MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- **CLASE 3: REGRESIÓN.**
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

Clase 3 - Parte 2: Métodos Estadísticos para IA

Análisis de residuos y
coeficiente de determinación

Objetivos de esta presentación

- En esta presentación continuaremos con el modelo de **regresión lineal simple**, pero las preguntas serán ahora: ¿Cuánto de mis datos son explicados por un modelo lineal? ¿Será necesario pensar en un modelo no lineal? ¿Puedo extender lo aprendido para modelos no lineales?
- Para la última pregunta **extenderemos** los resultados para algunos problemas particulares.
- Las primeras preguntas las podremos responder aprendiendo:
  - **Análisis de residuos**
  - **Coeficiente de determinación**

Resumen

Análisis de residuos

Coeficiente de determinación

Extensión para casos no lineales

Resumen

Análisis de residuos

Coeficiente de determinación

Extensión para casos no lineales

Análisis de residuos

- El **análisis de residuos** es una herramienta fundamental para validar un modelo, especialmente en regresión lineal y otros modelos estadísticos.
- Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo.

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \qquad (1)$$

para $i = 1, \ldots, n$. Se suele también observar el residuo estandarizado: $d_i = e_i / \sqrt{\hat{\sigma}^2}$, donde $\hat{\sigma}^2$ es estimado por la Ec. (10) de la presentación 3.1.

- Al analizar los residuos, se busca verificar si se cumplen ciertas **suposiciones clave** que justifican la validez del modelo.

Suposiciones que deben cumplirse en el análisis de residuos para justificar la validez del modelo

- Se asume que la relación entre las variables independientes y la variable dependiente es **lineal**.
- Los residuos deben ser **independientes entre sí**.
- La varianza de los residuos debe ser **constante** a lo largo de todos los niveles de las variables independientes[1].
- Se asume que los residuos están **normalmente distribuidos**, especialmente si se van a hacer test de hipótesis o construir intervalos de confianza.

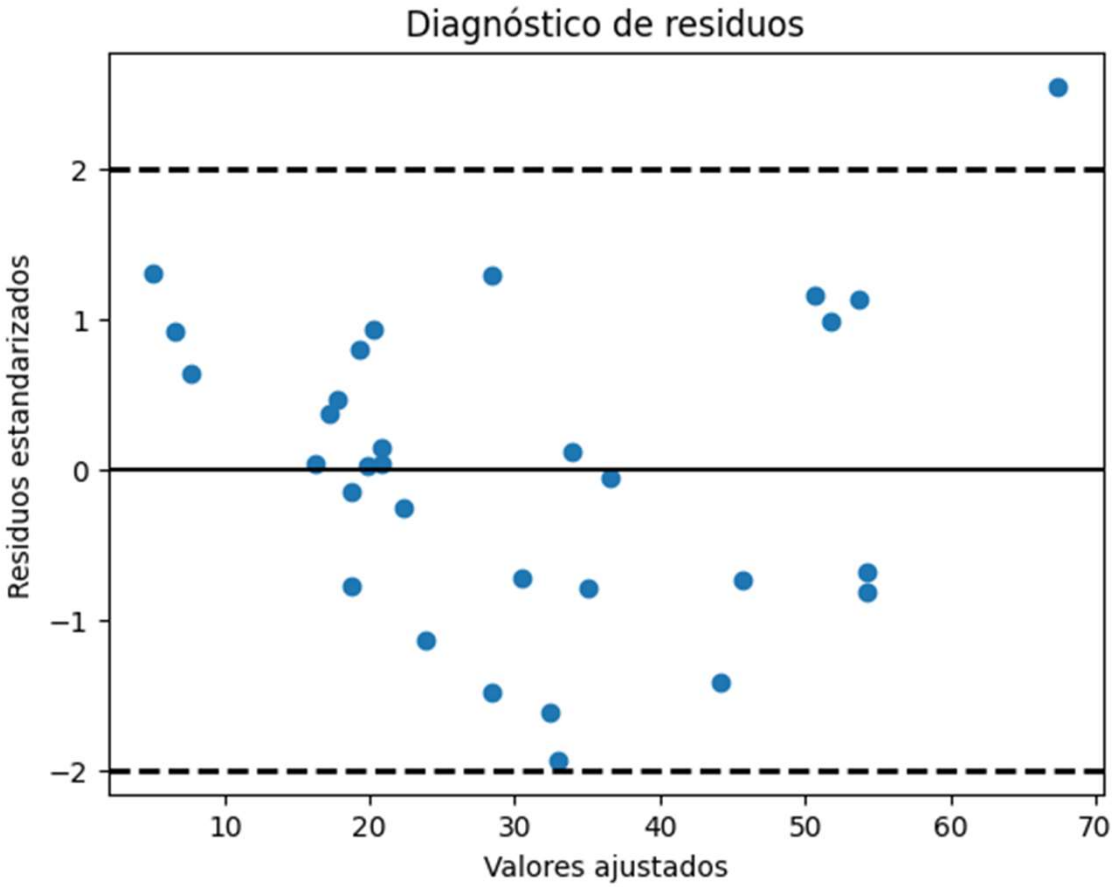
[1]Esto se denomina **homocedasticidad**.

