UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN
MATEMÁTICAS APLICADAS Y COMPUTACIÓN

REGRESIÓN LINEAL. ESTADÍTICA II. SEMESTRE VI.

Análisis de la pobreza en México 2014 y 2016

Modelo de Regresión Lineal

$$y = \beta_0 + \sum_{i=1} \beta_i x_i + \epsilon_i.$$

Barrientos Aguilar Juan Carlos -> 100%
García Totozintle Sergio Antonio -> 100%
Pérez Acosta Asiel -> 100%
Sarmiento Ibarra Jonatan -> 100%

Contenido

1. Tabla de ilustraciones

2. Resumen

En este trabajo se buscó encontrar si existe una relación lineal entre los datos de pobreza en México recabados por del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) entre los años 2014 y 2016, con la ayuda del lenguaje de programación R que es especializado en el análisis estadístico se obtuvieron los resultados pertinentes para aprobar dicha relación. Se generó la gráfica de dispersión y la gráfica de la regresión lineal para poder observar que cierta aproximación es válida (verificando si el modelo de regresión generado es óptimo). Para brindar una mayor certeza sobre los resultados, se realizaron las diversas pruebas para demostrar los supuestos de nuestro modelo tales como los son la prueba de Pearson, Breusch-Pagan, Durbin-Watson y Anderson-Darling. Los resultados arrojados indicaron que dicha aproximación lineal no es óptima dado que, uno de los supuestos (homocedasticidad) no se cumplió.

Comentado [JV1]: coma

Comentado [JV2]: coma

Comentado [JV3]: por medio de un modelo de regresión lineal simple estimado por el método de mínimos cuadrados.

3. Palabras clave

Regresión Lineal

Pruebas de:

i) Coeficiente de correlación de Pearson
ii) Breusch-Pagan
iii) Anderson-Darling
iv) Durbin-Watson.

Comentado [JV4]: Pobreza tendría que ser una palabra clave

4. Introducción

En México, la pobreza ha sido un problema recurrente durante décadas, afectando a millones de personas en todo el país. Entre los años 2014 y 2016, el índice de pobreza en México disminuyó ligeramente esto según datos del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). Es importante señalar que, aunque el índice de pobreza disminuyó en general, este vario según la entidad federativa.

Que el índice disminuyera del 2014 al 2016 no necesariamente implica que exista una relación lineal entre estos años. Por lo tanto, es necesario llevar a cabo un análisis estadístico para determinar si existe una relación lineal entre los diferentes años en términos del índice de pobreza en México. Para lograr este objetivo, se analizó los datos disponibles del CONEVAL sobre la pobreza en México en los años 2014 y 2016, utilizando el lenguaje de programación R para identificar cualquier relación entre ambos periodos.

Comentado [JV5]: analizaron

Comentado [JV6]: Están tema y objetivo. Faltó la estructura del recporte.

5. Marco teórico

5.1 Formulario

$H_0 \rightarrow$ Hipótesis nula: es una hipótesis que el investigador pretende refutar, rechazar o anular.

$H_1 \rightarrow$ Hipótesis alternativa: es lo que el investigador quiere probar (que se demuestre dicha hipótesis).

Nivel de significancia ($\alpha = 0.05$): límite para juzgar un resultado como estadísticamente significativo. Para todo test, si:

$$P - value < \alpha$$

Entonces, se rechaza H_0.

(Cabana, 2021)

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}, \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}}$$

(Carollo, 2012)

$$SCE_R = SCPE = \sum_{j=1}^{c}\sum_{i=1}^{c}\left(Y_{ij} - \bar{Y}_j\right)^2$$

$$SCE_F = SCE = \sum_{j=1}^{c}\sum_{i=1}^{c}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

$$gl_R = (n-k), gl_F = (n-c)$$

SCE_R: suma de cuadrados del error (residuos) del modelo ajustado.

SCE_F: suma de cuadrados del error (residuos) del modelo completo.

gl_R: grados de libertad del modelo ajustado.

gl_F: grados de libertad del modelo completo

(Vela Peon, 2011)

5.2 Análisis de regresión lineal

El objetivo de la regresión lineal es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) y y un conjunto de variables independientes (variables explicativas) $x_1, x_2, \ldots, x_n$. El modelo de regresión lineal simple tiene la siguiente expresión:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde: Y es la variable dependiente.
β_0 es la ordenada en el origen (intercepto).
β_1 es la pendiente de la recta (indica cómo cambia Y al incrementar X en una unidad).
ε es el término de error o residuo.

