“Análisis de regresión lineal”

M.Sc. Henry Luis López García

@Hen1985 hlopez@unan.edu.ni

Contenidos

- Modelo de regresión lineal
- Estimación de lo (β_0,β_1)
- Propiedad de los estimadores por mínimos cuadrados
- Estimación de (σ^2)
- Prueba de significación del modelo de regresión
- Diagnósticos de los residuos
- Coeficiente de correlación
- Estimación de (ρ)
- Prueba de significancia del coeficiente de correlación muestral (ρ)

Regresión lineal

La regresión es una técnica estadística para investigar y modelar la relación entre variables.

Propósito de la regresión lineal:

- Describir la regresión lineal entre y & x
- Determinar cuanta variación en y puede ser explicada por la relación con x
- Predecir valores nuevos de y usando nuevos valores de x

Regresión lineal

Consideremos el modelo de regresión lineal simple un modelo con solo un regresor x que tiene una relación con una respuesta y, donde la relación es una línea recta.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (ecuación\ 1)$$

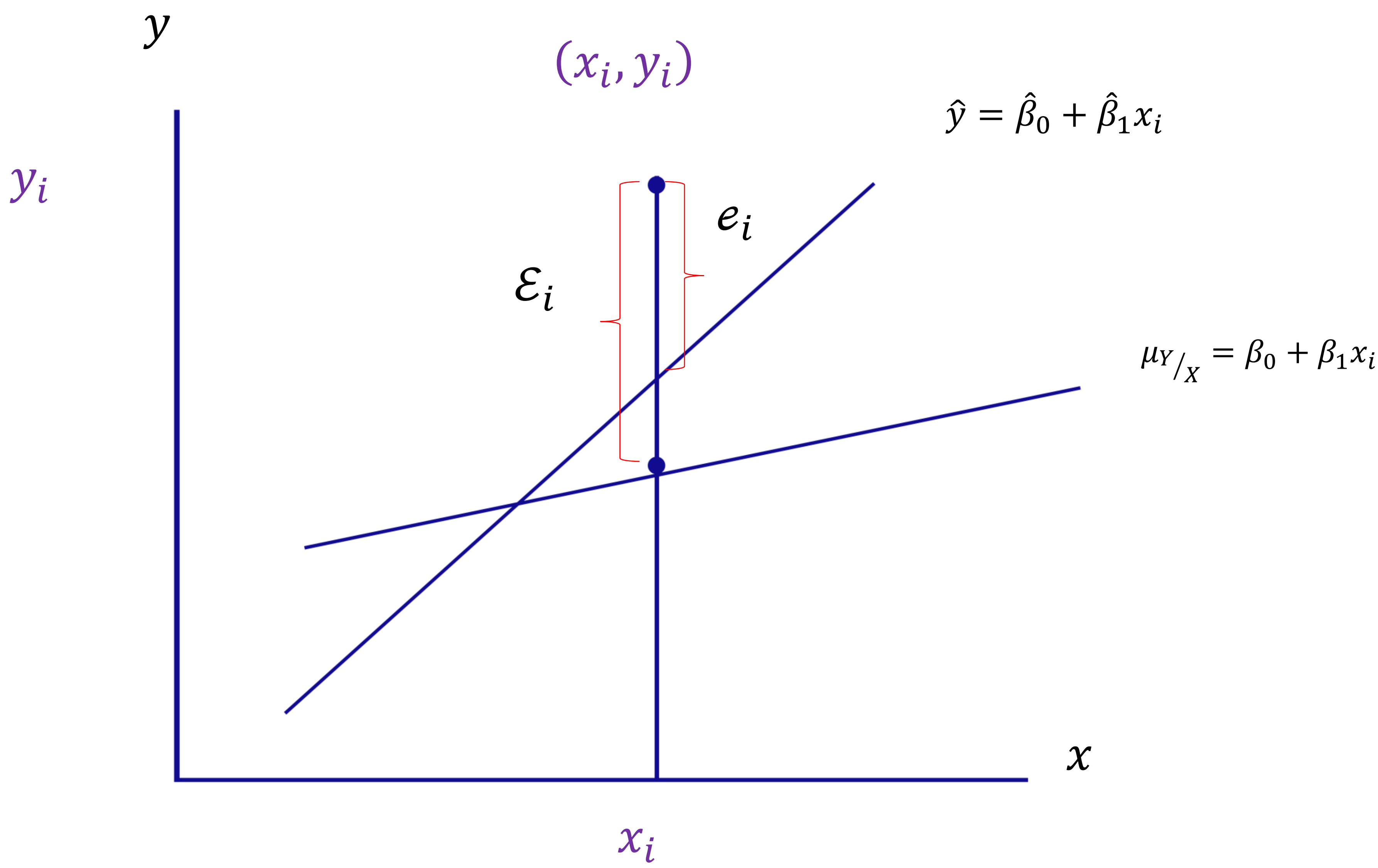
Por tanto observando el modelo:

