Aplicación de regresión lineal a La Población con Empleo del Ecuador (ENEMDU).

Application of linear regression to the Population with Employment of Ecuador (NSEUU).

Erik Arnold Pastor Fuentes.
Terry Cristopher Jara Romero.
Universidad de Guayaquil

Resumen

En el presente artículo se ha utilizado la aplicación de regresión lineal para obtener ecuaciones pronostico, las cuales nos ayudaran a realizar aproximaciones en un **x** trimestre o año, dar como resultado el número aproximado de habitantes que se encontraran con empleo (no incluye tareas informales), empleando los datos obtenidos desde el año 2014 hasta el primer trimestre del 2018, estos datos fueron extraídos del ENEMDU, INEC.

Palabras clave: ENEMDU (Encuesta nacional de Empleo, Desempleo y Subempleo.), INEC (Instituto Nacional de Estadística y Censo), Regresión lineal.

Abstract

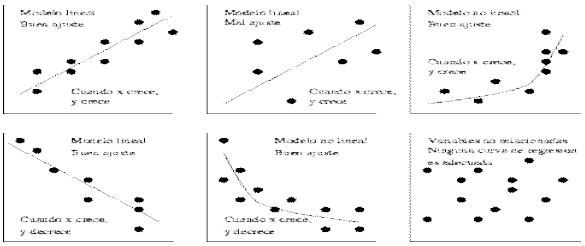
In the present article, the linear regression application has been used to obtain forecast equations, which will help us to make approximations in a x quarter or year, resulting in the approximate number of inhabitants who find employment (does not include informal tasks) , using the data obtained from 2014 until the first quarter of 2018, these data were extracted from the ENEMDU, INEC.

Key words: National Survey of Employment, Unemployment and Underemployment (NSEUU), National Institute of Statistics and Census (NISC), Linear regression.

Introducción

La regresión lineal es un método el cual permite determinar modelos matemáticos para poder realizar cierto tipo de proyecciones o aproximaciones, en función de la relación entre una variable dependiente **y,** una variable independiente **x**, el término regresión se utilizó por primera vez en el estudio de variables antropométricas, al comparar la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio, tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, "regresaban" al promedio, la constatación empírica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno. El término lineal se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática ya sea esta cuadrática, polinomial, variables múltiples, funciones conocidas.

Figura 1. Tipos de relación entre 2 variables.



Los modelos lineales son una explicación simplificada de la realidad, mucho más ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadística. El modelo de regresión lineal es aplicado en un gran número de campos, desde el ámbito científico hasta el ámbito social, pasando por aplicaciones industriales ya que en multitud de situaciones se encuentran comportamientos lineales, las diferentes aplicaciones en las que se puede ver inmersa la regresión lineal: en física, química, producción, estudios de población, etc. ya que cada uno de estos se puede representar mediante una función que involucre a diferentes puntos predispuestos en un diagrama de dispersión.

¿Cuándo utilizar la regresión lineal?

La regresión lineal es un modelo óptimo para cierto tipo de patrones que presenten tendencia (creciente o decreciente), es decir en pocas palabras, patrones que presenten una relación de linealidad entre la variable dependiente **(y)** y el tiempo (variable independiente **(x)**. Se debe tener muy en cuenta la gráfica de dispersión de los datos si estos nos indican una relación de linealidad entre los puntos de dispersión, esto será un gran indicador de que el método de regresión para obtener una ecuación pronóstico de esos puntos es una regresión lineal. Es muy importante tener en cuenta el tipo de regresión a utilizar lo cual va a depender del comportamiento de los datos dados en el diagrama de dispersión, ya que de una u otra manera si se aplica un tipo de regresión la cual no es la indicada para un grupo de datos el error en el momento de evaluar las aproximación se va a incrementar y por ende el coeficiente de correlación medido en los estándares dado nos va a indicar que se ha utilizado un tipo de regresión errónea en el caso de la lineal.

