MODELO DE REGRESIÓN ORTOGONAL EN PRONÓSTICOS Y CORRECCIÓN DE DATOS ATÍPICOS EN HIDROLOGÍA

MODELO DE REGRESIÓN ORTOGONAL EN PRONÓSTICOS Y CORRECCIÓN DE DATOS ATÍPICOS EN HIDROLOGÍA

ORTHOGONAL REGRESSION MODEL IN FORECASTS AND CORRECTION OF ATYPICAL DATA IN HYDROLOGY

Mario Soto
mariosoto456@gmail.com
Universidad Autónoma Tomás Frías. Potosí, Bolivia.

RESUMEN

El modelo de regresión ortogonal (MRO), es un método estadístico, muy aplicado el área de la ingeniería, contrariamente poco difundido en el área de ciencias sociales, donde los errores de medición son comunes en ambas variables. El tipo de investigación es correlacional, la investigación se aplicó en el campo en hidrología cuyo datos se obtuvieron por medio de simulación de dos estaciones X y Y, cuyos caudales son medidos en m^3/s. El objetivo es estudiar el modelo de regresión ortogonal, para pronosticar datos faltantes y corregir datos atípicos o anómalos que se ajusten adecuadamente al modelo.

El modelo de regresión ortogonal permitió corregir los datos atípicos, pronosticar los datos faltantes y mejorar el coeficiente de correlación que fue de $R = 0.35$ con un coeficiente de determinación $R^2 = 0.1225$ a $R = 0.93$ con un coeficiente de determinación de $R^2 = 0.87$

PALABRAS CLAVES

Atípico, coeficiente de determinación, correlación, pronostico, regresión ortogonal

ABSTRACT

The orthogonal regression model (ORM) is a statistical method, widely applied in the area of engineering, contrary to little spread in the area of social sciences, where measurement errors are common in both variables.

The type of research is correlational, the research was applied in the field of hydrology, where data were obtained by simulating two stations X and Y, whose flows are measured in m3/s. The objective is to study the orthogonal regression model, to predict missing data and correct atypical or anomalous data that fit the model adequately.

The orthogonal regression model allowed correcting atypical data, predicting missing data and improving the correlation coefficient, which was from $R = 0.35$ with a coefficient of determination $R^2 = 0.1225$ to $R = 0.93$ with a coefficient of determination of $R^2 = 0.87$.

KEY WORDS

Atypical, coefficient of determination, correlation, forecast, orthogonal regression

I. INTRODUCCIÓN

Para que una investigación tenga resultados satisfactorios en todas las áreas del conocimiento, es importante acudir a la Estadística, mediante ella recopilar información confiable y analizar estadísticamente toda esa información. Pero en muchos estudios o investigaciones, por diferentes factores no se cuenta con toda la información o faltan datos, pero también se presentan anomalías en los datos que se denominan datos atípicos.

El modelo de regresión ortogonal, conocida también como regresión de Deming, examina la relación lineal entre dos variables cuantitativas sean estas discretas o continuas, una variable "*Y*" de respuesta y otra variable "*X*" denominado predictor, ambas variables en el modelo de regresión ortogonal contienen error de medición. (Recio, 2021)

La aplicación del modelo de regresión ortogonal hace que el investigador pueda obtener un modelo más confiable que muestre la verdadera relación entre las variables estudiadas, lo que conduce a mejores predicciones y análisis. (García, 2017)

Hidrología (del griego hydor-, agua) es la disciplina científica dedicada al estudio de las aguas de la Tierra, incluyendo su presencia, distribución y circulación a través del ciclo hidrológico, y las interacciones con los seres vivos. También trata de las propiedades químicas y físicas del agua en todas sus fases. (Velez, 2000)

En la actualidad la hidrología tiene un papel muy importante en el planeamiento del uso de los Recursos Hidráulicos, y ha llegado a convertirse en parte fundamental de los proyectos de ingeniería que tienen que ver con suministro de agua, disposición de aguas servidas, drenaje, protección contra la acción de ríos y recreación. (Gutierrez, 2014)

El objetivo es estudiar el modelo de regresión ortogonal, como un método adecuado para corregir datos anómalos o atípicos, que se presentan en las variables cuantitativas en estudio, estos datos si no son corregidos afectan enormemente a los resultados que generalmente son desastrosos o sesgados.