En el análisis de regresión lineal, se parte de la suposición de que la relación entre las variables independientes y la variable dependiente es lineal. Por ende, para realizar la estimación del modelo de regresión lineal simple, se busca una línea que mejor se ajuste a los datos. A esta línea se le conoce como la línea de regresión y se representa por la ecuación:

$$\hat{Y} = b_0 + b_1 X$$

Forma general:

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_k X_k$$

Donde: $\hat{Y}$ es la variable dependiente.
b_0 es la intersección de la línea de regresión con el eje Y (conocido como: término constante o intercepto).
b_1 es la pendiente de la línea de regresión (también conocida como el coeficiente de regresión).
X es la variable independiente.

Para encontrar la línea de regresión, utilizaremos el método de mínimos cuadrados. Este método consiste en minimizar la suma de los cuadrados de los errores (residuos):

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Es decir, la suma de los cuadrados de las diferencias entre los valores reales observados (Y_i) y los valores estimados ($\hat{Y}_i$).

Con dicho método, tenemos que:

$$b_1 = \frac{S_{xy}}{S_x^2}$$
$$b_0 = \bar{Y} - b_1 \bar{X}$$

Donde: $\bar{X}$ y $\bar{Y}$, son las medias correspondientes a X y Y.
S_x^2 es la varianza muestral de X.
S_{xy} es la covarianza muestral entre X y Y.

El coeficiente de regresión (b_1) nos da información muy valiosa sobre el comportamiento de la variable Y frente a la variable X, de manera que:

Si $b_1 = 0$, para cualquier valor de X la variable Y es constante (es decir, no cambia).

Si $b_1 > 0$, esto nos indica que al aumentar el valor de X, también aumenta el valor de Y.

Si $b_1 < 0$, esto nos indica que al aumentar el valor de X, el valor de Y disminuye.

Para garantizar que un modelo de regresión lineal ajustado es válido, debemos comprobar que se cumplen los siguientes supuestos del modelo: linealidad, independencia, homocedasticidad y normalidad.

(Montgomery C., Peck A., & Vining, 2012); (Carollo, 2012)

Comentado [JV7]: Desde la primera vez que se cita una fuente con más de dos autores se usa et al. (Montgomery et al., 2012).

Implementación en el lenguaje R.

En R tenenos la función "$lm()$" para generar una regresion lineal simple. Asignamos a una variable el modelo generado por "$lm()$", es decir:

$$> nombreVariable = lm(Y \sim X)$$

Siendo Y la variable dependiente y X la variable independiente. Utilizamos la función "$summary()$" para ver los detalles del modelo creado con "$lm()$", el argumento de "$summary()$" es la variable a la que asignamos el modelo.

$$> summary(nombreVariable)$$

(Mendoza, 2020)

5.3 Gráfica de dispersión

Para obtener la gráfica de dispersión empleamos el lenguaje R en el entorno de Desarrollo RStudio. Instalamos e importamos la librería "$ggplot2$" para poder emplear la función:

$$> plot(X, Y)$$

Ahora bien, para observar la aproximación (línea recta) del modelo sobre dichas muestras, después de ejecutar la función "$plot$", ejecutamos la función "$abline()$":

$$> abline(modelo)$$

El argumento de la función "$abline()$" es el objeto generado por "$lm()$".

(Mendoza, 2020)

5.4 Prueba de falta de ajuste

Está diseñada para evaluar si una relación curvilineal podría ajustarse mejor a los datos que un modelo lineal. Para ello la SCE (suma de cuadrados de errores) se descompone en dos partes: el componente de error puro y el componente de falta de ajuste. Estos dos componentes son utilizados para construir un estadístico de prueba F particular con el fin de contrastar la hipótesis siguiente:

H_0: la relación es lineal, H_1: la relación no es lineal

La prueba requiere observaciones repetidas en al menos uno de los niveles de X. Las observaciones de X e Y son independientes y se encuentran normalmente distribuidas, además las distribuciones de Y tienen la misma varianza.

$$F^* = \frac{(SCE_R - SCE_F)gl_F}{SCE_F(gl_R - gl_F)}$$

La regla de decisión está dada por:

$$\text{Si } F^* > F^{\alpha}_{c-k,n-c} \text{ rechazar } H_0$$

Donde: c: # de niveles distintos de X
p: # de variables en la ecuación de regresión
n: # de observaciones