En este artículo se va a describir de manera minuciosa el análisis de la regresión en donde están involucradas una variable dependiente (**y**), y una variable independiente (x), en donde existirá una relación entre ellas la cual se va a representar mediante una línea recta que será un ecuación (pronostico) indicada bajo el formato de la pendiente de una recta en este estudio aplicaremos la regresión lineal en base a los datos de la población que

Tabla1. Datos de población con empleo en Total Nacional, Urbano y Rural

	Población con Empleo																	
AÑO(X)	2014	2014,3	2014,6	2014,9	2015	2015,3	2015,6	2015,9	2016	2016,3	2016,6	2016,9	2017	2017,3	2017,6	2017,9	2018	2018,3
TOTAL NACIONAL	6.664.241	6.706.314	6.643.458	6.866.776	6.921.107	7.091.116	7.098.584	7.274.221	7.140.636	7.412.671	7.415.099	7.637.986	7.463.579	7.728.968	7.781.560	7.842.471	7.712.177	7.802.374
URBANO	4.481.130	4.501.505	4.529.978	4.638.310	4.647.582	4.630.745	4.707.715	4.854.005	4.840.314	4.882.929	4.889.895	5.005.457	4.971.669	5.048.482	5.125.446	5.174.135	5.169.942	5.129.893
RURAL	2.183.111	2.204.809	2.113.480	2.228.466	2.273.525	2.460.371	2.390.869	2.420.216	2.300.322	2.529.742	2.525.203	2.632.529	2.491.910	2.680.487	2.656.114	2.668.336	2.542.236	2.672.481

se encuentra con empleo en el Ecuador donde tenemos datos del Total Nacional este total también aparece dividido en Rural y Urbano ,en base a estos datos se procederá a encontrar y establecer la ecuación pronóstico para poder realizar proyecciones para saber de manera aproxima la población que se va a encontrar con empleo en un año o trimestre del algún años que no se encuentre en la base de datos provista es decir con la ecuación pronostico se encontrara un valor de **y** en función de **x**.

Aplicación del modelo de regresión lineal.

Con el propósito de entender y aplicar este método en relación al tema a tratar comenzó con las investigaciones pertinentes para recolectar información clara y verídica de los datos de población con empleo del Total Nacional, Urbano y Rural en fuentes oficiales desde el año 2014 hasta el primer trimestre del 2018, es importante tener una gran cantidad de datos para que no se altere el diagrama de dispersión, por tal motivo se tomaron datos trimestrales (3meses) del periodo de tiempo determinado, con esta base de dato mayor facilita tomar la decisión del método de regresión a utilizar, toda esta información fue proporcionada por instituciones estatales ENEMDU (Encuesta nacional de Empleo, Desempleo y Subempleo), INEC (Instituto Nacional de Estadística y Censo), que nos proporcionar datos veraces, para así tener certeza de los resultados previos a realizar mediante los cálculos pertinentes, como se los puede apreciar en las siguientes tablas con sus respectivas graficas de dispersión.

Con los datos mostrados en las tablas, se va a establecer una función o ecuación matemática pronostico la cual se va ajustar a los datos indicados y va a describir la relación entre las variables por medio de una regresión de cada una de las tablas.

Existen 3 puntos claves al momento de realizar el análisis de regresión estos son:

- Decidir qué clase de curva describen los puntos en una gráfica.
- De acuerdo a la gráfica determinar el tipo de ecuación que mejor se ajuste a los datos.
- Encontrar la ecuación pronóstico, y verificar datos de proximidad.