II. MATERIALES Y METODOS

El modelo de regresión ortogonal llamado también como "mínimos cuadrados totales" o "regresión de Deming", la regresión ortogonal *mal llamada correlación ortogonal*, examina la relación lineal entre dos variables continuas: una predictor X y otro de respuesta o explicativa Y, este método considera el error en la medida de ambas variables, mientras que regresión lineal tiene en cuenta únicamente los errores de la variable respuesta o explicativa Y. Esta característica justifica en qué casos emplear una u otra: si en la nuestra el error sólo afecta en las Y, aplicaremos regresión lineal, si el error afecta también a la X, emplearemos la regresión ortogonal. (Soto, 2022)

En las siguientes figuras, se muestran las soluciones de la regresión lineal y la ortogonal para un mismo conjunto de puntos. Nótese como pueden llegar a diferir ambas soluciones, de ahí la importancia de la correcta aplicación de uno u otro método en función de cada caso.

Figura 8.6

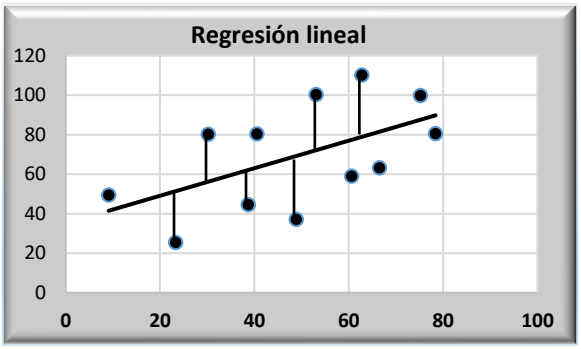
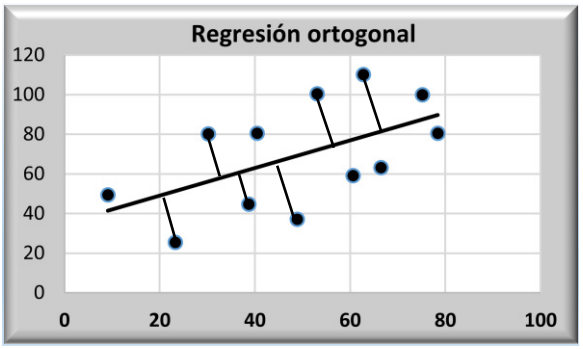


Figura 8.7



En este método el intervalo de confianza en torno a la recta de regresion es $2\sqrt{\lambda_1}$ es decir que más del 95% de los datos de la muestra se encuentran dentro de este intervalo de confianza o banda de seguridad, marcando el nivel de significancia del 5% en torno a la recta de regresión.

Es importante señalar que en MRO el coeficiente de correlación es acepatable cuando es mayor o igual a 0.8. En caso de que no se alcanse este valor, se debe eliminar los valores que caen fuera del intervalo de confianza o banda de seguridad, deviendo realizar los calculos nuevamente hasta alcanzar o superar el coeficiente señalado.

Este metodo se utiliza pricipalmente en hidrologia, pero su aplicación no solo se limita a este campo, tambien se aplica en la medicina, en estudios sociales, turismo, etc.

Las formulas siguientes que son analogas a las anteriores presentadas en el metodo de regresión lineal (MRL) son utilizadas en el metodo de regresión ortogonal (MRO):

Coordenadas del centro de gravedad

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

Varianza

$$\sigma_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n} \quad ; \quad \sigma_Y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{Y})^2}{n} \tag{2}$$

Covarianza

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n} \tag{3}$$

Coeficiente de correlación

$$R = r = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \quad ; \quad -1 \le R \le 1 \tag{4}$$

- **Ecuaciones operativas del completado de datos**

Para obtener la ecuación de linea recta, donde la suma de las distancias de cada punto de coordenadas x_i, y_i sea mininma, se aplica el metodo de los mínimos cuadrados, obteniendo una ecuación de segundo grado, la que permite obtener dos raices que verifican la relación $\lambda_2 > \lambda_1 > 0$

La ecuación del segundo grado es:

$$\lambda^2 - (\sigma_X^2 + \sigma_Y^2)\lambda + (\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2) = 0 \tag{5}$$

Ecuación de la linea recta

$$Y - \bar{Y} = m(X - \bar{X}) \Rightarrow Y = m(X - \bar{X}) + \bar{Y} \tag{6}$$

Pendiente

$$m = b_1 = \frac{\sigma_{XY}}{\lambda_2 - \sigma_Y^2} \tag{7}$$

Entonces la ecuación de la linea recta ortogonal resulta

$$Y = \frac{\sigma_{XY}}{\lambda_2 - \sigma_Y^2}(X - \bar{X}) + \bar{Y} \quad ; \quad \text{luego} \quad Y = m(X - \bar{X}) + \bar{Y} \Rightarrow Y = \underbrace{\bar{Y} - m\bar{X}}_{b_0} + mX$$