Observar que, SCE_F es la suma de cuadrados del error puro. Es común encontrar la siguiente notación: $SCE_F = SSEP, SCE_R = SSE$

Se define entonces, $SSLF = SSE + SSEP$, donde, $SSLF$ es la suma de cuadrados de falta de ajuste. Tenemos entonces,

Comentado [JV8]: Ojo: de acuerdo con las normas APA debemos citar la fuente en cada párrafo, a menos de que al principio indiquen que toda la sección se toma del mismo autor.

$$F^* = \frac{(SCE_R - SCE_F)gl_F}{SCE_F(gl_R - gl_F)} = \frac{\frac{SSE - SSPE}{(n-2)-(n-c)}}{\frac{SSE}{n-c}} = \frac{MSLF}{MSPE}$$

(Vela Peon, 2011)

5.5 Coeficiente de correlación

El coeficiente de correlación r mide la dependencia lineal que existe entre dos variables.

Propiedades:

No tiene dimensión, siempre toma valores en $[-1,1]$.

Si las variables son independientes, entonces $r = 0$.

Si existe una relación lineal exacta entre X y Y, entonces $r = 1$ (relación directa) o $r = -1$ (relación inversa).

Si $r > 0$, esto indica una relación directa entre las variables (es decir, si aumenta X, también aumenta Y).

Si $r < 0$, la correlación entre las variables es inversa (si aumentamos una, la otra disminuye).

Tenemos que: $r = \frac{S_{xy}}{S_x S_y}$

(Carollo, 2012); (Morales, 2011)

Para calcular el coeficiente de correlación en R, tenemos la siguiente función:

$$cor(variableDependiente, variableIndependiente)$$

(Hernandez , 2017)

5.6 Anova

La prueba de ANOVA es utilizada para determinar si hay una relación significativa entre la variable dependiente y la variable independiente. La prueba de ANOVA compara la variación entre los grupos con la variación dentro de los grupos. Si la variación entre los grupos es mayor que la variación dentro de los grupos, entonces hay una relación significativa entre la variable dependiente y la variable independiente.

(Draper & Smith, 1998)

5.7 Supuestos

Para garantizar que la aproximación $\hat{Y} = b_0 + b_1 X$ sea válida, se deben cumplir las siguientes condiciones (supuestos):

(Carollo, 2012)

Comentado [JV9]: ANOVA
Son siglas del análisis de varianza

5.7.1 Linealidad

El supuesto de la linealidad implica que la relación entre la variable independiente X y la variable dependiente Y sea lineal. Esto lo podemos verificar con el test de Pearson. Hipótesis del test:

$$H_0: No\ existe\ correlación\ lineal\ entre\ X\ y\ Y$$
$$H_1: Existe\ correlación\ lineal\ entre\ X\ y\ Y$$

Implementación en R:

$$> cor.test(Y, X, method = "Pearson")$$

X variable independiente y Y variable dependiente

(Chiroque, 2020)

5.7.2 Independencia

Comentado [JV10]: Es un error grave. Independencia no es lo mismo que no correlación. Y lo que se comprueba con Durbin Watson es la no correlación.

Los residuos (errores) deben ser independientes entre sí (no deben estar correlacionados). El contraste de *Durbin-Watson* (D-W) se utiliza para realizar una prueba de autocorrelación AR(1) sobre un conjunto de datos. Este contraste se centra en el estudio de los **residuos de Mínimos Cuadrados Ordinarios** (MCO). Hipótesis del test D-W:

H_0: No existe autocorrelación entre los residuos (los residuos son independientes)

H_1: Existe autocorrelación entre los residuos (los residuos son dependientes)

(Rodo, 2019)

Implementación en R. Instalar e importar la librería "$lmtest$".

Aplicamos el test de D-W:

$$> dwtest(variableDependiente \sim variableIndependiente)$$

(Parra, 2017)

Si el estadístico de D-W se encuentra entre 1.5 y 2.5, entonces podemos concluir que los residuos son independientes, en caso contrario, afirmamos que existe una dependencia entre los mismos.

(Carollo, 2012)

5.7.3 Homocedasticidad

Para cada valor de la variable X, la varianza de los residuos $\varepsilon_i = Y_i - \hat{Y}_i$ debe ser la misma (es decir, que el ajuste es igual de preciso independientemente de los valores que tome X). Para probar dicho supuesto podemos aplicar el test de "*Breusch-Pagan*". Implementación en R.

