PROPUESTA DE PROYECTO DE ESTADÍSTICA: UN MODELO DE REGRESIÓN LINEAL SIMPLE PARA PRONOSTICAR LA CONCENTRACIÓN DE CO_2 DEL VOLCÁN MAUNA LOA

* CLAUDIO ALFREDO LÓPEZ MIRANDA, CÉSAR AUGUSTO ROMERO RAMOS

RESUMEN

Este trabajo aplica un modelo predictivo de regresión lineal para analizar la contaminación atmosférica de dióxido de carbono (CO_2) producida por el volcán Mauna Loa de Hawái. Los datos fueron extraídos de un repositorio de internet que contiene múltiples casos de geología, climatología, física, etcétera. El modelo se utilizó para predecir la tendencia de emisiones de CO_2 con respecto al tiempo; se estimó la contaminación promedio de dicha tendencia, la cual descubrimos ha crecido aproximadamente 0.1 partes por millón por mes; así comotambién se obtuvieron los intervalos de predicción para una emisión puntual que existió en un momento determinado. Se recomienda el trabajo para estudiantes de ciencias exactas y naturales, como prototipo de artículo de investigación donde se aplique específicamente el modelo de regresión lineal simple; aunque la estructura también puede servir en otras áreas donde se enseñen los modelos de regresión.

Palabras clave: *Regresión lineal simple, estadística aplicada.*

DR. CLAUDIO ALFREDO LÓPEZ MIRANDA
Departamento de Matemáticas, Universidad de Sonora
Correo: claudio@mat.uson.mx
EST. CÉSAR AUGUSTO ROMERO RAMOS
Departamento de Física, Universidad de Sonora
Correo: romero.rca_81@hotmail.com

*Autor para correspondencia: Claudio Alfredo López Miranda
Correo electrónico: claudio@mat.uson.mx
Recibido: 04 de septiembre del 2014
Aceptado: 24 de noviembre del 2014
ISSN: 2007-4530

MOTIVACIÓN Y JUSTIFICACIÓN

Las carreras de Ciencias Exactas y Naturales, como las licenciaturas en Matemáticas, Física, Geología y Ciencias Computacionales, tienen una orientación fuerte hacia la investigación teórica y aplicada, por lo que es deseable que sus estudiantes comiencen a desarrollar su capacidad investigadora desde el inicio de la carrera, por ejemplo, elaborando un proyecto de estadística al final de un curso, en la forma de un artículo de revista. Se sugiere usar datos reales de algún experimento observacional, pruebas de laboratorio, o de algún repositorio en internet. Se recomienda que tanto el tema como la "recolección" de datos se deje al criterio del estudiante, así habrá mayor entusiasmo y profundidad en la investigación.

Generalmente el modelo de Regresión Lineal (RL) se estudia en los cursos hasta el final del semestre, por lo que un proyecto de RL serviría como estudio integral de la mayoría de los temas ya que incluiría gran parte de las técnicas de estadística, tanto descriptiva como de estadística inferencial.

La estructura de este trabajo está planeada como guía para el estudiante en la elaboración de un artículo, o como ejemplo para exponerse en clase. El trabajo va dirigido a estudiantes de Ciencias Exactas y Naturales, aunque la estructura general puede utilizarse en otras áreas como Ingeniería Química, Biología, Ciencias Sociales, Nutrición, o prácticamente cualquier área donde se curse la materia de estadística y se estudie la parte descriptiva y de inferencia de los modelos de regresión lineal simple.

INTRODUCCIÓN

El Mauna Loa ("Montaña Grande") es uno de los cinco volcanes que forman la Isla de Hawái en el Océano Pacífico (Figura 1). Históricamente es considerado el volcán más grande sobre la tierra, tanto en masa como en volumen; es un volcán activo de 75,000 km^3 de volumen. Tiene 700,000 años haciendo erupciones, y debió salir sobre el nivel del mar hace 400,000 años. En este trabajo analizaremos las mediciones de gas invernadero monitoreadas a lo largo del tiempo, desde 1965 hasta 1980, con el fin de estimar la tendencia de emisiones de CO_2. Cabe mencionar que nos limitamos a investigar sólo el crecimiento de emisiones (la tendencia) con respecto al tiempo, y no en el contexto de pronósticos para *series de tiempo*.

Figura 1. Ríos de lava sobre el Mauna Loa.