Como $b_0 = \bar{Y} - m\bar{X}$ entonces $Y = b_0 + mX$ como la pendiente es $b_1 = m$, se obtiene el siguiente modelo:

$$Y = b_0 + b_1 X \tag{8}$$

En el caso, donde la linea de regresión corta el eje de las X en un punto $X_0 > 0$, para todos los puntos $X < X_0$, los valores de Y seran negativos. Para resolver problema, se recomineda utilizar la siguiente expresión para el llenado de datos faltantes $Y = \bar{Y}\left(\frac{X}{\bar{X}}\right)^{\left(\frac{m\bar{X}}{\bar{Y}}\right)}$

- **Fórmula para resolver la ecuación de segundo grado**

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{9}$$

2.1 Datos faltantes, valores ausentes o missing values

Los datos faltantes, conocido también como valores faltantes, son aquellos que no se almacena en una determinada variable de interés, como ser: errores en la transcripción de datos, resistencia a responder ciertas preguntas en una encuesta. Es un problema común en las investigaciones y puede afectar significativamente en los resultados. (Dagnino, 2014)

2.2 Datos atipicos

Los datos atípicos son observaciones cuyos valores son muy diferentes a las observaciones del mismo grupo de datos, extrañamente son observaciones muy grandes o muy pequeños. Estos datos pueden tener un efecto exagerado o desproporcionado en los resultados estadísticos, como la media que puede conducir a interpretaciones sesgadas. Los datos atípicos son originados por: errores de procedimiento y acontecimientos extraordinarios. (Perez, 2019)

Los datos se obtuvieron mediante simulación para dos estaciones que lo denominamos X y Y, cuyos caudales son medidos en m^3/s en el departamento de Potosí, para el procesamiento de los datos se ha utilizado los siguientes paquetes EXCEL y MINITAB.

III. RESULTADOS

Los datos hidrologicos representan a caudales medios anuales medidos en m^3/s que se obtuvieron mediante simulación, para dos Estaciones X y Y con datos faltantes y datos atipicos como se muestra en la siguiente tabla:

El objetivo en primera instanacia es estimar o pronosticar los datos faltantes con el modelo de regresión ortogonal y posteriormente corregir los datos atipicos empleando el Modelo de Regresión Ortogonal.

Tabla 1. Caudales medios anuales (1999-2018)

Obs.	*Año*	*X*	*Y*	
1	1999	5.3	4.15	
2	2000	3.8	3.25	
3	2001	¿?	4.6	***Se anula***
4	2002	12.95	2.17	
5	2003	8.45	6.32	
6	2004	4.95	¿?	***Se anula***
7	2005	8.82	9.12	
8	2006	4.65	5.38	
9	2007	3.05	14.17	
10	2008	2.49	3.87	
11	2009	7.3	5.35	
12	2010	12.28	10.64	
13	2011	10.63	10.87	
14	2012	9.7	10.98	
15	2014	5.32	7.33	
16	2013	¿?	6.36	***Se anula***
17	2015	6.87	5.16	
18	2016	6.32	7.18	
19	2017	1.45	2.67	
20	2018	11.27	12.46	

Eliminando momentáneamente las observaciones que no tienen su par ordenado correspondiente, o presentan datos faltantes, son las siguientes: observación **3** año 2001, observación **6** año 2004 y la observación **16** año 2013. La Tabla 1, se reduce a 17 como se presenta en la siguiente tabla. Pero se mantiene los datos atípicos que después serán corregidos.

Tabla 2

nº	*Obs*	*Año*	*X*	*Y*	*XY*	*X2*	*Y2*
1	1	1999	5.3	4.15	21.995	28.09	17.2225
2	2	2000	3.8	3.25	12.35	14.44	10.5625
3	**4**	**2002**	**12.95**	**2.17**	28.1015	167.7025	4.7089
4	5	2003	8.45	6.32	53.404	71.4025	39.9424
5	7	2005	8.82	9.12	80.4384	77.7924	83.1744
6	8	2006	4.65	5.38	25.017	21.6225	28.9444
7	**9**	**2007**	**3.05**	**14.17**	43.2185	9.3025	200.7889
8	10	2008	2.49	3.87	9.6363	6.2001	14.9769
9	11	2009	7.3	5.35	39.055	53.29	28.6225
10	12	2010	12.28	10.64	130.6592	150.7984	113.2096
11	13	2011	10.63	10.87	115.5481	112.9969	118.1569
12	14	2012	9.7	10.98	106.506	94.09	120.5604
13	15	2014	5.32	7.33	38.9956	28.3024	53.7289
14	17	2015	6.87	5.16	35.4492	47.1969	26.6256
15	18	2016	6.32	7.18	45.3776	39.9424	51.5524
16	19	2017	1.45	2.67	3.8715	2.1025	7.1289
17	20	2018	11.27	12.46	140.4242	127.0129	155.2516
			120.65	**121.07**	**930.0471**	**1052.2849**	**1075.1577**