Hipótesis del test:

$$H_0: Homocedasticidad$$
$$H_1: Heterocedasticidad$$

Instalar e importar la librería "$lmtest$". Enseguida aplicamos el test Breusch-Pagan con la función: $> bptest(nombreDelModelo)$

(Santibañez, 2018)

5.7.4 Normalidad

Para cada variable X, los residuos $\varepsilon_i = Y_i - \hat{Y}_i$ tienen distribución normal con media cero. Para este caso se tiene el test de *Anderson-Darling* (A-D), es una modificación del test de *Kolmogórov-Smirnov*, se diferencia de este ya que da más peso a las colas de la distribución.

(Saes Magdaleno, 2017)

Hipótesis del test A-D:

H_0: Los residuos (errores) se distribuyen normalmente

H_1: Los residuos no se distribuyen normalmente.

Implementación en R. Instalar e importar la librería "$nortest$":

Ahora, se aplica el test de A–D: $> ad.test(nombreDelVectorDeDatos)$

(Zach, 2019)

Comentado [JV11]: La información es completa en relación con la teoría que vimos sobre el modelo de regresión lineal. Se excedieron en el número de cuartillas, deben desarrollar más su capacidad de síntesis.
La penalización en puntos es esencialmente por un error grave de concepto: confundir independencia con no linealidad.

6 Mapa conceptual

Comentado [JV12]: Su mapa tiene los conceptos esenciales de la teoría del modelo de regresión y está organizado jerárquicamente, como debe ser. Solamente hizo falta incluir conectores entre los conceptos.

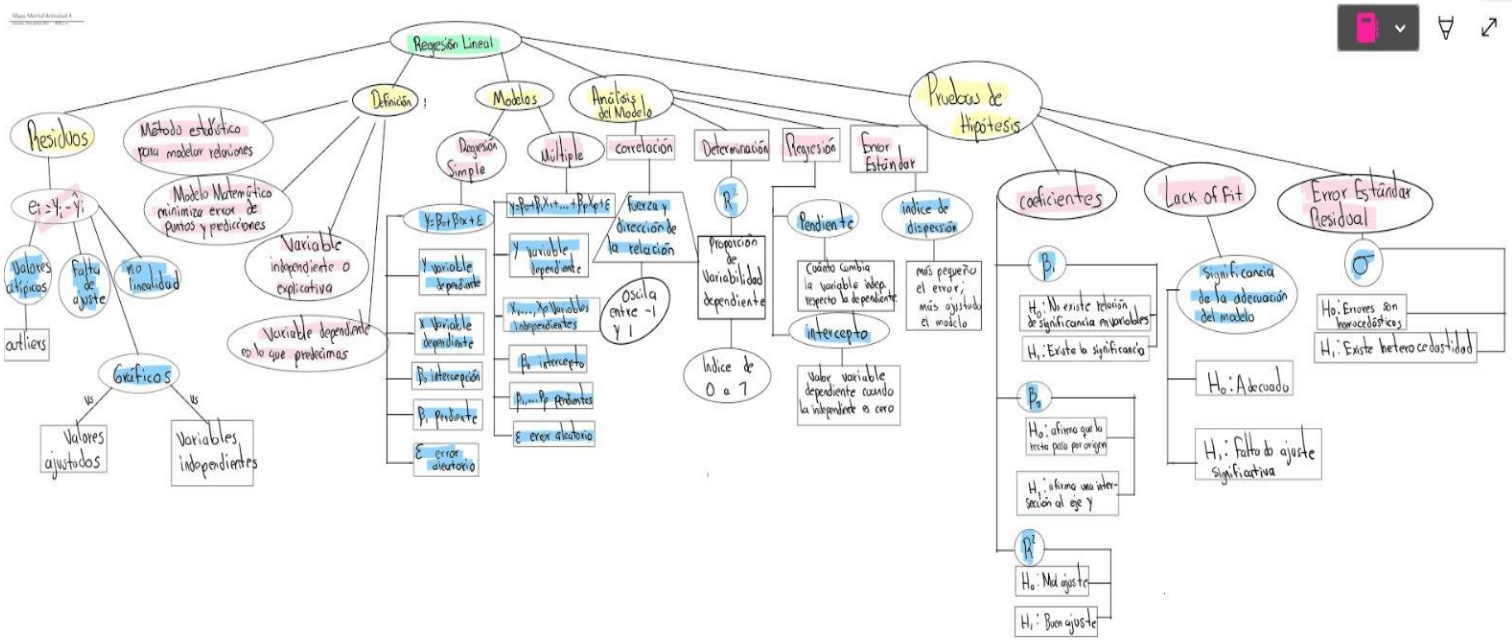


Ilustración 1: Mapa Mental de Regresión Lineal

7 Desarrollo

Compruebe si hay una relación aproximadamente lineal entre los datos de 2014 y los de 2016 y, en su caso, obtenga una ecuación de regresión lineal para modelarla.