- Necesitamos estimar dos parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.
- $\hat{\beta}_0$ es el intercepto, la media de la distribución de probabilidad de y cuando x es 0.

Regresión lineal

- $\hat{\beta}_1$ es a menudo llamado la pendiente, mide la tasa de cambio en y por una unidad de cambio en x.

- La estimación de los parámetros es a través de los mínimos cuadrados, lo que resuelve este modelo por medio minimizar e_i realmente $\sum_{i:1}^{n} e^2{}_i$.



y
(x_i, y_i)
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$
y_i
e_i
ε_i
$\mu_{Y/X} = \beta_0 + \beta_1 x_i$
x
x_i

Estimación de (β_0y β_1)

- Los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$ son desconocidos y se deben de estimar con los datos de la muestra, ahora supongamos que hay n pares de datos: $(y_1, x_1), (y_2, x_2), \ldots\ldots (y_n, x_n)$.

- Entonces para estimar $\hat{\beta}_0$ y $\hat{\beta}_1$ se usa el método de mínimo cuadrados, esto es, estimar $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima, según la ecuación puede escribirse

Estimación de (β_0y β_1)

- Considerando la $(y = \beta_0 + \beta_1 x + \varepsilon)$ es un modelo de regresión poblacional mientras que la ecuación 2, es un modelo muestral de regresión, escrito en términos de los n pares de datos(y_i, x_i) i =1, 2, ,.......,n

$s(\beta_0, \beta_1)$=$\sum_{i:1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ $\sum_{i:1}^{n} e^2{}_i$

- Los estimadores por mínimos cuadrados, de β_0 y β_1, que se designarán por $\hat{\beta}_0$ y $\hat{\beta}_1$, deben de satisfacer

Estimación de (β_0y β_1)

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\widehat{\beta}_0,\widehat{\beta}_1} = -2 \sum_{i:1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\widehat{\beta}_0,\widehat{\beta}_1} = -2 \sum_{i:1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0$$

Se simplifican estas dos ecuaciones se obtiene

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i:1}^{n} x_i = \sum_{i:1}^{n} y_i \qquad \hat{\beta}_0 \sum_{i:1}^{n} x_i + \hat{\beta}_1 \sum_{i:1}^{n} {x^2}_i = \sum_{i:1}^{n} y_i x_i$$

Ecuaciones normales de mínimos cuadrados

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (ecuación\ 3)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x^2{}_i - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (ecuación\ 4)$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Ecuaciones normales por mínimos cuadrados

Una forma cómoda de escribir la ecuación 4

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Suma corregida de los productos cruzados de las x_i & y_i

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

Suma corregida de cuadrados de las x_i

$$S_{xx} = \sum_{i=1}^{n} x^2{}_i - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Residual (e_i)

- La diferencia entre el valor observado y_i & el valor ajustado correspondiente $\hat{y}_i$ se le llama residual, matemáticamente, el $i - ésimo$ residual es

$$e_i = y_i - \hat{y} = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), \quad \mathrm{i} = 1, 2, \ldots \ldots, n$$

$(ecuación\ 5)$

- Los residuales tienen un papel importante en la adecuación del modelo de regresión ajustado, y para detectar diferencias respecto a las hipótesis básicas.

Propiedad de los estimadores

- Los es estimadores por mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen algunas propiedades importantes, estas se describen:

- La suma de los residuales en cualquier modelo de regresión que contenga una ordenada al origen $\hat{\beta}_0$ siempre es igual a cero, esto es.

$$\sum_{i:1}^{n}(y_i - \hat{y}_i) = \sum_{i:1}^{n} e_i = 0$$

Propiedad de los estimadores

- La suma de los valores observados y_i es igual a la suma de los valores ajustados $\hat{y}_i$

$$\sum_{i:1}^{n} y_i = \sum_{i:1}^{n} \hat{y}_i$$

- La línea de regresión de mínimos cuadrados siempre pasa por el **centroide** de los datos, que es el punto (y_i, x_i).