Con los 17 datos se procede a calcular el coeficiente de correlación entre las dos Estaciones

Coordenada del centro de gravedad

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{120.65}{17} = 7.097058824 \qquad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{121.07}{17} = 7.12176471$$

Covarianza $\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y} \Rightarrow \sigma_{XY} = \frac{1}{17}(930.0471) - \left(\frac{120.65}{17}\right)\left(\frac{121.07}{17}\right) = 4.1650699$

$$\sigma_{XY}^2 = (4.1650699)^2 = 17.3478072 \approx 17.348$$

Varianza

$$\sigma_X^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2 = \frac{1052.2849}{17} - \left(\frac{120.65}{17}\right)^2 = 11.5308678$$

$$\sigma_Y^2 = \frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2 = \frac{1075.1577}{17} - \left(\frac{121.07}{17}\right)^2 = 12.5250381$$

Desviación estándar o típica

$$\sigma_X = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} = \sqrt{11.5308678} = 3.39571315$$

$$\sigma_Y = \sqrt{\frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2} = \sqrt{12.5250381} = 3.53907305$$

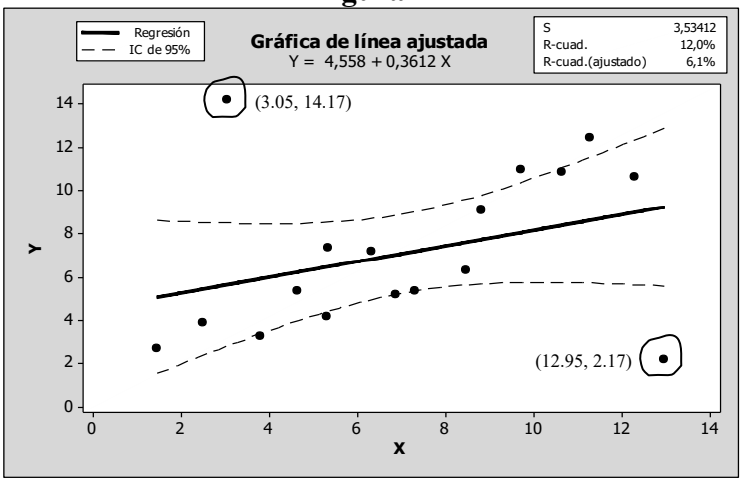
Coeficiente de correlación

$$R = r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{4.1650699}{(3.39571315)(3.53907305)} = 0.346578622 \approx 0.35 \ \blacksquare$$

El coeficiente de correlación entre la estación X y la estación Y, como se puede observar es muy baja positiva $R = 0.35$, por consiguiente ***no se puede estimar o completar*** los datos faltantes serian sesgados.

En torno a la *recta de regresión*, se puede establecer un intervalo de confianza, de tal forma que el 95% de los datos se encuentren dentro de este intervalo, existen dos puntos alejados del resto que son datos atípicos, los mismos afectan considerablemente en el cálculo del coeficiente de correlación y estas no se ajustan al modelo. (Ver figura 1).

Figura 1



Los ***datos anómalos o atípicos*** ocasionan problemas serios en el análisis estadístico, reportando correlaciones muy bajas, seguidamente se dejan de lado estos dos datos detectados en el anterior gráfico, los mismos corresponden a las observaciones **4** año **2002** y **9** año **2007**, (Ver Tabla 1) luego el número de observaciones de 17 se reduce a 15.