Solución ↓

Tenemos las siguientes muestras X (variable independiente) y Y (variable dependiente) correspondientes a los años 2014 y 2016 respectivamente. Los datos están expresados en miles de personas.
Los datos se encuentran en un archivo tipo Excel llamado: $PobrezaMexico.xlsx$

	lugar	x	y
1	Aguascalientes	442.866	369.7
2	Baja California	984.945	789.1
3	Baja California Sur	226.234	175.6
4	Campeche	390.959	405.0
5	Coahuila	885.786	745.9
6	Colima	244.938	248.7
7	Chiapas	3960.988	4114.0
8	Chihuahua	1265.546	1150.0
9	Distrito Federal	2502.468	2434.4
10	Durango	761.244	643.3
11	Guanajuato	2683.282	2489.7
12	Guerrero	2315.421	2314.7
13	Hidalgo	1547.812	1478.8
14	Jalisco	2780.223	2560.6
15	México	8269.852	8230.2
16	Michoacán	2708.631	2565.9
17	Morelos	993.730	965.9
18	Nayarit	488.828	470.1
19	Nuevo León	1022.734	737.8
20	Oaxaca	2662.748	2847.3
21	Puebla	3958.812	3728.2
22	Querétaro	675.679	635.7
23	Quintana Roo	553.015	471.0
24	San Luis Potosí	1338.100	1267.7
25	Sinaloa	1167.066	929.7
26	Sonora	852.081	831.4
27	Tabasco	1169.789	1228.1
28	Tamaulipas	1330.707	1156.2
29	Tlaxcala	745.137	701.8
30	Veracruz	4634.239	5049.5
31	Yucatán	957.908	901.9
32	Zacatecas	819.788	780.3

Ilustración 2: Datos a analizar

El análisis de regresión lineal se hará con el lenguaje R en el entorno de desarrollo RStudio.

7.1 Modelo de regresión lineal

Necesitamos instalar e importar la librearía "$readxl$":

$$> install.packages("readxl")$$
$$> library(readxl)$$

Cargamos las muestras:

$$> PobrezaMexico <- read\_excel("PobrezaMexico.xlsx")$$

Creamos el modelo de regresión lineal con la función "$lm()$":

$$> modeloPM = lm(y \sim x, data = PobrezaMexico)$$

Donde "y" es la variable dependiente, "x" la variable independiente y "$PobrezaMexico$" es la base de datos (donde se encuentran las muestras).
Con la función "$summary()$" podremos observar una descripción general del modelo de regresión creado:

```
> summary(modeloPM)

Call:
lm(formula = y ~ x, data = PobrezaMexico)

Residuals:
    Min      1Q  Median      3Q     Max
-213.53 -102.61    6.28   50.17  419.32

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -93.48277   34.49184   -2.71    0.011 *
x             1.01930    0.01449   70.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 134 on 30 degrees of freedom
Multiple R-squared:  0.994,     Adjusted R-squared:  0.9938
F-statistic:  4945 on 1 and 30 DF,  p-value: < 2.2e-16
```

Ilustración 3: Summary del modelo

Con esto se obtiene la ecuación de regresión lineal

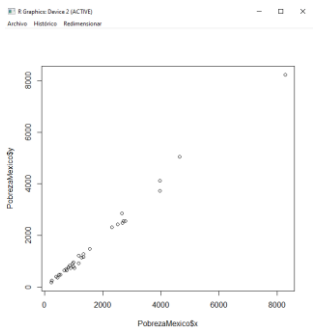
$$Y = B_0 + B_1X + \varepsilon \quad \rightarrow \quad Y = -93.48277 + 1.01930X + \varepsilon$$

Ahora, procedemos a obtener las evidencias estadísticamente significativas para garantizar que nuestro modelo de regresión lineal (aproximación) sea óptimo.