Propiedad de los estimadores

- La suma de los residuales, ponderados por el valor correspondiente de la variable regresora, siempre es igual a cero:

$$\sum_{i:1}^{n} x_i e_i = 0$$

- La suma de los residuales, ponderados por el valor ajustado correspondiente, siempre es igual a cero:

$$\sum_{i:1}^{n} \hat{y}_i e_i = 0$$

Estimación de σ^2

- Además de estimar $\hat{\beta}_0\ y\ \hat{\beta}_1$, se requiere un estimado se σ^2 para probar hipótesis y formar estimados de intervalos pertinentes al modelo de regresión, el estimado de σ^2 se obtiene de la suma cuadrado residuales, o suma cuadrado del error:

$$MSr_{Res} \sum_{i:1}^{n} e^2{}_i = \sum_{i:1}^{n} (y_i - \hat{y}_i)^2$$

- Se puede deducir una formula fácil de MSr_{Res} sustituyendo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimación de σ^2

$$SS_{Res} = \sum_{i:1}^{n} {y^2}_i - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

$$SS_T = \sum_{i:1}^{n} {y^2}_i - n\bar{y}^2 = \sum_{i:1}^{n} (y_i - \bar{y})^2$$

Es justo la suma de cuadrado corregida, de las observaciones de las respuestas por lo que:

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$$

Estimación de σ^2

- La suma de cuadrado de residuales tiene $n-2$ grados de libertad, porque dos grados de libertad se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{y}_i$, por lo que un estimador insesgado de σ^2 es

$$MS_{Res} = \frac{SS_{Res}}{n-2} = \hat{\sigma}^2$$

- La cantidad MS_{Res} se le llama cuadrado medio residual la raíz cuadrada de $\hat{\sigma}^2$, se llama el error estándar de la regresión y tiene la misma unidad que la variable de respuesta y.

Prueba de significancia de regresión

- La siguientes hipótesis se relacionan con la significancia de la regresión, el no rechazar la $\mathrm{H_0}: \beta_1 = 0$, implica que no hay relación lineal entre x & y. Esto puede implicar que x tiene muy poco valor para explicar la variación de y y que el mejor estimador x es $\hat{y} = \bar{y}$, o que la verdadera relación entre x & y no es lineal.

$$\mathrm{H_0}: \beta_1 = 0$$

$$\mathrm{H_1}: \beta_1 \neq 0$$

Prueba de significancia de regresión

- El procedimiento de prueba para $\mathrm{H}_0: \beta_1 = 0$, se puede establecer, tan solo usando el estadístico t, en la siguiente ecuación:

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}, donde \qquad se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$$

- Rechazando $\mathrm{H}_0: \beta_1 = 0$ si, $\quad |t_0| > t_{\left(\frac{\propto}{2}, n-2\right)}$

Prueba de significancia de regresión

- El procedimiento de prueba para $\mathrm{H}_0: \beta_0 = 0$, se puede establecer, tan solo usando el estadístico t, en la siguiente ecuación:

$$t_0 = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}, donde \qquad se(\hat{\beta}_0) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

- Rechazando $\mathrm{H}_0: \beta_0 = 0$ si, $\quad |t_0| > t_{\left(\frac{\alpha}{2}, n-2\right)}$

Supuesto del modelo

1. Debe haber una relación lineal y aditiva entre la variable dependiente (respuesta) y la(s) variable(s) independiente(s) (predictora). Una relación lineal sugiere que un cambio en la respuesta Y debido a un cambio unitario en X^1 es constante, independientemente del valor de X^1. Una relación aditiva sugiere que el efecto de X^1 sobre Y es independiente de otras variables.
2. No debe haber correlación entre los términos residuales (error). La ausencia de este fenómeno se conoce como Autocorrelación.
3. Las variables independientes no deben estar correlacionadas. La ausencia de este fenómeno se conoce como multicolinealidad.
4. Los términos de error deben tener varianza constante. Este fenómeno se conoce como homocedasticidad. La presencia de varianza no constante se denomina heteroscedasticidad.
5. Los términos de error deben tener una distribución normal.