Tabla 3 Sin datos atípicos

nº	*Obs.*	*Año*	*X*	*Y*	*XY*	*X2*	*Y2*
1	1	1999	5.3	4.15	21.995	28.09	17.2225
2	2	2000	3.8	3.25	12.35	14.44	10.5625
3	5	2003	8.45	6.32	53.404	71.4025	39.9424
4	7	2005	8.82	9.12	80.4384	77.7924	83.1744
5	8	2006	4.65	5.38	25.017	21.6225	28.9444
6	10	2008	2.49	3.87	9.6363	6.2001	14.9769
7	11	2009	7.3	5.35	39.055	53.29	28.6225
8	12	2010	12.28	10.64	130.6592	150.7984	113.2096
9	13	2011	10.63	10.87	115.5481	112.9969	118.1569
10	14	2012	9.7	10.98	106.506	94.09	120.5604
11	15	2014	5.32	7.33	38.9956	28.3024	53.7289
12	17	2015	6.87	5.16	35.4492	47.1969	26.6256
13	18	2016	6.32	7.18	45.3776	39.9424	51.5524
14	19	2017	1.45	2.67	3.8715	2.1025	7.1289
15	20	2018	11.27	12.46	140.4242	127.0129	155.2516
			104.65	**104.73**	**858.7271**	**875.2799**	**869.6599**

Seguidamente se calcula el coeficiente de correlación

Coordenada del centro de gravedad

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{104.65}{15} = 6.97666667 \quad ; \quad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{104.73}{15} = 6.982$$

Covarianza $\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y} \Rightarrow \sigma_{XY} = \frac{1}{15}(858.7271) - \left(\frac{104.65}{15}\right)\left(\frac{104.73}{15}\right) = 8.537386667$

$$\sigma_{XY}^2 = (8.537386667)^2 = 72.8869711$$

$$\sigma_X^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2 = \frac{875.2799}{15} - \left(\frac{104.65}{15}\right)^2 = 9.678115557$$

Varianza

$$\sigma_Y^2 = \frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2 = \frac{869.6599}{15} - \left(\frac{104.73}{15}\right)^2 = 9.229002666$$

Coeficiente de correlación

$$R = r = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{8.537386667}{\sqrt{(9.678115557)(9.229002666)}} = 0.903341919 \approx 0.90 \quad \blacksquare$$

Una vez eliminado los datos atípicos el coeficiente de correlación mejoró considerablemente $R = 0.90$ con este resultados, estamos en condiciones de generar el Modelo de Regresión Ortogonal, misma nos permitirá estimar los datos faltantes para ambos estaciones así como corregir los datos atípicos.

1. Modelo de Regresión Ortogonal Y vs X

Utilizando la ecuación de segundo grado tenemos:

$$\underbrace{\lambda^2}_{a} - \underbrace{(\sigma_X^2 + \sigma_Y^2)}_{b}\lambda + \underbrace{(\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2)}_{c} = 0$$

$a = 1$

$b = -(\sigma_X^2 + \sigma_Y^2) = -(9.678115557 + 9.229002666) = -18.90711822$

$c = \sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2 = (9.678115557)(9.229002666) - 72.8869711 = 16.43238318$

Reemplazando los valores anteriores en la siguiente formula tenemos:

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$\lambda = \frac{-(-18.90711822) \pm \sqrt{(-18.90711822)^2 - 4\times 1\times 16.43238318}}{2\times 1} =$$

$$\lambda = \frac{-(-18.90711822) \pm 17.08067875}{2}$$

$$\lambda_1 = 0.913219734 \quad , \quad \lambda_2 = 17.99389849$$

Pendiente $\boxed{b_1 = m = \frac{\sigma_{XY}}{\lambda_2 - \sigma_Y^2}}$

$$b_1 = m = \frac{\sigma_{XY}}{\lambda_2 - \sigma_Y^2} = \frac{8.537386667}{17.99389849 - 9.229002666} = 0.974043141$$

$$b_0 = \bar{Y} - m\bar{X} \Rightarrow b_0 = \frac{104.73}{15} - 0.974043141\left(\frac{104.65}{15}\right) = 0.186425686$$

$$Y = b_0 + b_1 X \Rightarrow Y = 0.186425686 + 0.974043141X$$

Modelo Y vs X $Y = 0.1864 + 0.9740X$ ■

2. Modelo de Regresión Ortogonal *X* vs *Y*

En este caso utilizaremos las siguientes fórmulas:

Pendiente $\boxed{b_1' = m = \frac{\sigma_{XY}}{\lambda_2 - \sigma_X^2}}$

$$b_1' = m = \frac{\sigma_{XY}}{\lambda_2 - \sigma_X^2} = \frac{8.537386667}{17.99389849 - 9.678115557} = 1.026648571$$

$$\boxed{b_0' = \bar{X} - m\bar{Y}}$$

$$b_0' = \bar{X} - m\bar{Y} = 6.97666667 - 1.026648571(6.982) = -0.19139365$$

$$X = b_0' + b_1' Y \Rightarrow X = -0.19139365 + 1.026648571Y$$

Modelo X vs Y $X = -0.191 + 1.027Y$ ■

OTRA FORMA

a) Modelo de Regresión Ortogonal Y vs X

Utilizando la *ecuación de la línea* recta tenemos:

Ecuación de la línea recta $\boxed{Y-\bar{Y}=m(X-\bar{X})}$

$$\bar{X}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{104.65}{15}=6.97666667 \;\; ; \;\; \bar{Y}=\frac{\sum_{i=1}^{n}y_i}{n}=\frac{104.73}{15}=6.982$$

$$\sigma_{XY}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{X}\bar{Y}\Rightarrow\sigma_{XY}=\frac{1}{15}(858.7271)-\left(\frac{104.65}{15}\right)\left(\frac{104.73}{15}\right)=8.537386667$$

$$\lambda_2=17.99389849$$

$$m=\frac{\sigma_{XY}}{\lambda_2-\sigma_Y^2}=\frac{8.537386667}{17.99389849-9.229002666}=0.974043141$$

Reemplazando en la ecuación de la línea recta tenemos:

$$Y-\bar{Y}=m(X-\bar{X})\Rightarrow Y-6.982=0.974043141(X-6.97666667)$$
$$Y=0.974043141X-6.795574317+6.982\;\Rightarrow\; Y=0.186425683+0.974043141X$$

Modelo Y vs X $\boxed{Y=0.1864+0.9740X}$ ■

b) Modelo de Regresión Ortogonal *X* vs *Y*

Utilizando la *ecuación de la línea* recta tenemos:

Ecuación de la línea recta $\boxed{X-\bar{X}=m(Y-\bar{Y})}$

$$\bar{X}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{104.65}{15}=6.97666667 \;\; ; \;\; \bar{Y}=\frac{\sum_{i=1}^{n}y_i}{n}=\frac{104.73}{15}=6.982$$

$$\sigma_{XY}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{X}\bar{Y}\Rightarrow\sigma_{XY}=\frac{1}{15}(858.7271)-\left(\frac{104.65}{15}\right)\left(\frac{104.73}{15}\right)=8.537386667$$

$$\lambda_2=17.99389849$$

$$m=\frac{\sigma_{XY}}{\lambda_2-\sigma_X^2}=\frac{8.537386667}{17.99389849-9.678115557}=1.026648571$$

Reemplazando en la ecuación de la línea recta tenemos:

$$X-\bar{X}=m(Y-\bar{Y})\Rightarrow X-6.97666667=1.026648571(Y-6.982)$$
$$X=1.026648571Y-7.168060323+6.97666667\ \Rightarrow\ X=-0.19139365+1.026648571$$

Modelo X vs Y $X=-0.191+1.027Y$ ■

Con *MINITAB* se obtienen los mismos resultados:

Análisis de regresión ortogonal: Y versus X

Relación Error - Varianza (Y/X): 1

Ecuación de regresión

Y = 0,186 + 0,974 X

Coeficientes

Predictor	Coef	SE Coef	Z	P	IC de 95% aprox.
Constante	**0,18643**	0,969802	0,1922	0,848	(-1,71435; 2,08720)
X	**0,97404**	0,128492	7,5806	0,000	(0,72220; 1,22588)

Análisis de regresión ortogonal: X versus Y

Relación Error - Varianza (X/Y): 1

Ecuación de regresión

X = - 0,191 + 1,027 Y

Coeficientes

Predictor	Coef	SE Coef	Z	P	IC de 95% aprox.
Constante	**-0,19139**	1,01920	-0,1878	0,851	(-2,18898; 1,80619)
Y	**1,02665**	0,13546	7,5791	0,000	(0,76116; 1,29214)

Con los modelos obtenidos de manera analítica y también por medio de *MINITAB* se procede a pronosticar o completar los datos faltantes y corregir los datos anómalos o atípicos de la siguiente forma:

3.1 Completado de datos. Una vez determinado ambos Y vs X y X vs Y, se completó los datos faltantes o pronósticos de la siguiente forma:

Para el año 2001, $Y = 4.6$ entonces $X_{2001} = -0.191 + 1.027(4.6) = 4.5332 \approx 4.53$

Para el año 2004, $X = 4.95$ entonces $Y_{2004} = 0.186 + 0.974(4.95) = 5.0077 \approx 5.01$