El objetivo de este trabajo es presentar un artículo que contemple algunos de los pasos que deben considerarse a la hora de aplicar un modelo de regresión lineal simple, lo cual se ejemplifica mediante la deducción de un modelo lineal para pronosticar la concentración de CO_2 del volcán Mauna Loa. Los métodos estadísticos aquí presentados tienen un nivel de dificultad de acuerdo al programa de un primer curso de estadística, por lo que algunos aspectos de la teoría de regresión lineal no son contemplados de manera exhaustiva, como por ejemplo el análisis residual o el de varianza. Es común que a la hora de intentar aplicar un modelo de regresión lineal simple, el estudiante o investigador novato se limite a usar un paquete de cómputo para obtener una estimación de los parámetros del modelo, a interpretarlos, a graficar los puntos junto con la recta de regresión, y a usar el modelo resultante para pronosticar valores de la variable de respuesta, todo ello sin validar en momento alguno si el modelo es correcto o no. Por lo tanto, para que la aplicación resulte lo más realista y completa posible, se debe por un lado, de verificar primeramente si los supuestos del modelo son válidos antes de proceder a obtener el modelo mismo; y por otro lado, una vez verificados los supuestos, se debe de explotar la gama de herramientas estadísticas para estudiar la precisión de las estimaciones, incluir por lo menos un análisis parcial de varianza y de residuos, estudiar los intervalos de confianza y de predicción, así como distintas pruebas de hipótesis.

En consecuencia, el trabajo está presentado en el siguiente orden. Primero explicamos el concepto de regresión lineal; luego realizamos la exploración gráfica de los datos para detectar que el patrón lineal es factible. Después verificamos mediante gráficas y pruebas formales los supuestos del modelo, a saber, la hipótesis de homocedasticidad de la varianza y el supuesto de normalidad de los residuos. Al concluir que los supuestos se cumplen, se procede con la estimación de parámetros del modelo β_0 y β_1. Una vez que se tienen las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$, es posible construir dicho modelo; sin embargo, antes de usarlo para hacer pronósticos, se debe estudiar la precisión en la estimación de los parámetros de la recta (en particular la pendiente) así como validar mediante una prueba si dicha pendiente es significativa. Si el modelo que relaciona Y y X, resulta significativo, esto es, si estadísticos como el coeficiente de determinación R^2 explica en gran medida la variabilidad en la respuesta (CO_2), y el coeficiente de correlación R arroja una fuerte dependencia lineal entre la variable regresora (tiempo) y la variable de respuesta (CO_2), entonces tendrá sentido el análisis inferencial que se realice, siempre y cuando la relación entre X e Y, no sea solamente una correlación espuria o de falta de causalidad. Posterior a verificar lo anteriormente expuesto, se computan los intervalos de confianza para la pendiente de la recta β_1, se calcula una estimación puntual y los intervalos de confianza para un valor promedio $E(Y/x)$, de la variable de respuesta, así como una predicción puntual y un intervalo de predicción de Y. Finalmente se utiliza el modelo como predictor-pronosticador.

EL MODELO DE REGRESIÓN LINEAL

Los modelos de RL simple o bivariada, se utilizan como modelos de predicción o pronóstico. El caso más típico es cuando la variable predictora, regresora o independiente X es una variable controlada (no aleatoria), mientras que la variable de respuesta o dependiente Y resulta una variable aleatoria que tiene una distribución aproximadamente normal para cada valor x de X, pero con varianza constante σ^2. Dicha varianza se debe al error aleatorio en cada medición. Los modelos de RL surgieron desde 1889 cuando Francis Galton [1] los utilizó para pronosticar la estatura de los hijos a través de la estatura de los padres. El término "regresión" se usó en principio para indicar que ciertos fenómenos presentan continuamente mediciones altas y bajas, pero que dichas mediciones eventualmente "regresan" a un promedio desconocido pero esperado, el cual depende del momento en que se mide x. Cuando el promedio indica un desempeño pobre, entonces se dice que hay una regresión a la medianía, tal como lo usó Francis Galton [1] en su artículo pionero "*Regression towards mediocrity in hereditary stature*", donde establece que hijos de padres altos no son tan altos como sus padres, e hijos de padres bajos no son tan bajos como sus padres.