Residuos normalmente distribuidos

- Si el modelo es correcto y los errores son normales, entonces los residuos estandarizados d_i se comportan como una variable aleatoria con distribución **normal estándar**, es decir: $d_i \sim N(0,1)$.
- Por lo tanto, aproximadamente el **95%** de los valores caen dentro del intervalo: [−1.96,+1.96]. Este intervalo es, de hecho, el **intervalo de confianza al 95%** para una observación de una variable normal estándar. En general, se toma el intervalo [-2,2].
- Si un residuo estandarizado **cae dentro de** [-2,2], es **consistente con el comportamiento esperado** de un error bajo la suposición de normalidad.
- Si un residuo estandarizado **cae fuera de ese rango**, entonces estaría **fuera del intervalo de confianza al 95%**, y eso **sugiere** (aunque no prueba) que:
  - Esa observación puede ser **atípica**
  - El modelo puede no ajustarse bien ahí.
  - La suposición de normalidad no se cumple en ese punto.

Residuo estandarizado

- Para graficar lo anterior en el script del ejemplo 19 continuamos con el estudio de la relación entre el volumen del árbol de guinda y el diámetro del tronco.
- En la figura se grafica d_i vs. $\hat{y}_i$.Para obtener el residuo y el residuo estandarizado en Python:




```
residuos_estandarizados = influence.resid_studentized_internal
valores_ajustados = modelo.fittedvalues
```

- Se ve que el valor correspondiente a $\hat{y}_{31}$= 67,41 pies cúbicos es un valor atípico.

Residuo estandarizado

En general, como receta se suelen dibujar las siguientes gráficas:

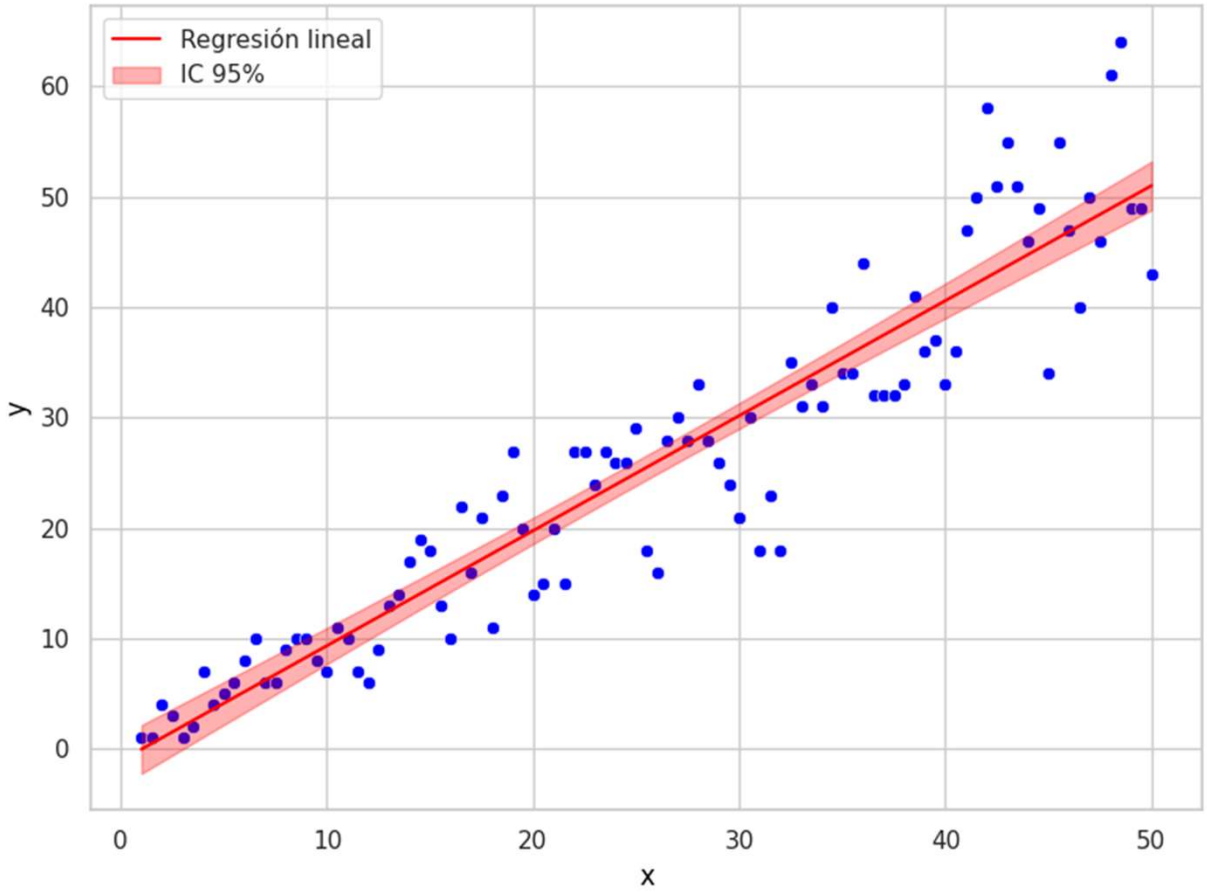
1. d_i vs. $\hat{y}_i$
2. d_i vs. x_i
3. d_i vs. el instante de tiempo en que fue adquirida la muestra (aunque a menudo no tenemos esa información).