Figura 2. Grafica de dispersión T. Nacional

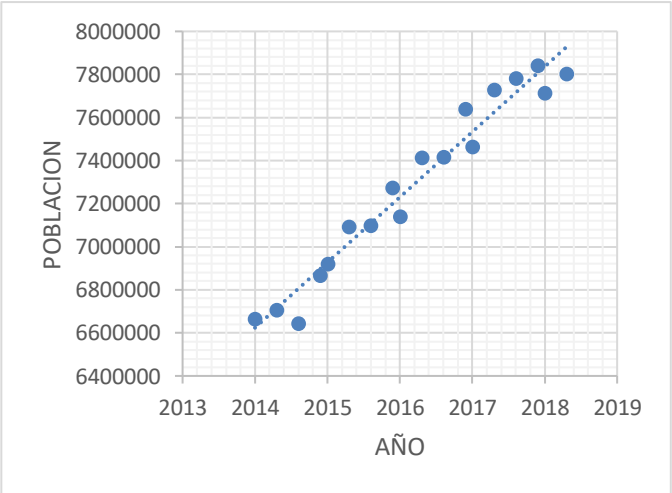


Figura 3. Grafica de dispersión Urbano

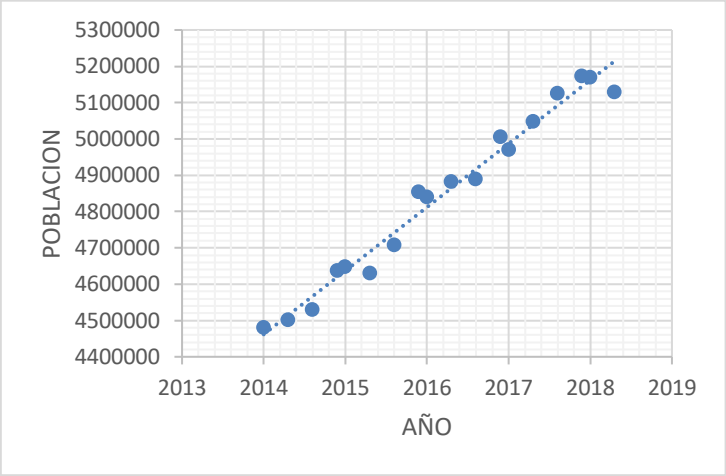
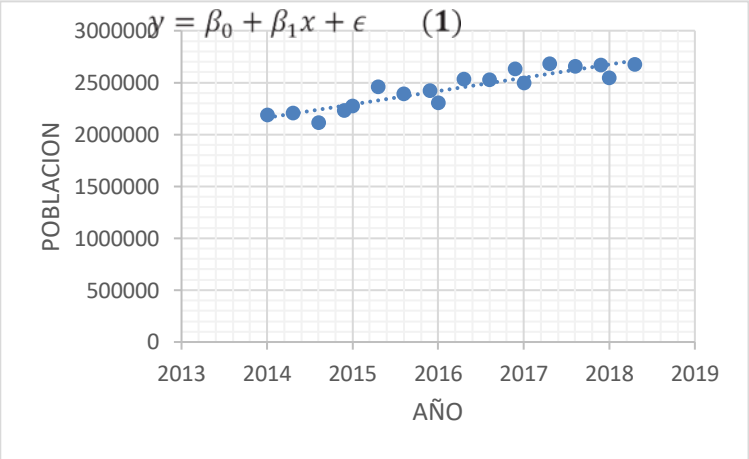


Figura 4. Grafica de dispersión Rural



En el análisis realizado y de acuerdo a lo ya dicho y lo observado en la **Figura 2,3,4 .** En la cual se puede apreciar el comportamiento de los datos en la gráfica desde el 2014 hasta el primer trimestre del 2018 y lo establecido en la **Figura 1.** Se puede determinar un comportamiento lineal.

Modelo de regresión lineal:

$$\hat{y} = \alpha + \beta x \ (1)$$

En donde $\boldsymbol{\alpha}$ estará representada de la siguiente manera:

$$\alpha = \bar{y} - \beta * \bar{x} \qquad \textbf{(F1)}$$

$\bar{y} =$ Promedio de variable dependiente.

$\bar{x} =$ Promedio de variable independiente

En donde $\boldsymbol{\beta}$ estará representado de la siguiente manera:

$$\boldsymbol{\beta} = \frac{\boldsymbol{\Sigma x * y + \Sigma x * \Sigma y}}{\boldsymbol{\Sigma x^2 + (\Sigma x)^2}} \qquad \textbf{(F2)}$$

Los valores de los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ ecuación **(1)** no se conocen y deben de estimarse a partir de los datos de la muestra obtenida, estos coeficientes se calculan con valores conocidos y se los conoce como regresores.

Para el valor de los regresores se utiliza el método fundamentado en teorema los mínimos cuadrados, este método emplea los datos de la muestra (población) para determinar características de la recta que van hacer mínima la suma de los cuadrados de las desviaciones.

$$\textbf{min}\boldsymbol{\Sigma}(\boldsymbol{y_i - \hat{y}})\text{^}\boldsymbol{2} \ \textbf{(2)}$$

En dónde;

$\boldsymbol{y_i} =$ Valor observado de la variable dependiente para la i-esima.

$\hat{\boldsymbol{y}} =$ Ecuación pronostico determinada de tablas de datos.

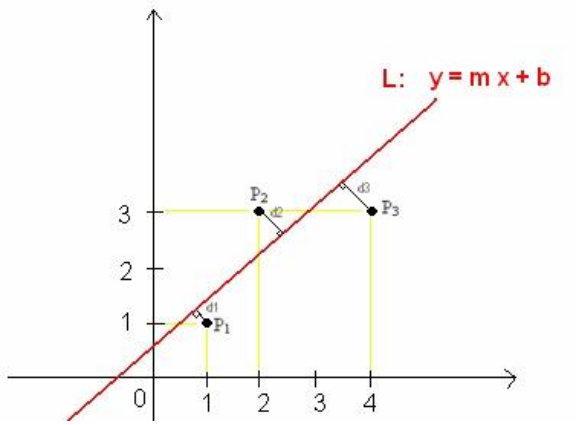
Reemplazando la ecuacion pronostico **(1)**, en **(2)**.

