Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Acatlan
Matemáticas Aplicadas y Computación

“Regresión Lineal: Análisis de la pobreza en México”

Integrantes :

-Cano Verduzco Monica 100%
-Palomares Olegario Alexis 100%
-Arreola Calderón Jesús Enrique 100%

Estadística II

Jaime Vergara Prado

Grupo : 2651

Índice

Tabla de imágenes

Introducción:

Implementación del modelo de regresión lineal, para el análisis de pobreza en México en los años 2014 y 2016

El objetivo será por medio de la regresión lineal analizar los datos e identificar y cuantificar la relación entre distintas variables y la tasa de pobreza en México durante los años 2014 y 2016, con el fin de entender los factores que contribuyen a la pobreza.

Palabras clave:

- Variable dependiente
- Regresión lineal simple
- Regresión lineal múltiple
- Coeficientes

Mediante los datos ya proporcionados, calcularemos los residuos de los mismos datos, el anova y el cálculo de los errores, para una vez teniendo este dato podremos estimar un intervalo de confianza, con estos datos podremos verlo en una gráfica y así poderlo visualizar de una mejor manera, del mismo modo podremos hacer una prueba F que sirve para saber si el conjunto si es significativo o no.

Todo esto lo hacemos para poder ver la relación de nuestros datos en relación a su año, teniendo la variable dependiente el año 2016 y la independiente el 2014, conociendo la relación de los datos y tomando en cuanta nuestros intervalos de confianza, más adelante podremos realizar pronósticos, para eso nos sirve conocer todas estas pruebas.

Marco teórico:

> “la ley dela regresión universal” fue introducido por Francis Galton. Estudio las alturas de padres he hijos a partir de mil registros familiares, llego a la conclusión de que

padres muy altos tenían una atendencia a tener hijos que heredaban parte de esa estatura, pero que revelaban también una tendencia a regresar a la media.

La primera forma de regresiones lineales fue el método de los mínimos cuadrados, por Legendre en 1805(pag.3 Prezi)

Este método se emplea para ajustar rectas a conjuntos de datos presentados como puntos en un plano. Proporciona un criterio para obtener la mejor recta que represente los puntos dados.

Una variante inicial de la regresión fue la regresión cuantil, la cual modela la relación entre un conjunto de variables predictoras e intervalos específicos (o "cuantiles") de una variable objetivo, generalmente la mediana. Sus ventajas respecto a la regresión de mínimos cuadrados ordinarios incluyen la falta de suposiciones sobre la distribución de la variable objetivo y una mayor resistencia a la influencia de observaciones atípicas

Mínimos cuadrados

En términos históricos, los mínimos cuadrados constituyen una técnica de análisis numérico enmarcada dentro de la optimización matemática. Busca encontrar la función que mejor se ajuste a un conjunto de datos, minimizando el error cuadrático. En su forma más básica, intenta minimizar la suma de los cuadrados de las diferencias entre los puntos generados por la función y los datos correspondientes. Esta técnica, también conocida como mínimos cuadrados promedio (LMS), se aplica cuando hay un solo dato medido y utiliza el método del descenso por gradiente para minimizar el residuo cuadrado. Aunque LMS minimiza el error cuadrático esperado con un mínimo de operaciones por iteración, requiere numerosas iteraciones para converger

Regresión lineal

En el contexto del aprendizaje automático, la tarea de la regresión implica predecir un parámetro (Y) a partir de otro parámetro conocido (X). Los modelos de regresión lineal son ampliamente utilizados en diversas áreas de investigación debido a su rapidez y facilidad de interpretación.Existen dos tipos principales de regresión lineal:

- Regresión lineal simple

En una regresión lineal, se trata de establecer una relación entre una variable independiente y su correspondiente variable dependiente. Esta relación se expresa como una línea recta. No es posible trazar una línea recta que pase por todos los puntos de un gráfico si estos se encuentran ordenados de manera caótica. Por lo tanto, sólo se determina la ubicación óptima de esta línea mediante una regresión lineal. Algunos puntos seguirán distanciados de la recta, pero esta distancia debe ser mínima. El cálculo de la distancia mínima de la recta a cada punto se denomina función de pérdida.

$$y=\beta_0+\beta_1 x+\varepsilon$$

$$\beta_0, \beta_1 \rightarrow son\ dos\ constantes\ desconocidas$$
$$\varepsilon \rightarrow la\ funcion\ de\ perdida$$

$$y \rightarrow una\ variable\ independiente$$

Limitaciones de la regresión lineal simple:

La regresión lineal simple establece que existe una relación entre las variables, pero no revela una relación causal: Y depende de, pero no implica que genere a Y.