Estas figuras nos ayudarán a detectar comportamientos que nos separen de las suposiciones formuladas anteriormente.
La figura anterior se comporta de manera correcta, salvo por el valor atípico, que en principio puede indicar algún error en la medición. Podría ser descartado, aunque primeramente habría que revisar el experimento.
El punto 3 es interesante para verificar que la medición no cambia con el tiempo, por ejemplo, es posible que la varianza crezca con el tiempo.

Residuo estandarizado: comportamiento de σ^2

Veamos otro ejemplo…

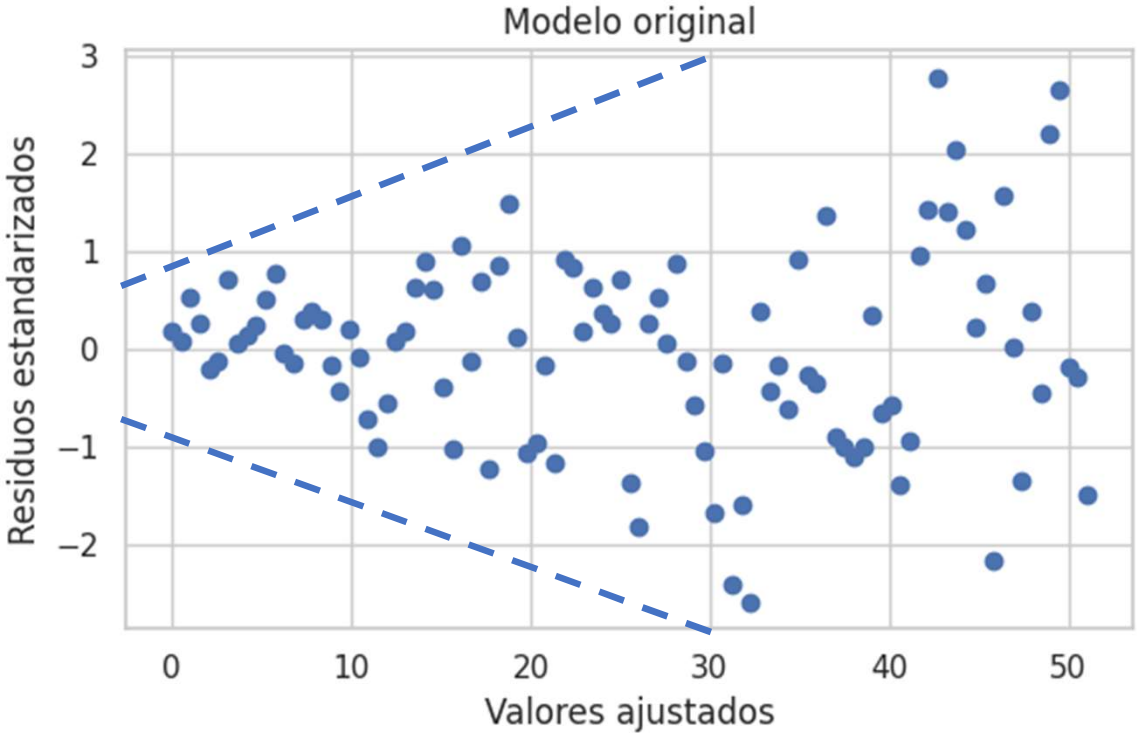
- En el script del ejemplo 20 se observan los siguientes datos:



- Ya se estimaron los parámetros, ahora verificaremos la suposición de varianza constante (σ^2 = cte).

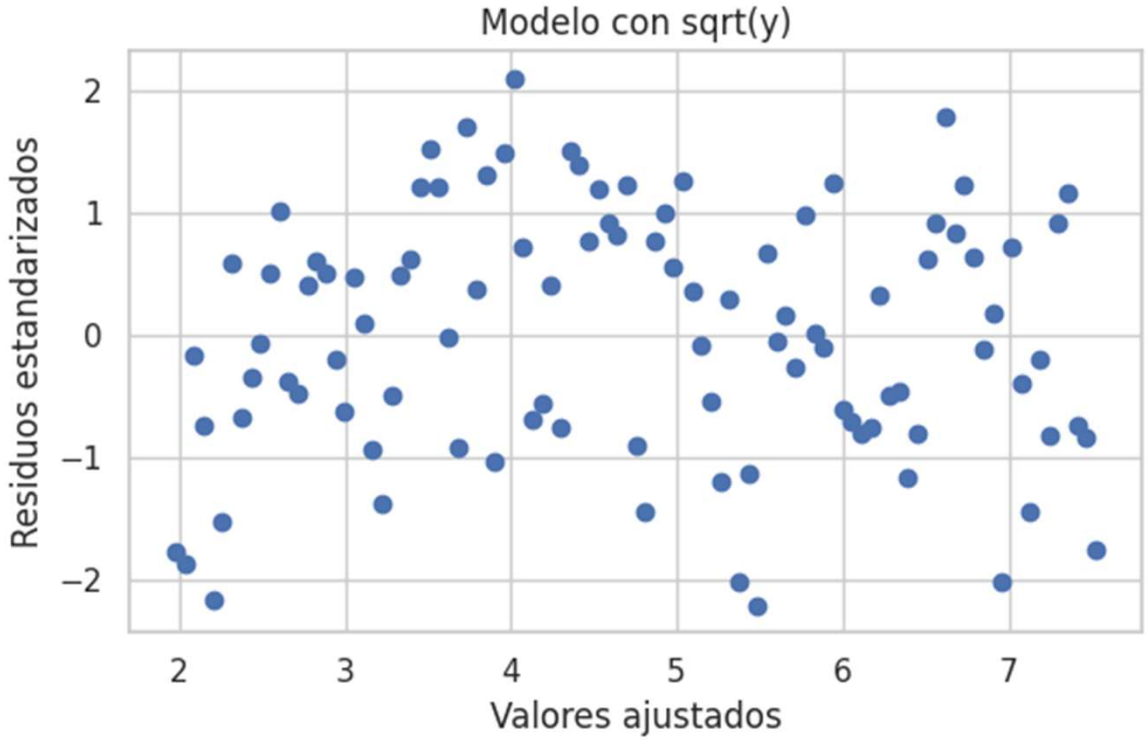
Residuo estandarizado: comportamiento de σ^2

- Si graficamos d_i vs. $\hat{y}_i$, se obtiene la siguiente figura.
- Las líneas punteadas marcan una tendencia: la dispersión parece crecer. Es decir, cuando crece $\hat{y}_i$ crece la varianza. Esto se puede corregir.
- Este problema es habitual cuando los datos provienen de variables aleatorias cuya varianza depende de la media. Este es el caso del ejemplo 20, donde los datos provienen de una v.a. Poisson $(E\{X\} = V\{X\} = \lambda)$.