$\boldsymbol{\Sigma(y_i - (\alpha + \beta x))\text{^}2}$ **(3)** esta ecuacion nos ayudara a determinar el error en funcion de las variables x, y de la funcion pronostico, para que la funcion pronostico represente el total de datos de manera much mas representatva se le debe de sumar el error

que se comete en la aproximacion, este error no es mas que la distancia desde cualquie punto de la grafica hacia la recta (d1,d2,d3), como se presenta en la **Figura 5.** La ecuacion (3) nos permite minimizar el miembro de la ecuacion para esto se debe calcular las derivadas parciales de esta expresion respecto a cada uno de los coeficientes de regresion es decir hay que derivar con respect a α y β e igualar a cero cada una de las derivadas parciales, realizado este procedimiento obtendremos un Sistema de ecuaciones que puede ser representado como un Sistema matricial. De la siguiente manera:

$$\Sigma yi = n\alpha + \beta xi$$
$$\Sigma xiyi = \alpha * \Sigma xi + \beta xixi$$

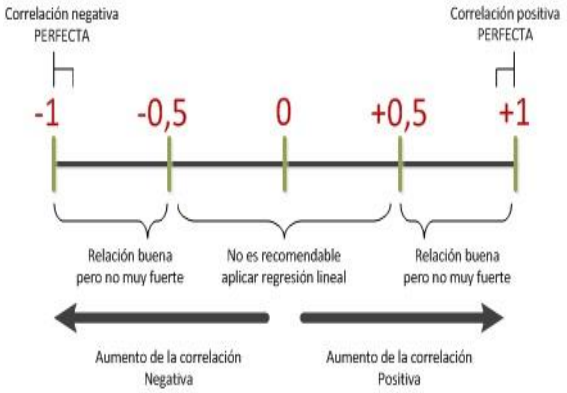
Figura 5. Representacion del error en la grafica.



Coeficiente de correlacion.

El coeficiente de correlacion, es una medida que nos va a indicar el grado de asociacion de los datos de las variables(x,y), esta medida nos va a indicar el tipo de relacion o dependencia, con este coeficinte tambien podemos determinar si el metodo de regresion que usamos es el correcto de acuerdo a la escala ya dada **Figura 6.** Si la correlacion esta entre (-0,5;0,5) esto nos indica que existe una correlacion, en la cual no es recomendable aplicar regresion lineal esto quiere decir que la ecuacion pronostico no se ajusta de manera adeacuada a los datos dispersos en el diagrama de dispercion.

Figura 6. Escala de aceptacion del coeficiente de correlacion.

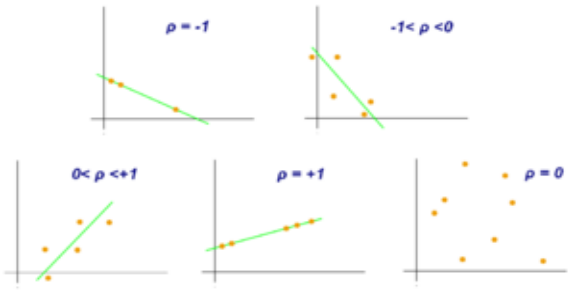


Fórmula para calcular el coeficiente de correlación.

$$r = \frac{\Sigma(x - \hat{x})(y - \hat{y})}{\sqrt{\Sigma(x - \hat{x})} * \sqrt{\Sigma(y - \hat{y})}}$$

Coeficiente de determinación.

El coeficiente de determinación se lo simboliza con la letra r^2 y no es más que el coeficiente de correlación al cuadrado, lo que el coeficiente de determinación nos indica en la regresión lineal, es probar cierto tipo de hipótesis, este coeficiente ayudara a determinar la calidad del modelo (ecuación



pronóstico), para replicar resultados, y la proporción de variación de los resultados.

Procedimiento para encontrar las ecuaciones pronostico.

