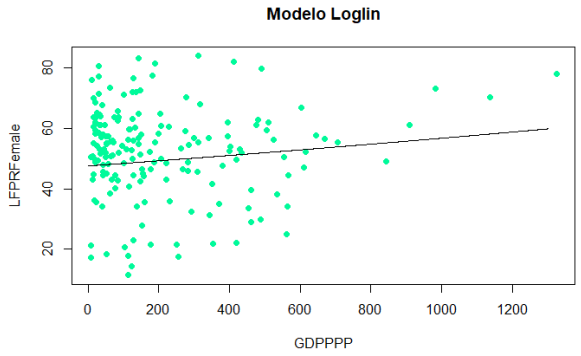

Figura 12: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `Hloglin<-hatvalues(rlloglin)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hloglin[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que los datos 150 al 175 son atípicos (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `dloglin<-cooks.distance(rlloglin)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `dloglin[i]>1`, en caso de imprimir `i`

vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

h) Repetir los incisos a), b), c) y d) para un modelo loglog.

a) *Solución:* Usando el comando `rlloglog<-lm(log(LFPRFemale) log(GDPPPP),data = datos)` y `summary(rlloglog)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = -0.0114 \quad \text{y} \quad \hat{\beta}_0 = 3.9542$$

entonces por cada unidad porcentual que el PIB-PPA aumenta, el porcentaje de mujeres mayores de 15 años que forman parte de la población económicamente activas aumenta en $0.000114\,\%$.

□

b) *Solución:* Evaluando esto con `exp(predict(rlloglog, newdata=data.frame(GDPPPP=122.5469)))` ya que sabemos que $Y = \exp(\beta_0 + \beta_1 x)$, entonces obtenemos el valor

$$\hat{y} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot \ln(122.5469)\right) = 49.37189$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Regresión loglog"),pch=19,col="lightsalmon1")` obtenemos la siguiente gráfica de los datos y `lines(sort(predict(rlloglog, newdata=data.frame(GDPPPP=1:1300))))` el ajuste del modelo

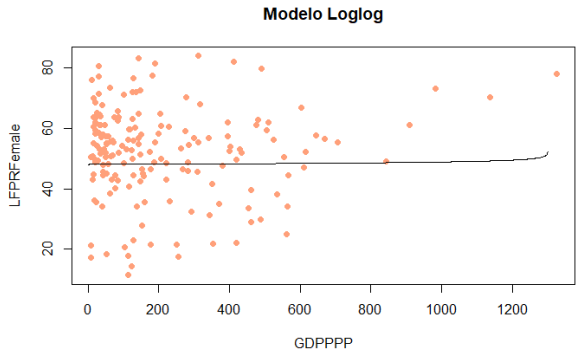

Figura 13: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `Hloglog<-hatvalues(rlloglog)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hloglog[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que los datos 107 al 175 son atípicos (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `dloglog<-cooks.distance(rlloglog)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `dloglog[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

i) Repetir los incisos a), b), c) y d) para un modelo reciproco.

a) *Solución:* Generamos los datos reciprocos con `datos$rcpGDPPPP<-1/datos$GDPPPP` usando el comando `rlrcp<-lm(LFPRFe = datos)` y `summary(rlrcp)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 166.88 \quad \text{y} \quad \hat{\beta}_0 = 49.34$$

FALTAAAAAAA INTERPRETACIOOOOOOON

□

b) *Solución:* Evaluando esto con `predict(rlrcp, newdata=data.frame(rcpGDPPPP=1/122.5469))` ya que sabemos que $Y = \exp(\beta_0 + \beta_1 x)$, entonces obtenemos el valor

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{122.5469} = 50.70529$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Modelo Reciproco"),pch=19,col="lightseagreen")` obtenemos la siguiente gráfica de los datos y `lines(sort(predict(rlrcp, newdata=data.frame(rcpGDPPPP=1:1300))))` el ajuste del modelo

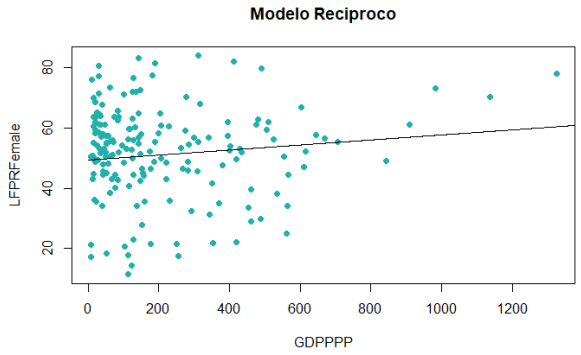

Figura 14: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `Hrcp<-hatvalues(rlrcp)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hrcp[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que los datos 155 al 175 son atípicos (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `dloglog<-cooks.distance(rlloglog)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `dloglog[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

j) Repetir los incisos a), b), c) y d) para un modelo log-reciproco.

a) *Solución:* Usando el comando `rllogrcp<-lm(log(LFPRFemale) rcpGDPPPP,data = datos)` y `summary(rllogrcp)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 2.5749 \quad \text{y} \quad \hat{\beta}_0 = 3.8568$$

Este modelo no tiene una interpretación clara de los parámetros.