Diagnósticos de los residuos

Como se pude considerar que un residual es la desviación entre los datos y el ajuste, también es una medida de variabilidad de la variable de respuesta que no explica el modelo de regresión. También conviene imaginar que los residuales son los valores realizados, u observados de los errores del modelo, por la que toda desviación de las premisas de los errores se debe reflejar en los residuos.

Diagnósticos de los residuos

- Residuales estandarizados, ya que la varianza aproximada de un residual se estima con MS_{Res}, el cuadrado medio de los residuales, un escalamiento lógico de los residuales sería el de los residuales estandarizados,

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \quad i = 1,2,3,\dots,n \qquad MS_{Res} = \frac{\sum_{i:1}^{n} e^2{}_i}{n-p}$$

- Los residuales estandarizados tiene media cero y varianza aproximadamente unitaria, en consecuencia un residual estandarizado grande $(d_i > 3)$.

Diagnósticos de los residuos

- Residuales estudentizados, Las violaciones de las premisas, del modelo, están con más probabilidad, en los puntos remotos, y pueden ser difíciles de detectar por inspecciones de los residuales ordinarios e_i por que en general sus residuales serán menores, entonces, un procedimiento lógico es examinar los residuales estudentizados,

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}, \quad i = 1,2,3, \dots, n \qquad h_{ii} = X(X^I X)^{-1} X^I$$

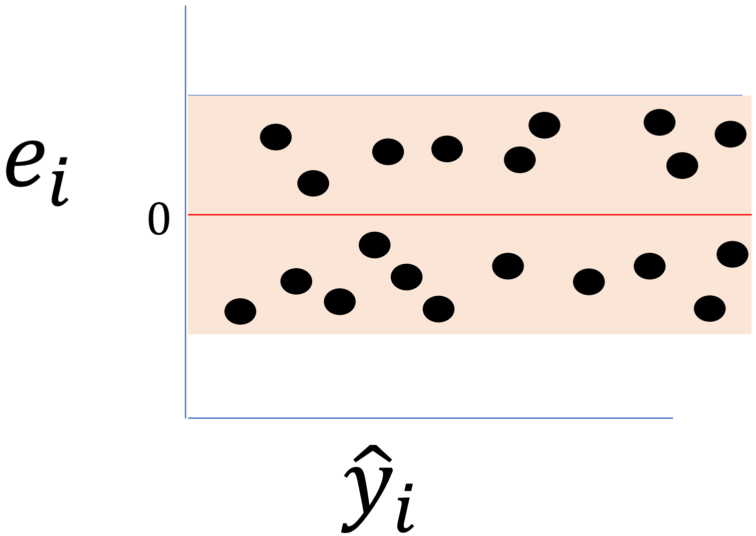
Agregar a X $\mathrm{V} = (1,1,1, \dots, 1)$

Diagnósticos de los residuos

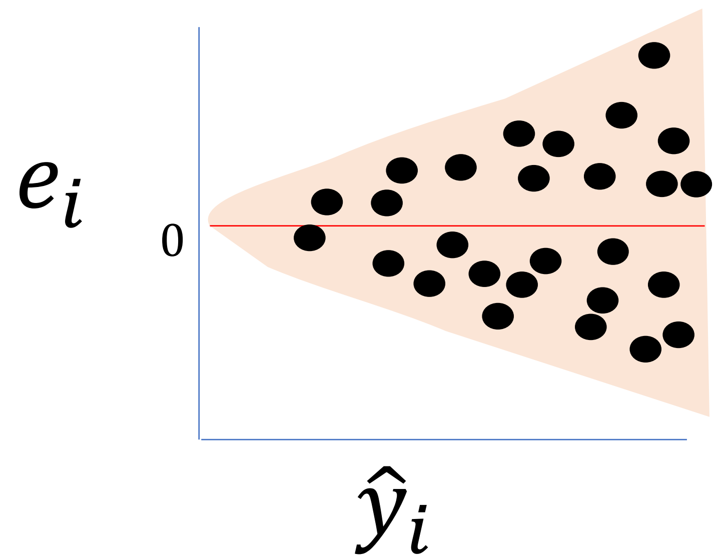
- Residuales PRESS, no es más que el residual ordinario ponderado por los elementos diagonal de la matriz de sombrero h_{ii}. Los residuales asociados con puntos para los h_{ii} es grande tendrán PRESS residuales grandes., estos serán por lo general puntos de gran influencia,

$$e_{(i)} = \frac{e_i}{1-h_{ii}}, \quad i = 1,2,3, \dots, n \qquad h_{ii} = X(X^I X)^{-1} X^I$$