- Regresión lineal múltiple

La regresión lineal múltiple encuentra la relación entre dos o más variables independientes y su correspondiente variable dependiente.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

$$Y \rightarrow es\ la\ variable\ dependiente.$$

$$X \rightarrow es\ una\ variable\ independiente.$$

$$\beta \rightarrow son\ coeficientes.$$

$$\varepsilon \rightarrow la\ función\ de\ pérdida.$$

Este tipo de regresión permite predecir tendencias y valores futuros. El análisis de regresión lineal múltiple ayuda a determinar el grado de influencia de las variables independientes sobre la variable dependiente, es decir, cuánto cambiará la variable dependiente cuando cambiemos las variables independientes. (Regresión Lineal: qué es, para qué sirve, por qué es importante, tipos y ejemplos de uso, s.f.)

El término lineal se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática 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 problema de la regresión consiste en elegir unos valores determinados para los parámetros desconocidos

Los valores escogidos como estimadores de los parámetros, son los coeficientes de regresión sin que se pueda garantizar que coincida n con parámetros reales del proceso generador. (Revista estadistica/ Regresion Lineal, s.f.)

Deducción del modelo de regresión lineal:

En su forma más general:

$$Y = f(X1, X2, X3 \ldots.., Xk; \beta)$$

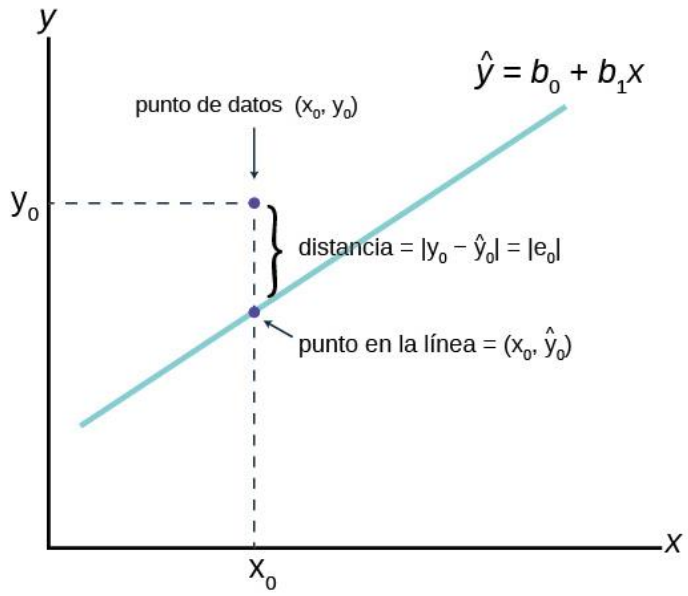

Figure 1. Grafica de la relación lineal

$$el término\ y\,0 - \hat{y}\,0 = e\,0$$

$$se\ denomina\ \text{error}\ o\ residual$$

La suma de los errores al cuadrado (SSE):

$$\hat{y} = b0 + b1\,x$$

$$donde\ b0 = y - - b1\,x -$$

$$y\,b1 = \frac{[\Sigma(x - x -)(y - y -)]}{\Sigma(x - x -)2} = \frac{[cov(x, y)]}{sx\,2}$$

Error estándar de la estimación:

$$s2\,a = \frac{[\Sigma(yi - \hat{y}i)^2]}{n} - k = \frac{[\Sigma ei^2]}{n} - k$$

Propiedades de estimadores:

Insesgamiento:

Un escenario ideal en la estimación puntual es que su estimador en promedio sea muy parecido al parámetro, luego un estimador se dice insesgado si y solo si se cumple que:

$$E(\hat{\theta}) = \theta$$

Observación: Si el valor esperado del estimador no es el parámetro, es decir, $E(\hat{\theta}) \neq \theta$, el estimador no es insesgado o se dice que tiene sesgo.