Para encontrar las ecuaciones pronósticos respecto a URBANO, RURAL y el TOTAL DE POBLACION que es la sumatoria de ambos respectivamente y con el cual se va a comparar la sumatoria de las ecuaciones pronóstico de RURAL, URBANO respecto al TOTAL NACIAONAL y determinar la dispersión de los resultados que en este caso se lo podrá determinar como un error entre sus partes y el total, así tendríamos aplicando la ecuación **(1),** y encontrando sus subtérminos α, β **(F1), (F2),** respectivamente se procederá a encontrar las ecuaciones pronóstico $\hat{y}$ de cada categoría:

- **Ecuaciones pronostico (RURAL).**

$$\boldsymbol{\alpha} = -255885317{,}1$$
$$\boldsymbol{\beta} = 128126{,}694$$

Reemplazando en ecuación **(1)**, tenemos;

$$\hat{y}\mathrm{R} = -255885317, +128126{,}694(\mathrm{x})$$

Tabla 2. Coeficientes de regresión e intervalos de confianza (Rural).

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	-255885317	28788655,9	-8,88840791	1,3785E-07	-316914541	-194856093	-316914541	-194856093
Variable X 1	128126,694	14278,7071	8,97326999	1,2129E-07	97857,1871	158396,201	97857,1871	158396,201

- **Ecuación pronostico (URBANO).**

$$\boldsymbol{\alpha} = -348136219{,}9$$
$$\boldsymbol{\beta} = 175073{,}5321$$

Reemplazando en ecuación **(1)**.
Tenemos;

$$\hat{\boldsymbol{y}}\mathrm{U} = -348136219{,}9 + 175073{,}5321(x)$$

Tabla 3. Coeficientes de regresión e intervalos de confianza (Urbano). v

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	-348136220	13925907,6	-24,9991764	2,9943E-14	-377657825	-318614615	-377657825	-318614615
Variable X 1	175073,532	6907,02462	25,3471707	2,4125E-14	160431,294	189715,77	160431,294	189715,77

Ecuación pronostico (TOTAL POBLACION).

$$\boldsymbol{\alpha} = -604021371{,}2$$

$$\boldsymbol{\beta} = 303200{,}1438$$

Reemplazando en ecuación **(1)**, tenemos;

$$\hat{y}\mathrm{N} = -604021371{,}2 + 303200{,}143(x)$$

Tabla 4. Coeficientes de regresión e intervalos de confianza (Total Nacional)

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	-604021371	32580998,6	-18,5390687	3,0713E-12	-673090003	-534952739	-673090003	-534952739
Variable X 1	303200,144	16159,6477	18,7627942	2,5553E-12	268943,221	337457,066	268943,221	337457,066

Como se describió anteriormente estas ecuaciones pronósticos no representan el muestral total debido a que cada una de estas ecuaciones pronostico se ajustan en una línea recta de tal manera que la distancia entre la recta y cualquier punto se aproximadamente igual y eso se lo determina como un error, entonces para que la ecuación pronostico represente de manera mucho más exacta las proyecciones se les debe de sumar el error cometido.

Factor de correlación.

El factor de correlación de cada uno de las categorías et dada por:

$$\boldsymbol{r} = \frac{\boldsymbol{\Sigma(x - \hat{x})(y - \hat{y})}}{\sqrt{\boldsymbol{\Sigma(x - \hat{x})}} * \sqrt{\boldsymbol{\Sigma(y - \hat{y})}}}$$

- **URBANO.**

$$r = 0{,}987776078$$

Tabla 5. Estadísticos de la regresión Urbano.

Coeficiente de correlación múltiple	0,98777608
Coeficiente de determinación R^2	0,97570158
R^2 ajustado	0,97418293
Error típico	38113,6631
Observaciones	18

- **RURAL.**

$$r = 0{,}913362204$$

Tabla 6. Estadísticos de la regresión Rural.

Coeficiente de correlación múltiple	0,9133622
Coeficiente de determinación R^2	0,83423052
R^2 ajustado	0,82386992
Error típico	78791,3555
Observaciones	18

- **TOTAL NACIONAL.**

$$r = 0{,}978021857$$

Tabla 7. Estadísticos de la regresión Total Nacional.

Coeficiente de correlación múltiple	0,97802186
Coeficiente de determinación R^2	0,95652675
R^2 ajustado	0,95380968
Error típico	89170,5765
Observaciones	18

De acuerdo a la regla de correlación el resultado obtenido en los factores de correlación calculada nos determina que existe una correlación en un intervalo entre 0,5 y 1 los que indica que existe una correlación positiva muy fuerte, lo que nos indica que la fuerza de correlación de los datos es muy buena y nos asegura que el método que se ha empleado para realiza la regresión ha sido el correcto por ende el error que se ha obtenido nos corrobora toda la información previa ya que son cifras que indican un error muy depreciable.