7.2 Gráfica de dispersión

Gráfica de dispersión

```
> plot(PobrezaMexico$x, PobrezaMexico$y)
```



Regresión lineal simple

```
> plot(PobrezaMexico$x, PobrezaMexico$y)
> abline(modeloPM)
```

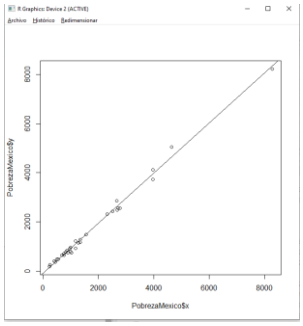


Ilustración 4: Graficas de dispersión en R

En la gráfica de dispersión podemos ver que las observaciones se ajustan satisfactoriamente dentro de una línea recta, pero también vemos que hay un par de datos bastante alejados.

Comentado [JV13]: Bastante es un término ambiguo. La sugerencia es decir exactamente cuáles son esos dos datos y la medida en la cual se alejan de la recta de regresión, destacando que son las que presentan el error más grande.

7.3 Coeficiente de correlación

Se emplea el test de Pearson.

```
> cor.test(PobrezaMexico$x,PobrezaMexico$y)

        Pearson's product-moment correlation

data:  PobrezaMexico$x and PobrezaMexico$y
t = 70.321, df = 30, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9937573 0.9985407
sample estimates:
      cor
0.9969804
```

Ilustración 5: Test de Pearson

Dado que, el coeficiente de correlación es muy cercano a 1 $(cor = 0.9969804)$, concluimos que existe una relación lineal muy fuerte entre las muestras $(X\ y\ Y)$ de pobreza respecto al año 2014 y 2016 en México.

7.4 Prueba de significancia de los coeficientes de correlación

Empleamos el "$t - test$"

```
> t.test(PobrezaMexico$x,PobrezaMexico$y)

        Welch Two Sample t-test

data:  PobrezaMexico$x and PobrezaMexico$y
t = 0.14317, df = 61.97, p-value = 0.8866
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -779.1304  899.3433
sample estimates:
mean of x mean of y
 1729.424  1669.317
```

Ilustración 6: Prueba T-test

Dado los resultados del test anterior, se logra observar que, $p - value = 0.8866 > \alpha = 0.05$. Por ende, concluimos que existe evidencia estadísticamente significativa para rechazar H_0. Entonces, podemos afirmar que las variables son dependientes, es decir, existe una relación lineal entre la variable x y y. Por tanto, la correlación es significativa.

Comentado [JV14]: Tenemos otros errores de concepto graves. Rechazamos la hipótesis nula cuando p-value es menor que alfa.
Por otro lado, el test t que usamos es el que se obtiene con el sumary del modelo, el que está en la Ilustración 3. Como podrán apreciar, ahí se rechaza la hipótesis nula, la cual establece que los coeficientes no son significativos (son cero).

7.5 Significancia de la regresión

Con la función "$summary()$" podemos probar dicha significancia.

```
> summary(modeloPM)

Call:
lm(formula = y ~ x, data = PobrezaMexico)

Residuals:
    Min      1Q  Median      3Q     Max
-213.53 -102.61    6.28   50.17  419.32

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -93.48277   34.49184   -2.71    0.011 *
x             1.01930    0.01449   70.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 134 on 30 degrees of freedom
Multiple R-squared:  0.994,     Adjusted R-squared:  0.9938
F-statistic:  4945 on 1 and 30 DF,  p-value: < 2.2e-16
```

Ilustración 7: Summary y la significancia de regresión

Se logra visualizar que, el $p - value < 2.2e^{-16} < \alpha = 0.05$ respecto a la prueba de F (F-statistic). Por ende, concluimos que existe evidencia estadísticamente significativa para rechazar H_0. Por tanto, podemos afirmar que la regresión es significativa.

7.6 Verificando los supuestos del modelo

7.6.1 Linealidad (Pearson)

Empleamos el test de Pearson.

```
> cor.test(PobrezaMexico$x,PobrezaMexico$y)

        Pearson's product-moment correlation

data:  PobrezaMexico$x and PobrezaMexico$y
t = 70.321, df = 30, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9937573 0.9985407
sample estimates:
      cor
0.9969804
```

Ilustración 8: Prueba de Pearson y linealidad

Dado que, $p - value < 2.2e^{-16} < \alpha = 0.05$, concluimos que existe evidencia estadísticamente significativa para rechazar H_0. Por tanto, podemos afirmar que existe una relación lineal entre las variables x y y, que además tiene un coeficiente de correlación muy bueno (cercano a 1).