El sesgo se define como sigue:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

De donde también se define una estadistica de error muy importante, el error cuadrático medio. notado como $M S E_\theta$ y escrito de la siguiente manera:

$$MSE_{\theta} = V(\theta) + [B(\hat{\theta})]^2$$

Consistencia:

Cuando el estimador no es insesgado en primera medida, lo que sería lo idóneo, se requiere al menos que su valor oscila cerca del valor del parámetro para tamaños de muestra grandes, es decir, un estimador es consistente cuando:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta$$
$$\lim_{n\to\infty} V(\hat{\theta}) = 0$$

Suficiencia:

Intuitivamente hablando un estimador es suficiente para un parámetro si toda la informaciōn acerca del parámetro està contenida en la muestra.
Formalmente seria: una estadistica $\hat{\theta}$ se dice suficiente para θ basada en una muestra aleatoria $x_1, x_2, \ldots, n$ de una población con función masa o de densidad de probabilidad $f_x(x, \theta)$ Si la distribución condicional de las variables aleatorias $x_1, x_2, \ldots, n$ dado $\hat{\theta}$ no depende del parámetro θ, es decir, $\hat{\theta}$ es un estimador suficiente de θ si:
$$

$$f_x(x_1, x_2, \ldots, n|\hat{\theta} = \theta) = g(\hat{\theta})$$

Donde $g(\hat{\theta})$ = $(x_1, x_2, \ldots, n)$

Eficiencia:

La eficiencia es un requisito de precisión. esto es, es más preciso aquel estimador que tenga menor varianza ya que tiene la capacidad de producir estimaciones más centradas. Así sean $\hat{\theta}$ y $\hat{\theta}$ ' dos estimadores insesgados para θ. estimadores basados en una muestra aleatoria$x_1, x_2, \ldots, n$ de una población con función masa o de densidad de probabilidad $f_x(x, \theta)$, se dice que $\hat{\theta}$ es estimador uniformemente mejor que $\hat{\theta}$ ' si:

$$V(\hat{\theta}) < V(\hat{\theta}')$$

Pruebas de comprobación de supuestos

La comprobación de supuestos en estadística es asegurarse de que las condiciones básicas de un modelo estadístico sean válidas para los datos que estamos analizando. Estos supuestos pueden incluir cosas como la normalidad de los datos o la consistencia de las varianzas entre grupos. Es importante verificar estos supuestos antes de confiar en los resultados de nuestro análisis, ya que si no se cumplen, podríamos obtener conclusiones incorrectas. Usualmente, lo hacemos mediante gráficos y pruebas estadísticas específicas. Si encontramos que un supuesto no se cumple, es posible que necesitemos ajustar nuestro modelo o considerar otras formas de análisis. En resumen, la comprobación de supuestos nos ayuda a asegurarnos de que nuestros análisis sean sólidos y confiables.

> Uno de los supuestos de un anova y otras pruebas paramétricas es que las desviaciones estándar dentro del grupo de los grupos son todas iguales (exhiben homocedasticidad). Si las desviaciones estándar son diferentes entre sí (exhiben heterocedasticidad), la probabilidad de obtener un resultado falso positivo aunque la hipótesis nula sea verdadera puede ser mayor que el nivel alfa deseado.(H. McDonald, párrafo 1)

Para verificar ambos supuestos usamos el test de Breush-Pagan y el de White:

Test de Breush Pagan

- Si asumimos que existe una relación entre u^2 y X_j que puede ser lineal, es posible contrastar una restricción del tipo
$$u^2 = \delta_0 + \delta_1 X_1 + \ldots\ldots + \delta_k X_k + v$$
- **Contraste**: $H_0: \delta_1 = \delta_2 = \ldots = \delta_k = 0$
- Problema: el término de error no es observable, pero podemos utilizar los residuos MCO para esta regresión
- Después de regresar $\hat{u}^2$ sobre todas las X podemos usar el R^2 para construir el estadístico
- El estadistico F es igual que el estadístico que contrasta la significatividad global de la regresión, $F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \sim F_{k,n-k-1}$
- El estadístico LM es $LM = nR^2 \sim X_k^2$

Test de White

- Problema: el test de Breush-Pagan sólo detecta formas lineales de heterocedasticidad
- Para resolverlo, el test de White permite contrastar no linealidades utilizando los cuadrados y los productos cruzados de todos los regresores. Si $k = 3$,
$\hat{u}^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_1^2 + \delta_5 X_2^2 + \delta_6 X_3^2 + \delta_7 X_1 X_2 + \delta_8 X_1 X_3 + \delta_9 X_2 X_3 + v$
- **Contraste:** $H_0: \delta_1 = \delta_2 = \ldots = \delta_9 = 0$