Para el año 2013, $Y = 6.36$ entonces $X_{2013} = -0.191 + 1.027(6.36) = 6.34072 \approx 6.34$

Tabla 4 Datos pronosticados

Obs.	Años	Estación *X*	Estación *Y*
3	2001	**4.43**	4.6
6	2004	4.95	**5.0**
16	2013	**6.34**	6.36

3.2 Corrección de datos. Los datos anómalos o atípicos se corrigen, tomando como referencia la Estación ***X,*** para corregir la estación ***Y***, utilizando el siguiente Modelo de Regresión Ortogonal que fue determinado anteriormente:

$$Y = 0.1864 + 0.9740X$$

Para el año 2002, $X = 12.95 \Rightarrow Y_{2002} = 0.186 + 0.974(12.95) = 12.7997 \approx 12.80$

Para el año 2007, $X = 3.05 \Rightarrow Y_{2007} = 0.1864 + 0.974(3.05) = 3.1571 \approx 3.16$

Tabla 5 Datos corregidos

Obs.	Años	Estación *X*	Estación *Y*
4	2002	12.59	**12.80**
9	2007	3.05	**3.16**

Finalmente se procedió a calcular el coeficiente de correlación entre las estaciones X y Y, con todos los datos, es decir con los datos estimados y datos corregidos.

Tabla 6 Con pronósticos y corregidos

Obs.	*Año*	*X*	*Y*	*XY*	*X2*	*Y2*
1	1999	5.3	4.15	21.995	28.09	17.2225
2	2000	3.8	3.25	12.35	14.44	10.5625
3	2001	**4.43**	4.6	20.378	19.6249	21.16
4	2002	12.95	**12.8**	165.76	167.7025	163.84
5	2003	8.45	6.32	53.404	71.4025	39.9424
6	2004	4.95	**5**	24.75	24.5025	25
7	2005	8.82	9.12	80.4384	77.7924	83.1744
8	2006	4.65	5.38	25.017	21.6225	28.9444
9	2007	3.05	**3.16**	9.638	9.3025	9.9856
10	2008	2.49	3.87	9.6363	6.2001	14.9769
11	2009	7.3	5.35	39.055	53.29	28.6225
12	2010	12.28	10.64	130.6592	150.7984	113.2096
13	2011	10.63	10.87	115.5481	112.9969	118.1569
14	2012	9.7	10.98	106.506	94.09	120.5604
15	2014	5.32	7.33	38.9956	28.3024	53.7289
16	2013	**6.34**	6.36	40.3224	40.1956	40.4496
17	2015	6.87	5.16	35.4492	47.1969	26.6256
18	2016	6.32	7.18	45.3776	39.9424	51.5524
19	2017	1.45	2.67	3.8715	2.1025	7.1289
20	2018	11.27	12.46	140.4242	127.0129	155.2516
		136.37	**136.65**	**1119.5755**	**1136.6079**	**1130.0951**

Coordenada del centro de gravedad

$$\bar{X}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{136.37}{20}=6.8185 \qquad \bar{Y}=\frac{\sum_{i=1}^{n}y_i}{n}=\frac{136.65}{20}=6.8325$$

Covarianza $\sigma_{XY}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{X}\bar{Y}\Rightarrow\sigma_{XY}=\frac{1}{20}(1119.5755)-(6.8185)(6.8325)=9.39137375$

Varianza

$$\sigma_X^2=\frac{\sum x_i^2}{n}-\left(\frac{\sum x_i}{n}\right)^2=\frac{1136.6079}{20}-(6.8185)^2=10.33845275$$

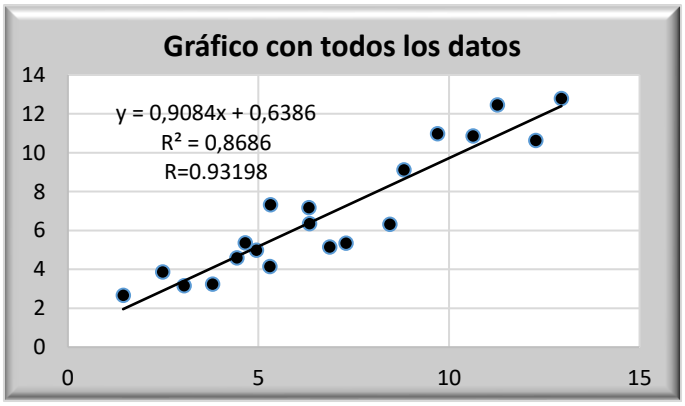
$$\sigma_Y^2=\frac{\sum y_i^2}{n}-\left(\frac{\sum y_i}{n}\right)^2=\frac{1130.0951}{20}-(6.8325)^2=9.82169875$$