7.6.2 Residuos con media cero (t-test)

```
> t.test(modeloPM$residuals)

        One Sample t-test

data:  modeloPM$residuals
t = -1.8107e-16, df = 31, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -46.49278  46.49278
sample estimates:
    mean of x
-4.127775e-15
```

Ilustración 9: t-test para media igual a cero

Dado que, $p - value = 1 > \alpha = 0.05$, no existe evidencia estadísticamente significativa para rechazar H_0. Por lo tanto, vemos que efectivamente, los residuos tienen media igual a cero.

Comentado [JV15]: Aquí sí interpretaron bien el p-value. Es extraño que haya dos interpretaciones diferentes.

7.6.3 Homocedasticidad (Breusch-Pagan)

Empleamos el test de "$Breusch - Pagan$"

```
> bptest(modeloPM)

        studentized Breusch-Pagan test

data:  modeloPM
BP = 6.332, df = 1, p-value = 0.01186
```

Ilustración 10: Prueba de Breusch-Pagan

Dado que, $p - value = 0.01186 < \alpha = 0.05$, concluimos que existe evidencia estadísticamente significativa para rechazar H_0. Por tanto, podemos afirmar que la varianza de los residuos no es constante, es decir nuestro modelo presenta heterocedasticidad, violando así éste supuesto.

Comentado [JV16]: No lleva acento.

7.6.4 Residuos no autocorrelacionados (Durbin-Watson)

Empleamos el test de "$Durbin - Watson$"

```
> dwtest(modeloPM)

        Durbin-Watson test

data:  modeloPM
DW = 2.3369, p-value = 0.8255
alternative hypothesis: true autocorrelation is greater than 0
```

Ilustración 11: Prueba de Durbin-Watson

Dado que, $p - value = 0.8255 > \alpha = 0.05$, concluimos que no existe evidencia significativa para rechazar H_0. Por tanto, podemos afirmar que no existen problemas de autocorrelación entre los residuos (los residuos no están autocorrelacionados).

7.6.5 Residuos normales (Anderson – Darling)

Empleamos el test de "$Anderson - Darling$"

```
> library(nortest)
> ad.test(modeloPM$residuals)

        Anderson-Darling normality test

data:  modeloPM$residuals
A = 0.47705, p-value = 0.2217
```

Ilustración 12: Prueba de Anderson-Darling

Dado que, $p - value = 0.2217 > \alpha = 0.05$, concluimos que no existe evidencia estadísticamente significativa para rechazar H_0. Por tanto, podemos afirmar que los residuos se distribuyen normalmente.

8 Conclusiones

A partir de los resultados obtenidos en la sección anterior, podemos ver que el modelo de regresión lineal que construimos parece explicar bastante bien la relación entre los datos de pobreza de los años 2014 y 2016, pero cuenta con un defecto muy claro: la heterocedasticidad.

Los supuestos del modelo son imprescindibles para asegurar que se trata de un modelo riguroso y bien construido. Esto nos lleva a pensar que no es el modelo más adecuado a la hora de tratar de hacer predicciones, creemos que este problema se debe a los datos casi atípicos que podíamos observar desde el diagrama de dispersión.

Podría ser posible plantear un modelo alternativo utilizando transformaciones en las variables o tal vez utilizando mínimos cuadrados ponderados para tratar de ajustar los residuos y conseguir así un modelo homocedástico.

Comentado [JV17]: ¿Cuáles datos atípicos? No mencionaron esto antes.

9 Referencias

Cabana. (2 de Diciembre de 2021). *Aprende con Eli*. Obtenido de https://aprendeconeli.com/porque-nivel-significacion-005/

Carollo, C. (2012). Regresion Lineal Simple. España, Santiago de Compostela, España. Obtenido de http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf

Chiroque, C. (8 de Diciembre de 2020). REGRESIÓN LINEAL: REQUISITOS. *CURSO DE MACHINE LEARNING EN R STUDIO*. Peru. Recuperado el Abril de 2023, de https://www.youtube.com/watch?v=FGpfhKwsluE&ab_channel=Datapol%C3%ADtica

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis.* Canada: Wiley.

Hernandez , R. (19 de Marzo de 2017). *RPubs*. Obtenido de https://rpubs.com/ronnyhdez/260186

Mendoza, J. (21 de Mayo de 2020). *Estadísticamente*. Obtenido de Programar en R: https://estadisticamente.com/como-hacer-regresion-lineal-en-r/

Montgomery C., D., Peck A., E., & Vining, G. G. (2012). *Introduction to linear regression analysis* . New Jersey: Wiley.