- El estadístico F y el LM nos permite contrastar si todas las X_j, X_j^2 y $X_j X_h$ son conjuntamente significativas

T-test

El t-test es una prueba estadística que se utiliza para determinar si hay una diferencia significativa entre las medias de dos grupos. Se utiliza comúnmente para comparar la media de una muestra con un valor conocido (t-test de una muestra) o para comparar las medias de dos muestras independientes (t-test de dos muestras). También se puede utilizar para comparar las medias de dos muestras relacionadas (t-test pareado).

- Correlación con la media cero: En el contexto del t-test, la correlación con la media cero se refiere a la suposición de que las diferencias entre las observaciones y la media muestral son independientes y tienen una media de cero.

Durbin-Watson Test

El test de Durbin-Watson es una prueba de autocorrelación que se utiliza para determinar si hay correlación serial de primer orden en los residuos de un modelo de regresión. Es decir, evalúa si existe una relación lineal entre los residuos en diferentes puntos en el tiempo.

.

- Para no correlaciones: Si el valor del estadístico de Durbin-Watson está cerca de 2, sugiere que no hay correlación serial significativa en los residuos del modelo.

.

Anderson-Darling Test

El test de Anderson-Darling se utiliza para evaluar si una muestra de datos proviene de una población con una distribución específica, como la distribución normal. Es una prueba de bondad de ajuste que se enfoca en diferencias en las colas de la distribución.

- Para distribuciones normales: En el contexto del test de Anderson-Darling, se utiliza para determinar si una muestra de datos sigue una distribución normal. Si el valor-p asociado al test es mayor que un nivel de significancia dado, no hay suficiente evidencia para rechazar la hipótesis nula de que los datos se distribuyen normalmente.

Mapa conceptual

Resumen:

La pobreza en México es un problema que puede llevar a causar grandes revueltas sociales en el país, es por eso que es necesario el estudio de esta ya que podemos estimar en años futuros como se comporta esta a lo largo del tiempo, nuestro años de estudio son del 2014 y 2016.
Esta problemática nos plantea una pregunta inicial ¿Qué relación existe entre los índices de pobreza del 2014 y 2016?. Para resolver este planteamiento es necesario realizar un estudio estadístico entre estas variables existentes, dicho estudio es saber que relación hay

entre estas variables para así poder estimar en años futuros como se comportará el índice de pobreza.
Entre las varias técnicas estadísticas que hay se usó el modelo ANOVA en conjunto con el software estadístico R que nos ayudó a agilizar los cálculos. Con estos métodos se buscó un modelo de regresión lineal adecuado. Para comprobar dicho modelo es necesario de pruebas más específicas como el test de Breush-Pagan, White, T-test, Durbin- Watson y el Anderson-Darling. Dichos test tienen que cumplir la condición del p-value para verificar de manera correcta el modelo la cuál tienen que ser menor que un alpha. Se propone una media cero.
Uno de los principales hallazgos que se encontró fue que la media fue distinta de cero lo cuál significa que existe el modelo. El modelo pasa la prueba T-test, Anderson-Darling, Breush-Pagan pero en la prueba Anderson- Darling no se rechaza ni se aprueba nuestra hipótesis nula.
Por lo cuál podemos afirmar que existe un modelo que podemos llegar a utilizar.

Empleamiento del modelo:

Para el caso de estudio de la evolución de la pobreza y pobreza extrema nacional y en entidades federativas, 2010,2012, 2014 y 2016. Primero planteamos las hipótesis.