Errores.

- **Total nacional.**

$$\Sigma(y_i - (-604021371{,}2 + 303200{,}143(x))$$

Error Generado= -1,07847E-06

- **RURAL.**

$$\Sigma(y_i - (-255885317{,}1 + 128126{,}694(x))$$

Error Generado= 9,00E-12

- **URBANO**.

$$\Sigma(y_i - (-348136219{,}9 + 175073{,}5321(x))$$

Error Generado= 1,86265E-09

Comparación de y pronóstico.

Como se determinó anteriormente el total nacional de población con empleo es igual a la suma de la población rural más la población urbana con empleo, teniendo la fórmula de pronóstico de los tres casos la relación entre el y pronóstico de la población total tiene que ser igual o casi igual a la suma de la ecuación del y pronostico del rural más el y pronostico del urbano, siendo así se presenta una función de funciones.

Ecuación a

$$F(\hat{y}\text{T. RU}) = F(\hat{y}\text{R}) + F(\hat{y}\text{U})$$

$$F(\hat{y}\text{T. RU}) = \left(-255885317, +128126{,}694(\text{x})\right) + \left(-348136219{,}9 + 175073{,}5321(x)\right)$$

$$F(\hat{y}\text{T. RU}) = -604021536{,}9 + 303200{,}226(\text{x})$$

Ecuación b

$$\hat{y}\text{N} = -604021371{,}2 + 303200{,}143(x)$$

Teniendo las dos funciones una del y pronostico total nacional (ecuación b) y la otra la suma de las funciones del y pronostico rural más el y pronostico urbano (ecuación a) podemos determinar que son muy semejantes entonces la relación es muy fuerte, en caso de querer calcular el total nacional de un x años o trimestre con cualquiera de las dos fórmulas es confiable hacerlo.

Conclusiones.

Utilizando la regresión lineal se puede ajustar una curva a través de una ecuación la cual se la conoce como pronóstico, respecto a un conjunto de datos los cuales describan un comportamiento con tendencia a ser una línea recta, involucrando el estudio de 2 variables cuantitativas (x,y), para lo cual se debe de encontrar un modelo matemático que relaciona una variable dependiente con una variable independiente, y de esta manera poder conocer la relación y la fuerza de cohesión de un conjunto de datos como en este caso de estudio.

Además es de mucha importancia para el estudio realizado que el coeficiente de correlación aborde un valor el cual se encuentre entre los parámetros ya establecidos y que se situé entre los valores de aceptación ya que este indicador muestra que el modelo matemático de regresión que se ha usado para el estudio de la población es el correcto y nos garantiza de cierta manera que los márgenes de errores no serán de gran ponderación para realizar el respectivo análisis de regresión de la POBLACIÓN CON EMPLEO en el Ecuador.

Donde se procederán a presentar variables de respuesta, y así poder analizar y predecir valores de la variable dependiente y evaluar el grado de relación entre las variables con el coeficiente de correlación. El uso de la regresión lineal nos va a permitir realizar proyecciones a futuro en función del tiempo de manera aproximada, respecto a cuál podría ser el número de personas que se encontraran con empleo dentro de cierto parámetro determinado mediante el modelo matemático encontrado a través de la regresión.

En este caso de estudio se ha divido a la distribución de la población en 2 partes:

- URBANO.
- RURAL.

Encontrando la y pronóstico para cada uno de estos sectores y poder realizar estimaciones de manera individual si así se lo dispone en cierto caso. Se ha manejado también un TOTAL NACIONAL el cual es la sumatoria de toda de los 2 sectores anteriores RURAL+URBANO, dando como resultado un TOTAL NACIONAL, el cual permitirá realizar estimaciones a nivel nacional y ya no por sectorización como en la situación anterior.

Referencias.

CANALE, C. (s.f.). *METODOS NUMERICOS PARA INGENIEROS* . McGRAW HILL.

FAIRES, R. L. (Junio 2003). *ANALISIS NUMERICO.*

G., S. (s.f.). *INTRODUCTION TO APPLIED MATHEMATHICS.* ED. Wesllesley Cambridge press.

Sánchez, A. N. (2002). *Métodos Numéricos Aplicados a la Ingeniería.*

Spiegel, M. R. (2010). *Probabilidad y estadística.*

Lista de Tablas.