Morales, P. (2011). *El coeficiente de correlacion.* Guatemala: Universidad Rafael Landívar. Obtenido de https://ice.unizar.es/sites/ice.unizar.es/files/users/leteo/materiales/01._documento_1_correlaciones.pdf

Parra, F. (15 de Junio de 2017). *RStudio Pubs*. Obtenido de Estadística y Machine Learning con R: https://rstudio-pubs-static.s3.amazonaws.com/293401_9dcbe5cf99a047e6908db0779e4cafe6.html

Rodo, P. (11 de Noviembre de 2019). *Economipedia*. Obtenido de Contraste de Durbin Watson: https://economipedia.com/definiciones/contraste-de-durbin-watson.html

Saes Magdaleno, L. (2017). *CONTRASTE DE HIPÓTESIS DE DISTRIBUCIONES.* España: Universidad de Cantabria.

Santibañez, J. (2018). *Sigma iimas*. Obtenido de Verificación del supuesto de homocedasticidad: http://sigma.iimas.unam.mx/jsantibanez/Cursos/Ciencias/2018_1/08_homocedasticidad.html

Vela Peon, F. (Octubre de 2011). Prueba de falta de ajuste (Lack-of-fit Test). Ciudad de México, Xochimilco, México.

Zach. (22 de abril de 2019). *Statology*. Obtenido de How to Conduct an Anderson-Darling Test in R: https://www.statology.org/anderson-darling-test-r/

Comentado [JV18]: En la versión vigente de las normas APA ya no se usa obtenido de. En los libros tampoco va ya la ciudad.
Las referencias siempre tienen algún elemento en cursiva.
Revisen las siguientes ligas, son necesarias para que aprendan a elaborar correctamente la referencia de los tipos de fuente usuales:
https://normas-apa.org/referencias/citar-un-blog/
https://normas-apa.org/referencias/citar-libro/
https://normas-apa.org/referencias/citar-revista/

Ilustraciones

García M. Tipos de pobreza. URL: https://www.asociacionproade.org/blog/tipos-de-pobreza/

Galindo A. (19 de octubre de 2015). La pobreza extrema. URL: https://www.las2orillas.co/la-pobreza-extrema-flagelo-de-nuestro-tiempo/

Figuereo J. 6 de abril de 2020. Fin de la pobreza ¿por qué será importante?. URL: https://www.qndiario.com/ods-fin-pobreza

Consecutivo	Lista de cotejo Aspecto por evaluar	Puntaje posible	Puntaje obtenido
Calidad de forma y secciones iniciales del reporte de proyecto			
A0	El reporte está libre de plagio porque incluye citas a fuentes documentales en cada párrafo del marco teórico y tabla automática de referencias, ambas de conformidad con las normas APA.	**Requisito**	
A1	El reporte está correctamente redactado y sin faltas de ortografía.	-0.5/0	
A2	El reporte incluye carátula, tabla de contenido automático, tabla automática de tablas, tabla automática de figuras y numeración de páginas.	-0.5/0	0
A3	El título del proyecto es descriptivo con hasta 12 palabras.	0/0.5	0.5
A4	El reporte de proyecto cuenta con un resumen de no más de 250 palabras que incluye objetivo, metodología, resultados y conclusiones básicas.	0/0.5	0.5
A5	Se incluyen tres palabras clave.	0/0.5	0.5
A6	La introducción del escrito incluye tema, objetivo y estructura del reporte. Responde a las preguntas ¿qué? (concepto de lo que se quiere hacer y antecedentes), ¿por qué? (problema a resolver) y ¿para qué? (objetivo del proyecto).	0/0.5	0
Metodología			
A7	El reporte de proyecto incorpora un marco teórico de máximo tres cuartillas, con una redacción propia del estudiante que articula coherentemente todos los conocimientos requeridos para la elaboración de la actividad.	0 a 2	1
A8	El marco teórico finaliza con un mapa conceptual que abarca todos los contenidos teóricos de la actividad.	-0.5/0	0
Conclusiones			
A9	Hay una conclusión por cada resultado presentado y en conjunto responden al objetivo del proyecto.	0/1	0.8
Dominio del contenido			
A10	El desarrollo del proyecto refleja el dominio del estudiante sobre los aspectos conceptuales y de aplicación de los temas explorados en la actividad.	0 a 4	3
A11	El desarrollo del proyecto se apoya en el software R en todos los aspectos estadísticos incluidos.	0 a 1	1