Por cuestiones didácticas en los cursos de estadística comúnmente nos enseñan primero a estimar los parámetros β_0 y β_1 de la recta de regresión (1) sin necesidad de verificar los supuestos del modelo (aunque en una aplicación real, el primer paso debe ser verificar la homocedasticidad de la varianza y la normalidad de los residuos). Dichos parámetros se estiman mediante el análisis de una muestra apareada de valores $(x,\ y)$ y aplicando el criterio de mínimos cuadrados:

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (1)$$

Donde β_0 es la ordenada en el origen y β_1 la pendiente de la recta. El valor ϵ representa un error de medición o ruido aleatorio, que de no existir, los valores $y's$ quedarían perfectamente sobre la recta. Es común suponer que el error ϵ es una variable aleatoria con distribución normal de media cero y varianza constante σ^2 e independiente del tiempo; asimismo, se supone que la variable de respuesta Y está distribuida normalmente también con varianza σ^2 para cada valor x. El supuesto de varianza constante constituye la hipótesis de homocedasticidad y se analizará más adelante.

Con la estimación de los parámetros se obtiene la ecuación de la recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$; donde $\hat{y}$ representa en nuestro estudio la cantidad estimada de la concentración de CO_2 para un tiempo particular x. Al valor $\hat{y}$ se le conoce en la literatura como el valor ajustado o simplemente ajuste; mientras que, a la recta se le conoce como la recta de ajuste o recta de regresión muestral. Así, y representará la concentración observada real mientras $\hat{y}$ será la concentración ajustada o pronosticada; a la diferencia entre ellas se le conoce como el residuo o error de estimación, denotado por $(y - \hat{y})$. Los residuos se representan gráficamente como el segmento vertical entre el punto correspondiente sobre la recta $(x, \hat{y})$ y el punto observado (x, y).

ANÁLISIS EXPLORATORIO DE DATOS

Primeramente se debe de explorar si la distribución de los datos se aproxima a un patrón lineal. Si el comportamiento se aleja de una línea recta el modelo lineal se descarta. En la figura 2 se ilustra el comportamiento de $n = 192$ mediciones de densidad de CO_2 en ppm en la atmósfera sobre Mauna Loa, desde enero de 1965 hasta diciembre de 1980. Observe la tendencia de crecimiento promedio lineal a grandes intervalos. El tiempo representa la variable explicativa X en meses, mientras que la concentración de CO_2 representa la variable de respuesta Y en ppm. Los datos se extrajeron del portal *DataMarket.com* usando la siguiente liga http://datamarket.com/data/set/22v1/co2-ppm-mauna-loa-1965-1980#!ds=22v1&display=line; una vez en el portal hacer clic sobre la pestaña *exportar* para extraer la serie de tiempo en el formato deseado (*Excel* por ejemplo).

Aunque no se muestran todos los detalles, se puede observar que la concentración de CO_2 se mantiene a la alza por un período de 5 a 6 meses, para después tener un período de tiempo similar a la baja. Esta forma pseudosinusoidal que adquiere la concentración de CO_2 es muy similar al comportamiento oscilatorio de muchos de los fenómenos de naturaleza geológica.

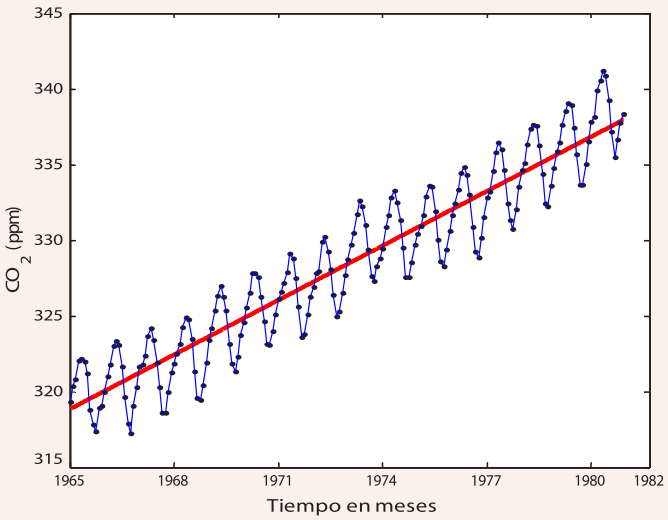

Figura 2. Concentración de CO_2 sobre el Mauna Loa respecto al tiempo.