$H_0: \theta = 0$ no hay autocorrelación
$H_a: \theta \neq 0$ existe autocorrelación
nivel de significancia $\alpha = 0.05$

Obtención del modelo:

```
entidades <- c("Aguascalientes", "Baja California", "Baja California Sur", "Campeche", "Coahuila", "Colima", "Chiapas", "Chihuahua", "Distrito Federal", "Durango
X <- c(442.9, 984.9, 226.2, 391.0, 885.8, 244.9, 3961.0, 1265.5, 2502.5, 761.2, 2683.3, 2315.4, 1547.8, 2780.2, 8269.9, 2708.6, 993.7, 488.8, 1022.7, 2662.7, 39
Y <- c(369.7, 789.1, 175.6, 405.0, 745.9, 248.7, 4114.0, 1150.0, 2434.4, 643.3, 2489.7, 2314.7, 1478.8, 2560.6, 8230.2, 2565.9, 965.9, 470.1, 737.8, 2847.3, 372

  MRL <- lm(Y~X)
```

Figure 2. Presentación de los datos en R

Utilizando la función summary en R para obtener el p-value:

```
summary(lm_model)
```

```
all:
m(formula = Y ~ X)

esiduals:
   Min      1Q  Median      3Q     Max
213.52 -102.57    6.26   50.15  419.35

oefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept) -93.46671   34.49160   -2.71    0.011 *
             1.01930    0.01449   70.32   <2e-16 ***
--
```

Figure 3. Obtencion del p-value en R

Como el p-value es menor al nivel de significancia designado, podemos rechazar la hipótesis nula, esto sugiere que hay una relación o efecto estadísticamente significativo en los datos.

Aún así se aplican 4 pruebas para verificar que se pueda plantear un modelo de regresión lineal:
Para la media distinta de cero, obtenemos los residuos, que son las diferencias entre los valores de la variable dependiente observados y los valores que predecimos a partir de nuestra recta de regresión.

```
res <- c(residuals(MRL))
plot(res)
abline(h=0,lty=2,col="red") ##Grafica para la media cero
t.test(res,alternative=c("two.sided"), conf.level = 0.95, mu = 0)
```

```
        One Sample t-test

data:  res
t = -1.0927e-16, df = 31, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -47.53345  47.53345
sample estimates:
    mean of x
-2.546791e-15
```

Figure 4. Código para representar los residuales en R

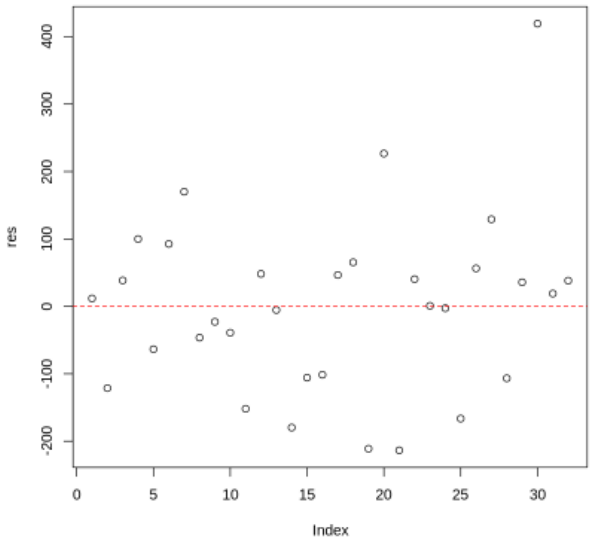

Figure 5. Gráfica de los residuales en R

Pasa la prueba con una media distinta de cero.

Ahora buscamos verificar la varianza constante, para ello usamos el Breusch-Pagan test.

```
bptest(MRL)
```

```
        studentized Breusch-Pagan test

data:  MRL
BP = 6.3333, df = 1, p-value = 0.01185
```

Figure 6. Breush- Pagan test en R

El cual nos da un p-value que niega la hipótesis nula.

Posteriormente realizamos el test de Durbin-Watson para comprobar la no correlación.

```
res <- c(residuals(MRL))
acf(res)
dwtest(MRL)
```

```
        Durbin-Watson test

data:  MRL
DW = 2.3366, p-value = 0.8253
alternative hypothesis: true autocorrelation is greater than 0
```

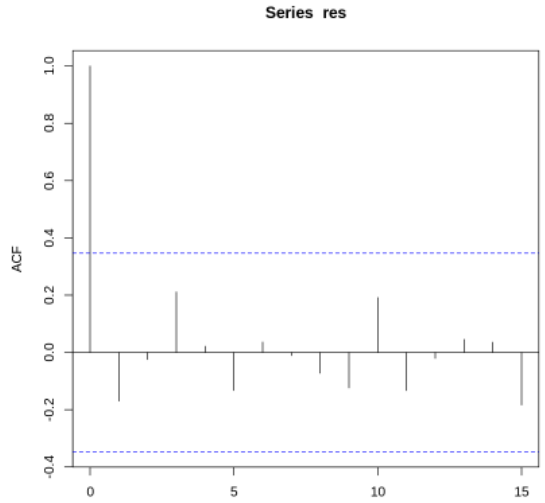

Figure 7. Durbin- Watson test en R

Podemos confirmar que existe verdadera correlación.