Residuo estandarizado: comportamiento de σ^2

- Para corregir esto y cumplir la suposición de varianza constante, se puede realizar una transformación para **estabilizar la varianza**.
- Se aplica entonces una función $(H(Y))$ a la variable dependiente: $\sqrt{y}$ y se realiza nuevamente la estimación.
- En la figura se muestra el residuo estandarizado obtenido.
- De todas maneras, aunque se estabiliza la varianza hay algunos puntos que parecen atípicos.



Resumen

Análisis de residuos

Coeficiente de determinación

Extensión para casos no lineales

Correlación y coeficiente de determinación

- Hasta el momento hemos pensado a la variable x (variable dependiente) como no aleatoria, y la aleatoriedad de Y se la dimos sumando un ruido aditivo ($E \sim \mathbf{N}(0, \sigma^2)$).
- Podemos también considerar a ambas variables como aleatorias, en ese caso X e Y tendrán una distribución conjunta (bidimensional) $f(X, Y)$, y por consiguiente, se puede definir el coeficiente de correlación como:

$$\rho = \frac{E\{XY\} - E\{X\}E\{Y\}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad (2)$$

donde σ_{XY} es la covarianza.

- Es posible hacer inferencia estadística y test de hipótesis sobre el coeficiente de correlación, y si además, consideramos a $f(X, Y)$ una distribución Normal bivariada[2], entonces para una dada $X = x$ los estimadores de máxima verosimilitud de β_0 y β_1 dan lo mismo que los estimadores estimados con la Ec. (5) vistos en la presentación 3.1.

[2]Wiki

Correlación y coeficiente de determinación

- Como dijimos, es posible hacer inferencia estadística (estimación de Máxima Verosimilitud, IC, Test de Hipótesis, etc.) con la variable aleatoria $\rho(X, Y)$.
- El Estadístico de Prueba que se define como:

$$R = \frac{S_{XY}}{(S_{XX} SS_T)^{1/2}} \tag{3}$$

donde SS_{XY}, SS_{XX}, y SS_T se definieron en la clase de la presentación 3.1 (Ecs. (6, 7), y (9)), se denomina **coeficiente de correlación muestral**[3].

- Note que:

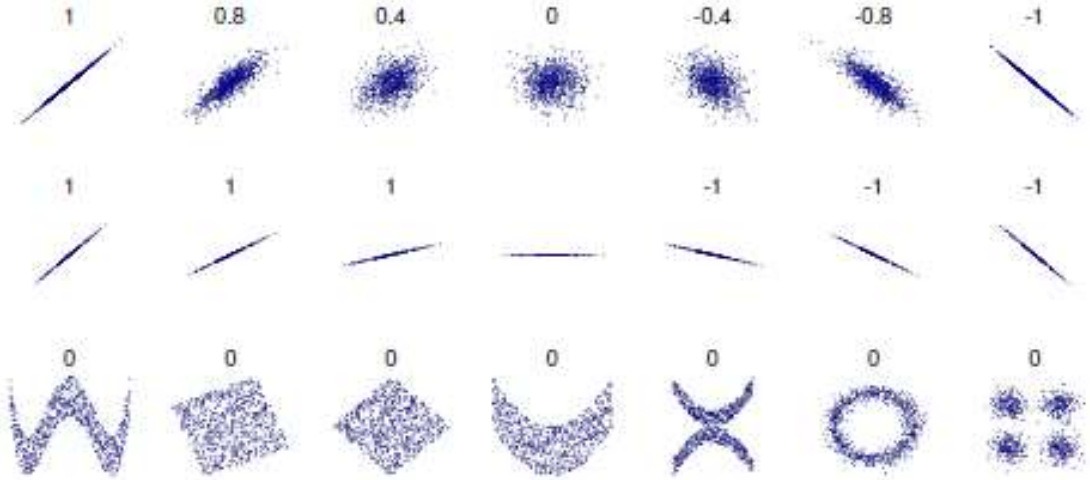
$$\hat{\beta}_1 = \left(\frac{SS_T}{S_{XX}}\right)^{1/2} R \tag{4}$$

[3]Note que hemos puesto todo con mayúsculas porque ahora las consideramos variables aleatorias.

Máxima verosimilitud (https://es.wikipedia.org/wiki/M%C3%A1xima_verosimilitud)

Correlación y coeficiente de determinación

- Según la Ec. (2), el coeficiente no depende de las unidades con que estén medidas las variables y puede tomar valores entre -1 y 1. La figura extraída de Wikipedia ejemplifica algunas salidas:



- A ρ se lo conoce también como coeficiente de correlación de Pearson. Además, a R se lo conoce como un estimador paramétrico de la correlación, porque si la distribución conjunta es Normal, entonces los patrones circulares tienden a cero, los elípticos a valores intermedios, hasta colapsar en una recta, donde toma valor uno[4].