ANÁLISIS DE HOMOCEDASTICIDAD DE LA VARIANZA

La homocedasticidad de la varianza se verifica primero de manera visual mediante una gráfica de puntos entre los valores predichos y los residuos, ambos estandarizados o

tipificados (Figura 3). Si la varianza es constante la gráfica no debe mostrar ningún patrón entre los residuos, como argumenta Lattin [2, p. 59]; por el contrario, si existe heterogeneidad en la varianza (i.e. la varianza depende del valor observado), la gráfica puede mostrar anchos distintos en la variabilidad, típicamente una gráfica en forma de embudo, sea hacia la izquierda, a la derecha o al centro. Observe en la figura 3 que no hay un patrón bien marcado de cambios, por lo que en apariencia el ancho o variación se aprecia muy similar [3, p.65].

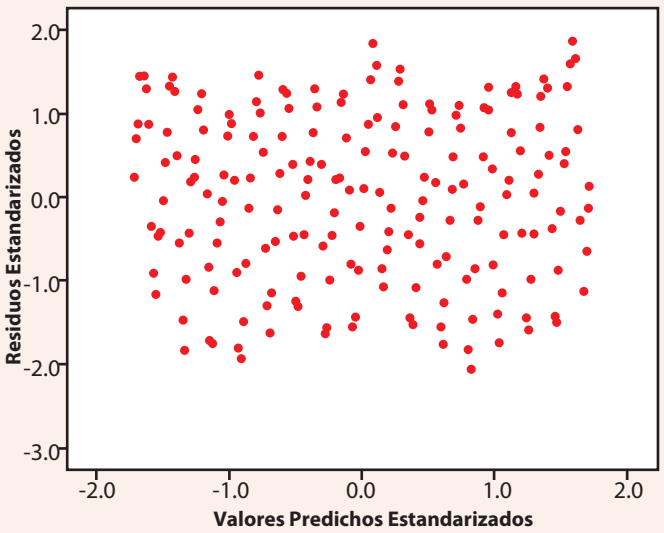

Figura 3. Distribución residual en función de los valores predichos, ambos estandarizados.

Sin embargo, para una prueba formal de homocedasticidad se debe aplicar un método estadístico formal, el cual consiste en probar si existe una correlación entre los residuos (en valor absoluto ya que de otra forma la correlación es cero) y el valor predicho no estandarizado. En este caso se obtuvo el coeficiente de correlación de Pearson igual a 0.025 con un p-valor de 0.731 para una prueba bilateral, con lo que se confirma que no hay ningún tipo de relación entre los residuos y los valores predichos, lo que da pie para asumir que σ^2 no debe variar entre un punto y otro. Esta prueba se realizó con SPSS en el menú *analizar/correlaciones/bivariadas* y seleccionado las dos variables involucradas, para un manual de estas pruebas de SPSS [4] ver el de C. Pérez.

ANÁLISIS RESIDUAL

Para mostrar que los residuos tienen una distribución aproximadamente normal con media cero y varianza σ^2, es común emplear un histograma de la distribución o bien una gráfica de probabilidad P-P (Figura 4). Observamos que los datos se dispersan con pequeñas desviaciones alrededor del patrón lineal esperado para una distribución normal. Pensamos que las desviaciones fluctuantes se deben a la oscilación de la serie de tiempo y algunos picos en la concentración de CO_2.

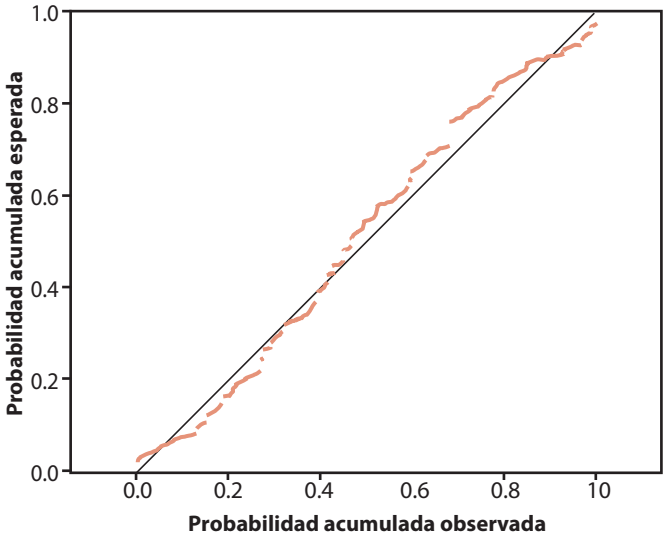

Figura 4. Gráfica de probabilidad normal P-P de residuos estandarizados.