Diagnósticos de los residuos

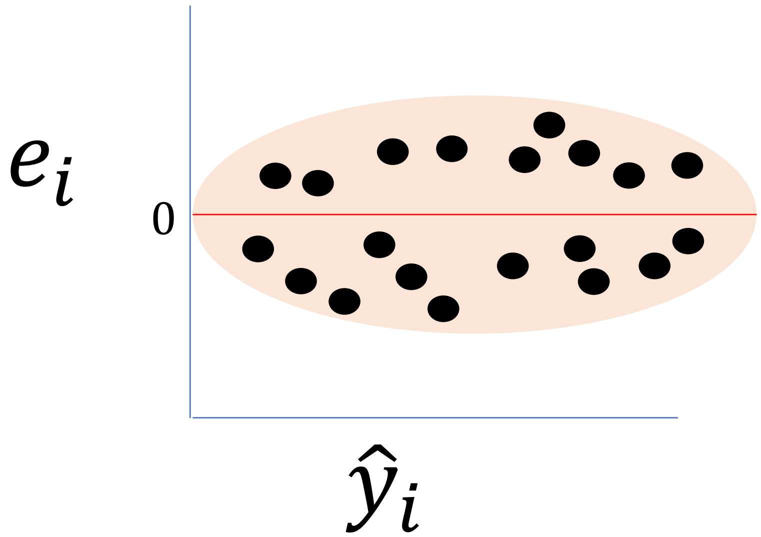


a)

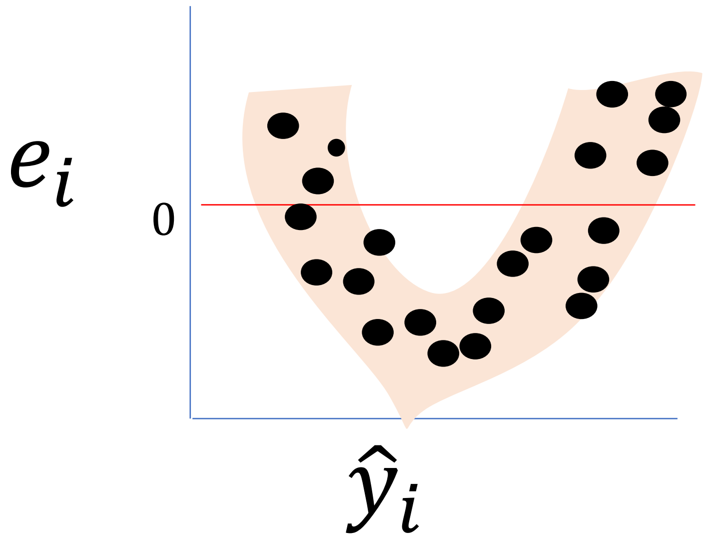


b)

Diagnósticos de los residuos



c)



d)

Bibliografía

1. Binek, R. (2015). Kosaciec szczecinkowaty Iris setosa [Image]. Retrieved from https://commons.wikimedia.org/wiki/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg#/media/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg
2. Chihara, L. M., & Hesterberg, T. C. (2018). *Mathematical Statistics with Resampling and R* (2nd ed.). Wiley.
3. Kloke, J., & McKean, J. W. (2014). *Nonparametric Statistical Methods Using R (Chapman & Hall/CRC The R Series Book 25) (English Edition)* (1.ª ed.). Chapman and Hall/CRC.
4. González, G. C., Liste, V. A., & Felpeto, B. A. (2011). *Tratamiento de datos con R, Statistica y SPSS* (1.ª ed.). Ediciones Diaz de Santos.
5. Rasch, D., Pilz, J., Verdooren, L. R., & Gebhardt, A. (2011). *Optimal Experimental Design with R (English Edition)* (1.ª ed.). Chapman and Hall/CRC.
6. Husson, F., Le, S., & Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R* (2nd ed.). CRC Press.
7. https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/?utm_source=twitter.com&utm_medium=social

“Análisis de regresión lineal”

M.Sc. Henry Luis López García

@Hen1985

hlopez@unan.edu.ni