Finalmente usamos el Anderson-Darling para verificar que se distribuye normalmente.

```
[14] res <- c(residuals(MRL))

     #plot(MRL,2)

     ad.test(res)
```

```
        Anderson-Darling normality test

data:  res
A = 0.4773, p-value = 0.2214
```

Figure 8. Anderson-Darling test en R

Como el p-value = 0.2214 esto indica que no hay suficiente evidencia para rechazar la hipótesis nula. Esto significa que los datos podrían seguir una distribución normal.

Conclusión: Decidimos plantear el modelo de regresión lineal dados los coeficientes de -93.4667 en el intercepto y 1.01929 en x, con la siguiente gráfica:

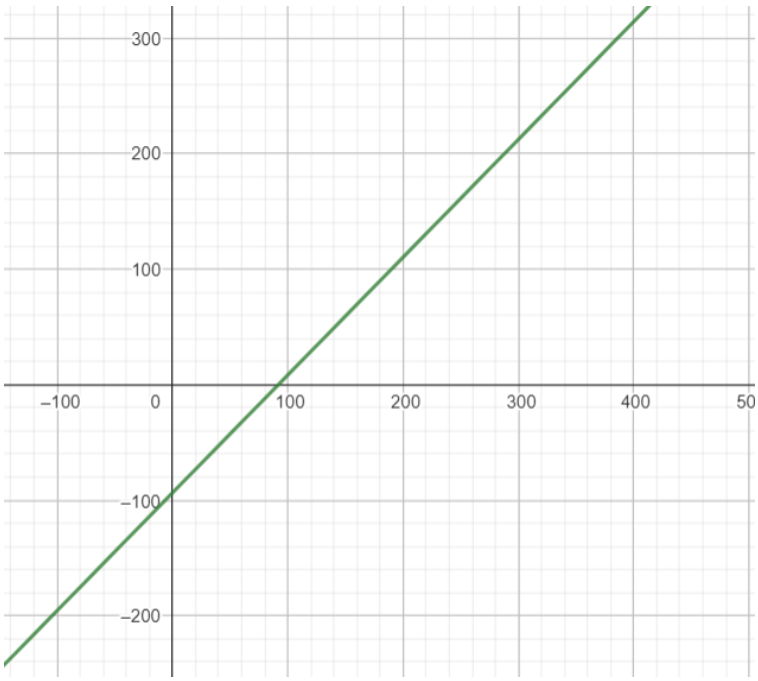

Figure 9. Gráfica de la regresión lineal

Bibliografía:

McDonald, J. H. (s.f.). Homocedasticidad y Heterocedasticidad. En Estadísticas Biológicas. Recuperado de https://espanol.libretexts.org/Estadisticas/Estadistica_Aplicada/Libro%3A_Estadisticas_Biologicas_(McDonald)/04%3A_Pruebas_para_una_variable_de_medici%C3%B3n/4.05%3A_Homocedasticidad_y_Heterocedasticidad

Lozano, A. (2010, 25 de noviembre). Regresion Lineal. Prezi. Recuperado de https://prezi.com/qpxysldecq4u/regresion-lineal/

IBM Documentation. (s.f.). IBM in Deutschland, Österreich und der Schweiz. Recuperado de https://www.ibm.com/docs/es/spss-statistics/saas?topic=regression-quantile

Grupo de trabajo G942 de la UNICAN. (2014-2015). Ppt_Ch6_G942_14-15.pdf [Archivo PDF]. Recuperado de https://ocw.unican.es/pluginfile.php/1098/course/section/691/Ppt_Ch6_G942_14-15.pdf

Regresión Lineal: qué es, para qué sirve, por qué es importante, tipos y ejemplos de uso. (s.f.). Ebac. Recuperado de https://ebac.mx/blog/regreson-lineal

Revista estadística/ Regresión Lineal. (s.f.). Issuu. Recuperado de https://issuu.com/crismarcontreras6/docs/revista_estadistica