No obstante, si deseamos ser más rigurosos podemos recurrir a procedimientos analíticos, como la prueba *Kolmogorov-Smirnov* para la normalidad. La tabla 1 muestra los resultados de dicha prueba obtenidos mediante SPSS (en el menú *analizar/pruebas no paramétricas/una muestra*), donde observamos un p-valor de 0.182, por lo que a un nivel de significancia bilateral menor al 18.2% no se rechaza la hipótesis de normalidad.

Tabla 1. Prueba de *Kolmogorov-Smirnov* para una muestra de residuos no estandarizados.

		Residuos no estandarizados
N		192
Parámetros normales	Media	.0000000
	Desviación típica	2.02109430
Diferencias más extremas	Absoluta	.079
	Positiva	.056
	Negativa	-.079
Sig. asintótica (bilateral)		.182

ESTIMACIÓN DE PARÁMETROS

De las secciones previas deducimos que el modelo lineal es factible, por lo que se procede a estimar los parámetros. Para una cantidad grande de datos la

estimación de los parámetros se realiza mediante un paquete de cómputo estadístico, por ejemplo, *Excel, SPSS, R, Matlab,* o *Calc* de *Open Office*, los cuales realizan el proceso automáticamente. La ecuación estimada de la recta resultó:

$$\hat{y}=0.10095x + 318.82 \qquad (2)$$

Lo anterior nos dice que al inicio de las mediciones (tiempo $x = 0$) la densidad de CO_2 se estima alrededor de $\hat{\beta}_0 = 318.82$ ppm; mientras que la pendiente positiva de la recta nos indica que la concentración estuvo creciendo aproximadamente (estimación) a $\hat{\beta}_1 = 0.1$ ppm por mes. No esperamos que la concentración siempre suba 0.1 ppm cada mes, en ocasiones estará por debajo de su valor esperado debido a la oscilación. Lo que estamos estimando con esta RL es la tendencia, la cual tiene sentido a grandes intervalos de tiempo. Dicho de otra forma, vemos que la serie de datos oscila alrededor de una media; y lo que la regresión lineal nos dice es cómo crece esta media.

La estimación de los parámetros del modelo de regresión y de otros valores de interés, se realiza a través de un conjunto típico de estadísticas, las cuales se resumen a continuación para referencias posteriores, entre ellas están $\bar{x} = 95.5$; $\bar{y} = 328.464$, la suma de los cuadrados (S_{xx} y S_{xy}) y la suma de productos Sxy:

$$\sum xy = 6{,}082{,}256;\ S_{xx} = 589{,}808;\ S_{yy} = 679.191;\ S_{xy} = 59{,}540.86 \quad (3)$$

Donde:

$$S_{xx} = \sum (x-\bar{x})^2;\ S_{yy} = \sum (y-\bar{y})^2 \text{ y } S_{xy} = \sum (x-\bar{x})(y-\bar{y}) \quad (4)$$

Para propósitos didácticos, a continuación mostraremos las fórmulas para obtener las cantidades de interés. Estas fórmulas se pueden consultar en cualquier libro de estadística como el de Devore [5, p.456] o el de Draper [3, pp. 23-33].

Comenzamos calculando $\hat{\beta}_1$ que representa la pendiente estimada de la recta, es decir, la razón de cambio mensual de CO_2; y $\hat{\beta}_0$ su ordenada en el origen que estima la concentración inicial de CO_2:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.10095;\ \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 318.8233 \qquad (5)$$

Con estos valores se obtiene la ecuación de la recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Así, y representará la concentración observada real mientras $\hat{y}$ será la concentración ajustada o pronosticada; recordando que a la diferencia entre ellas se le conoce como el residuo o error de estimación, denotado por $(y - \hat{y})$, el cual es muy importante ya que se utiliza para estimar el error de estimación y para el análisis residual anterior. Elevando al cuadrado los residuos y sumándolos obtenemos la suma de cuadrados del error (SCE), tal que $SCE = 777.9412$. Con esta SCE se obtiene $\hat{\sigma}^2$ en (6), que representa una estimación de la varianza del "*error de estimación*" y que se denota por s^2:

$$\hat{\sigma}^2 = s^2 = \frac{SCE}{n-2} = 4.106321 \qquad (6)$$

Nota: Cuando se utiliza un paquete de cómputo, debemos especificar si la constante β_0 tiene significado práctico, ya que de ello depende con que fórmula se estima el parámetro β_1. Se recomienda tener cuidado con su elección, nosotros supusimos $\beta_0 \neq 0$.