□

b) *Solución:* Evaluando esto con `predict(rllogrcp, newdata=data.frame(rcpGDPPPP=1/122.5469))` ya que sabemos que $Y = \exp(\beta_0 + \beta_1 x)$, entonces obtenemos el valor

$$\hat{y} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \frac{1}{122.5469}\right) = 48.31698$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Modelo Reciproco"),pch=19,col="deepskyblue2")` obtenemos la siguiente gráfica de los datos y `lines(sort(predict(rlrcp, newdata=data.frame(logrcpGDPPPP=1:1300))))` el ajuste del modelo

□

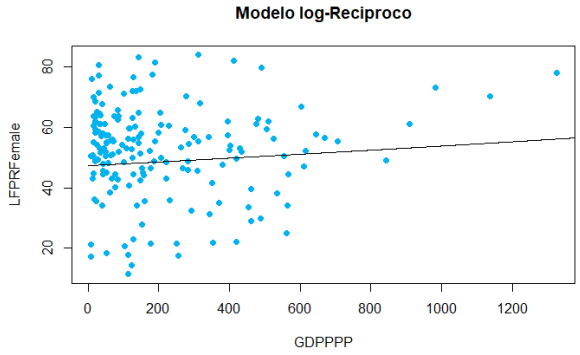

Figura 15: Gráfica de los datos con su recta ajustada

d) *Solución:* Usamos el comando `Hlogrcp<-hatvalues(rllogrcp)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hlogrcp[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que los datos 155 al 175 son atípicos (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `dlogrcp<-cooks.distance(rllogrcp)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `dlogrcp[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

k) Repetir los incisos a), b), c) y d) para un modelo polinomial de grado 3.

a) *Solución:* Generamos los datos al cubo con `datos$GDPPPP3<-(datos$GDPPPP)`$\hat{3}$ y usando el comando `rlp3<-lm(LFPRFemale = datos)` y `summary(rlp3)` obtuvimos los siguientes parámetros de la Regresión

$$\hat{\beta}_1 = 6.703 \times 10^{-9} \quad \text{y} \quad \hat{\beta}_0 = 51.70$$

Este modelo no tiene una interpretación clara de los parámetros.

□

b) *Solución:* Evaluando esto con `predict(rlp3, newdata=data.frame(GDPPPP3=122.5469))` obtenemos el valor

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 122.5469 = 51.69942$$

□

c) *Solución:* Usando `plot(datos$GDPPPP,datos$LFPRFemale,xlab = "GDPPPP", ylab = "LFPRFemale", title(main = "Modelo Polinomial de grado 3"),pch=19,col="springgreen4")` obtenemos la siguiente gráfica de los datos y `lines(sort(predict(rlrcp, newdata=data.frame(GDPPPP3=1:1300))))` el ajuste del modelo

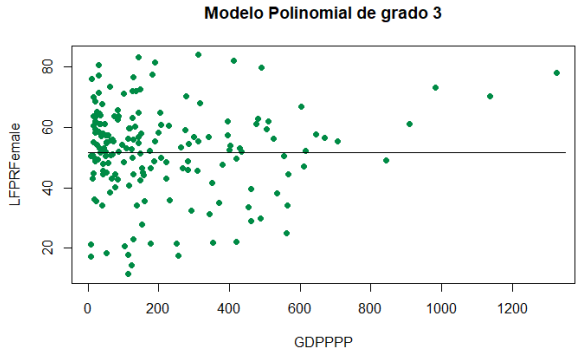

Figura 16: Gráfica de los datos con su recta ajustada

□

d) *Solución:* Usamos el comando `Hp3<-hatvalues(rlp3)` para obtener los niveles de influencia y agregarlos al conjunto de datos y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `Hp3[i]>=2/175`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, obtuvimos que los datos 169 al 175 son atípicos (los datos estando de manera ordenada).

Ahora, para la distancia de Cook utilizamos el comando `dp3<-cooks.distance(rlp3)` y utilizamos un comando `for if` que nos imprima el número de observación en caso de que `dp3[i]>1`, en caso de imprimir `i` vamos a tener un dato atípico. Al ejecutarlo, no se imprimió ningún indice por lo tanto, según el criterio de Cook, este modelo no presenta datos atípicos.

□

l) De acuerdo con el criterio del error cuadrático medio, ¿cuál de todos los modelos anteriores es el más adecuado para los datos?

Solución: A continuación presentamos los errores cuadráticos medios de todos los modelos que se generaron con la función `mse<-function(rl) mean(rl$residuals^2)`

Modelo	RS	Sin intercepto	linlog	loglin	loglog	Reciproco	log-reciproco	Polinomial 3
MSE	223.3986	1614.4	220.56	231.8277	229.4852	210.4273	218.215	221.689

Por lo tanto, el modelo con el MSE más pequeño es el modelo Reciproco, por lo tanto por el criterio del MSE tenemos que el modelo reciproco es el modejo que mejor ajusta a los datos.

□

m) De acuerdo con el criterio de Akaike, ¿cuál de todos los modelos anteriores es el más adecuado para los datos?

Solución: A continuación presentamos los estadísticos de Akaike obtenidos con la función `AIC()`

Modelo	RS	Sin intercepto	linlog	loglin	loglog	Reciproco	log-reciproco	Polinomial 3
AIC	1449.196	1793.304	1446.958	140.6597	142.5884	1438.728	138.3195	1447.852

Por lo tanto, siguiendo el criterio de Akaike el modelo que se acerca mas al ajuste original es el modelo log-reciproco.

□