Coeficiente de correlación

$$R=r=\frac{\sigma_{XY}}{\sqrt{\sigma_X^2\sigma_Y^2}}=\frac{9.39137375}{\sqrt{(10.33845275)(9.82169875)}}=0.93198312\approx 0.93\ \blacksquare$$

Una vez completado los datos faltantes y corregidos los datos anómalos o atípicos (Ver Tabla 6), se vuelve a calcular el coeficiente de correlación, como resultado se obtiene un coeficiente de correlación positiva muy alto $R = 0.93$, con un coeficiente de determinación de $R^2 = 0.8686$, que indica que los datos se ajustaron de manera apropiada al diagrama de dispersión.

Figura 2 Con todos los dato, pronosticados y corregidos



El diagrama de dispersión nos muestra que los datos están en torno a la línea estimada, por consiguiente el coeficiente de correlación es muy alta y positiva $R = 0.93$, mejoró sustancialmente con respecto al primer coeficiente de correlación que era $R = 0.35$. Consiguientemente el coeficiente de determinación llamado también medida de bondad de ajuste, que es el cuadrado del coeficiente de correlación $R^2 = 0.8686$ o $R^2 = 87\%$ (Ver tabla 7) nos indica que la línea estimada se ajusta aceptablemente al diagrama de dispersión.

Tabla 7 Resumen

Casos	Coeficiente de Correlación R	Coeficiente de Determinación R^2
Con datos atípicos $n = 17$	$R = 0.35$	$R^2 = 0.1225$
Sin datos atípicos $n = 15$	$R = 0.90$	$R^2 = 0.81$
Con pronósticos y datos corregidos $n = 20$	$R = 0.93$	$R^2 = 0.87$

IV. DISCUSIÓN

El modelo de regresión ortogonal que generalmente se aplica en hidrología, perfectamente se puedes aplicar en el área social, donde las encuestas se aplican a un determinado grupo de personas, cuya muestra se determina con un margen de error, en varias investigaciones las personas se resisten a responder o no responden definitivamente, también está la posibilidad de que mientan, se puede afirmar que los resultados del modelo de regresión ortogonal en la corrección de datos atípicos o anómalos así como en pronósticos es aceptable, con relación a otros modelos como ser el modelo de regresión lineal.

Por lo mencionado anteriormente se recomienda utilizar el modelo de regresión ortogonal en diferentes aplicaciones donde se presenten datos atípicos o anómalos principalmente.

V. CONCLUSIONES

El estudio del modelo de regresión ortogonal, fue muy útil y eficiente, principalmente para corregir los datos atípicos o anómalos así como para estimar los datos faltantes, donde el coeficiente de correlación que era $R = 0.35$ con datos atípicos, descartando los mismos el coeficiente de correlación mejora bastante es de $R = 0.\ 90$. Una vez corregido los datos atípicos tomando en cuenta el modelo de regresión ortogonal y estimado los datos faltantes, el coeficiente de correlación mejora ostensiblemente $R = 0.93$ muy alta positiva. . Por lo tanto el modelo de regresión ortogonal es muy útil y eficiente principalmente para corregir datos atípicos, con relación a otros modelos.

VI. AGRADECIMIENTOS

El autor desea agradecer a Huáscar Fedor Gonzales Guzmán, por el impulso y motivación transmitida para la finalización de este trabajo.

VII. REFERENCIAS BIBLIOGRÁFICAS

Referencias

Dagnino, J. (2014). Datos Faltantes (Missing Values). *Bioestadística y Epidemiología* , 33-334.

García, C. (2017). Regresión con Variables Ortogonales y Regresión Alzada con el Método STIRPAT. *Estudio de Econonomía Aplicada*, 717-734.

Gutierrez, C. (2014). *Hidrología Básica y Aplicada.* Quito: Abya-Yala.

Perez, L. (2019). Valores Atipicos en los datos ¿Como identificarlos y manejarlos? *Jardin Botanico Nacional*, 99-107.

Recio, J. (2021). Regresión Ortogonal. *Pensamiento Matematico*, 5-15.

Soto, M. (2022). *Estadística Aplica I, Descriptiva con Pronosticos* . Potosí: Navarro.

Velez, M. (2000). *Hidrología Para Ingenieros.* Medellin .