COEFICIENTE DE DETERMINACIÓN

El coeficiente de determinación R^2 es utilizado para medir que tanta variación de la concentración de CO_2 es explicada por el modelo de regresión, es decir que tanto de la variación se atribuye al crecimiento lineal (y no al error aleatorio en cada medición, ya que dicho error hace que varíe por sí misma la concentración). Para calcular R^2 definimos la suma total de los cuadrados de Y como:

$$STC = S_{yy} = \sum (y-\bar{y})^2 = 6{,}791.191, \qquad (7)$$

y obtenemos (tal como lo utiliza Devore [5, p. 463]),

$$R^2 = \frac{STC - SCE}{STC} = 1 - \frac{SCE}{STC} = 0.885448\,. \qquad (8)$$

Se observa en (8) que el numerador representa la diferencia entre la desviación total y la desviación del error, por lo que en realidad este numerador representa la desviación atribuida a la regresión, lo que nos indica que el 88.54% de la variación encontrada en la concentración de CO_2 es explicada por el modelo de regresión. Esto se corrobora con el coeficiente de correlación $r = \sqrt{R^2} = 0.94098$, que mide el grado de dependencia lineal entre el tiempo X y la concentración de CO_2 Y. Vemos una dependencia lineal positiva muy fuerte al quedar r próximo a uno.

ANÁLISIS DE LA PENDIENTE ESTIMADA

Una vez analizada la variabilidad y dependencia lineal, así como la estimación de la pendiente de la recta, es necesario discutir la precisión de $\hat{\beta}_1$ como estimación puntual de dicha pendiente y dar un intervalo de confianza. Recordemos que $\hat{\beta}_1$ es una variable aleatoria ya que depende de la muestra. Por lo tanto, además de la estimación puntual es necesario estimar su variabilidad esperada $\sigma_{\hat{\beta}_1}$, así como un intervalo de confianza para inferir el rango de valores en el que se espera la tasa mensual de CO_2.

La desviación estándar estimada $s_{\hat{\beta}_1}$ y su coeficiente de variación $CV_{\hat{\beta}_1}$, ver Jay L. Devore [5, p.470] son:

$$s_{\hat{\beta}_1} \equiv \frac{s}{\sqrt{S_{xx}}} = 0.002639, \qquad (9)$$

$$CV_{\hat{\beta}_1} \equiv \frac{s_{\hat{\beta}_1}}{\hat{\beta}_1} = \frac{0.002639}{0.10095} x100 = 2.6\%. \qquad (10)$$

De donde observamos una desviación estándar y un

coeficiente de variación bastante pequeño, indicando que la estimación es de muy buena precisión.

En cuanto a la distribución de $\hat{\beta}_1$ como variable aleatoria, sabemos que $\hat{\beta}_1$ es un estimador insesgado (*i.e.*, $E(\hat{\beta}_1) = \beta_1$) con distribución normal; por lo tanto, al tener una desviación estándar desconocida, usamos el estadístico T a continuación (11), el cual tiene una distribución t con $n-2$ grados de libertad, esto se debe a que $n-2$ es el divisor de s^2 en (6), y representa el número de grados de libertad asociado con la estimación (o la suma de cuadrados del error). Como lo explica Devore [5, p. 461], para obtener s^2 primero se deben estimar los parámetros β_0 y β_1, lo que hace que se pierdan 2 grados de libertad. Para más detalles ver las exposiciones de Devore [5, pp. 468-482], y Daper y Smith [3, pp.35-38].

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \qquad (11)$$

A partir de este estadístico el intervalo de confianza para β_1 es

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}. \qquad (12)$$

Sustituyendo valores obtenemos el intervalo bastante angosto [0.0958, 0.1061], lo cual indica que estimamos a β_1 con precisión y buen nivel de confianza del 95%.

PRUEBA DE HIPÓTESIS DE UTILIDAD DEL MODELO

Después de estudiar la precisión en la estimación, haremos un análisis inferencial respecto a la pendiente β_1 a través de su valor estimado $\hat{\beta}_1$, lo que algunos conocen como prueba de utilidad del modelo. Esta prueba consiste en establecer como hipótesis nula $H_0: \beta_1 = 0$ y como alternativa $H_a: \beta_1 \neq 0$ como lo explica Devore [5, p. 474], en otras palabras, proponer $\beta_1 \neq 0$ demuestra que la pendiente es significativa en el modelo, y por tanto, la variable predictora X debe incluirse. Si H_0 es cierta, el estadístico de prueba de (11) resulta $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$. Con n=192 este estadístico resulta $t = 38.25$ y el valor crítico de la región de rechazo está dado por $\pm t_{\alpha/2,n-2} \approx \pm z_{0.025} = \pm 1.96$. El estadístico está muy alejado del valor crítico, por lo tanto, con 95% de confianza rechazamos contundentemente H_0 y concluimos que nuestro modelo de regresión tiene pendiente significativamente distinta de cero, por lo que se dice que el modelo lineal es útil y adecuado. Aunque en nuestro caso de estudio la relación lineal es evidente (Figura 2), esta prueba se debe realizar en muchas aplicaciones para confirmar o rechazar la utilidad del modelo lineal. Es importante mencionar que para fines prácticos el no rechazar H_0 quiere decir que la relación lineal será significativa, aun cuando pudiera no existir necesariamente una condición de causalidad, como por ejemplo cuando una variable oculta correlaciona a dos variables entre sí.

PRONÓSTICO DE CONCENTRACIÓN PROMEDIO DE CO_2 Y VALORES Y

El análisis estadístico de las secciones anteriores nos permite confirmar en primer lugar que el modelo lineal es adecuado, que el modelo explica gran porcentaje de la variabilidad de CO_2 y que además conocemos el error de estimación, por lo tanto, aplicaremos el modelo como herramienta confiable de pronóstico. Lo que haremos es fijar un tiempo determinado x^* y calcular el ajuste $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, el cual puede ser considerado como una estimación puntual de la concentración promedio esperada en ese momento, es decir $E(Y/x^*)$, o como una predicción individual y de la concentración de CO_2 que resultará de una observación puntual en el tiempo $x = x^*$. Las dos afirmaciones anteriores se justifican mediante el siguiente cálculo:

$$E(Y/x^*) = E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^*, \quad (13)$$

ya que se supone que el error aleatorio ϵ tiene valor esperado igual a cero, además de varianza constante σ^2 . Por lo tanto, una estimación natural de $E(Y/x^*)$ sería $\hat{E}(Y/x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^* = \hat{y}$, que como se ve en el lado derecho es en sí misma es una estimación de y. Entonces, si tratamos a $\hat{y}$ como variable aleatoria (pues depende de $\hat{\beta}_0$ y $\hat{\beta}_1$), sabemos que hereda la distribución normal (al ser $\hat{\beta}_0$ y $\hat{\beta}_1$ vv.aa. normales), de acuerdo a Devore [5, p. 469], cuyo valor esperado es $E(\hat{Y}) = \beta_0 + \beta_1 x^*$; por tanto es un estimador insesgado de $E(Y/x^*)$, tal que su varianza resulta, para los detalles ver el desarrollo presentado por Devore [5, p. 478].

$$V(\hat{Y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]. \qquad (14)$$

La raíz cuadrada de (14) arroja la desviación estándar de $\hat{Y}$, sin embargo, al sustituir σ^2 por s^2 lo que obtendremos es su estimación denotada por $s_{\hat{Y}}$:

$$s_{\hat{Y}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}. \qquad (15)$$

La desviación estándar estimada $s_{\hat{Y}}$ se utiliza para construir los intervalos de confianza y de predicción. Por

ejemplo, el *intervalo de confianza* para el valor esperado $E(Y/x^*)$ de la concentración de CO_2 cuando $x = x^*$ se estima como lo presenta Devore [5, p. 479]:

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot s_{\hat{Y}}, \qquad (16)$$

el cual está basado en el siguiente estadístico presentado por Devore [5, p. 478] que tiene distribución t con $n-2$ grados de libertad:

$$T = \frac{\hat{Y} - E(\hat{Y})}{S_{\hat{Y}}}. \qquad (17)$$

Por otra parte, para calcular un *Intervalo de predicción* para una observación y futura de la concentración de CO_2 cuando $x = x^*$ tenemos la ecuación:

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s_{\hat{Y}}^2}, \qquad (18)$$

el cual está basado en el siguiente estadístico que también tiene una distribución t con $n-2$ grados de libertad, revisar Devore [5, p. 482]:

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{\sqrt{s^2 + s_{\hat{Y}}^2}}. \qquad (19)$$

En la figura 4 aparecen los intervalos de confianza para la concentración promedio esperada y un valor de predicción y. Observe que el intervalo de confianza para la concentración promedio esperada es mucho más estrecho que el intervalo de confianza para la predicción de observaciones, ya que a medida que el valor x se acerca al promedio de los datos ($\bar{x}$) la desviación del estadístico utilizado es menor (15).

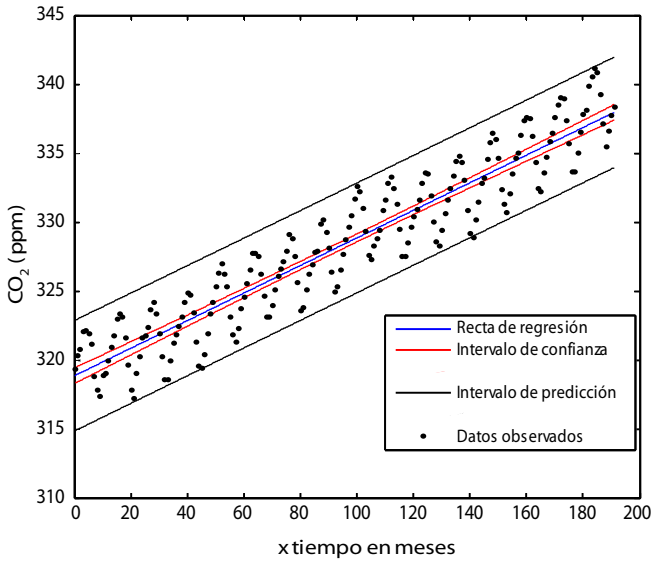

Figura 5. Comparación de intervalos de confianza para $E(Y/x^*)$ e intervalo de predicción y.

La figura 5 nos permite comparar el comportamiento de los intervalos de confianza en los distintos tiempos x's. Vemos que a medida que el tiempo se aleja del centro, ambos intervalos se expanden, siendo el intervalo de confianza para el promedio mucho más estrecho que el de predicción, tal como se dijo previamente. Las longitudes mínimas y máximas de los intervalos de confianza y de predicción resultaron respectivamente 0.5733, 1.1356, 7.9642 y 8.0243, las cuales se incluyeron en este análisis.

CONCLUSIONES

Este trabajo presentó una aplicación de los modelos de regresión lineal para estimar y pronosticar la tendencia de la concentración de CO2 emitida por el volcán Mauna Loa con respecto al tiempo. Se mostró que la técnica de regresión es útil y adecuada como modelo pronosticador. Además, se presentó el error de estimación y se analizó el comportamiento de un intervalo de confianza para la concentración promedio de CO2 y de un valor de predicción en cualquier momento. El trabajo fue presentado de manera didáctica para estudiantes de ciencias exactas y naturales, como prototipo de artículo de investigación donde se aplique el modelo de regresión lineal simple, aunque también puede servir para orientar a estudiantes de algunas áreas donde se enseñen tanto la parte descriptiva como de inferencia de este modelo de regresión.

BIBLIOGRAFÍA

1) F. Galton, «Regression towards mediocrity in hereditary stature,» *Anthropological Miscellanea*, 1889.
2) J. Lattin, J. D. Carroll, P. E. Green, «Analyzing Multivariate Data,» Belmont, CA: Duxbury Applied Series, 2002.
3) N. R. Draper, H. Smith, «Applied Regression Analysis,» New York: 3rd Ed., Wiley, 1998.
4) C. Pérez, «Técnicas de Análisis de Datos con SPSS,» Madrid: Pearson Prentice Hall, 2009.
5) J. L. Devore, «Probabilidad y Estadística para Ingeniería y Ciencias,» México: Séptima Edición, Cengage Learning, 2008.