[4]Existen otras maneras de evaluar correlación entre variables aleatorias de manera no paramétrica que más adelante comentaremos (e.g.: Spearman y Kendall).

Correlación y coeficiente de determinación

Algunos comentarios sobre la Ec. (4):

- Es importante destacar que mientras $\hat{\beta}_1$ brinda información de la pendiente, R contiene información de la relación lineal entre X e Y.
- Un valor que se suele observar es el denomiando **coeficiente de determinación**, que se define como R^2. Una forma de escribirlo es:

$$R^2 = 1 - \frac{SS_E}{SS_T} \tag{5}$$

que puede verse como una medida de cuánto de la variabilidad de los datos es explicada por el modelo de regresión lineal. De esta manera si una salida de R^2 es, por ejemplo, $r^2 = 0{,}85$ entonces diremos que el 85 % de la variabilidad es explicada por el modelo lineal.

Recordar que SS_E es la suma de los cuadrados de los errores y SS_T es la suma total de los cuadrados respecto a la media.

Correlación y coeficiente de determinación: Ejemplo

- En el script del ejemplo 19 se muestra lo mencionado hasta ahora.
- Para obtener el coeficiente de correlación:

```
# Cálculos manuales
Sxx = np.sum(xi**2) - (np.sum(xi)**2) / n
Sxy = np.sum(xi * yi) - (np.sum(xi) * np.sum(yi)) / n
SST = np.sum(yi**2) - n * (np.mean(yi)**2)

# Coeficiente de correlación muestral
r = Sxy / np.sqrt(Sxx * SST)
print(f"Salida coeficiente de correlación muestral: {r:.8f}")
```

- O se puede utilizar directamente la función de Python:

```
# También conocido como coeficiente de correlación de Pearson
pearson_r, p_value = stats.pearsonr(xi, yi)
print(f"Pearson r: {pearson_r:.8f}, p-valor: {p_value:.4e}")
```

- Para mostrar la salida de R^2:

```
# Mostrar resumen del modelo
print(modelo.summary())
# Extraer R²
R2 = modelo.rsquared
print(f"Coeficiente de determinación: {R2:.7f}")
```

Correlación y coeficiente de determinación: Ejemplo

- Si se ejecuta el ejemplo 19, se puede ver que el coeficiente de determinación es de 0,935, con lo que se puede concluir que el modelo lineal explica el 93% de la variabilidad de los datos.
- Bajo la condición de Normalidad, podemos realizar un test de hipótesis sobre $H_0 : \beta_1 = 0$ y $H_1 : \beta_1 \neq 0$, como hicimos en la presentación 3.1. Es equivalente a probar: $H_0 : \rho = 0$ y $H_1 : \rho \neq 0$, pero en este caso el Estadístico será:

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \tag{6}$$

cuya distribución es t-student con $n - 2$ grados de libertad.

```
# Estadístico t para r
T0r = r * np.sqrt(n - 2) / np.sqrt(1 - r**2)
p_valorR = 2 * (1 - t.cdf(np.abs(T0r), df=n - 2))
print(f"p-valor de R: {p_valorR:.4e}")
```

```
# Estadístico t para beta1
T0 = beta1_est / np.sqrt(varSigma_est / Sxx)
p_valor = 2 * (1 - t.cdf(np.abs(T0), df=n - 2))
print(f"p-valor de beta1: {p_valor:.4e}")
```

Resumen

Análisis de residuos

Coeficiente de determinación

Extensión para casos no lineales

Transformaciones

- En algunas ocasiones la relación entre las variables x (volvamos a x no aleatoria) e Y no es lineal. En ese caso, se pueden aplicar transformaciones, tal como lo hicimos para cumplir la condición de varianza constante en el ejemplo 20.

- Por ejemplo, tomemos el caso del ejemplo 21, donde se observan los datos de la correlación entre la tasa de energía por tiempo necesaria para que un primate mantenga sus funciones básicas[5] (RMR, de sus siglas en inglés *Resting Metabolic Rate*) y su peso.

- Para muchos animales (incluidos los humanos[6]) se cumple una ley de potencia del tipo:

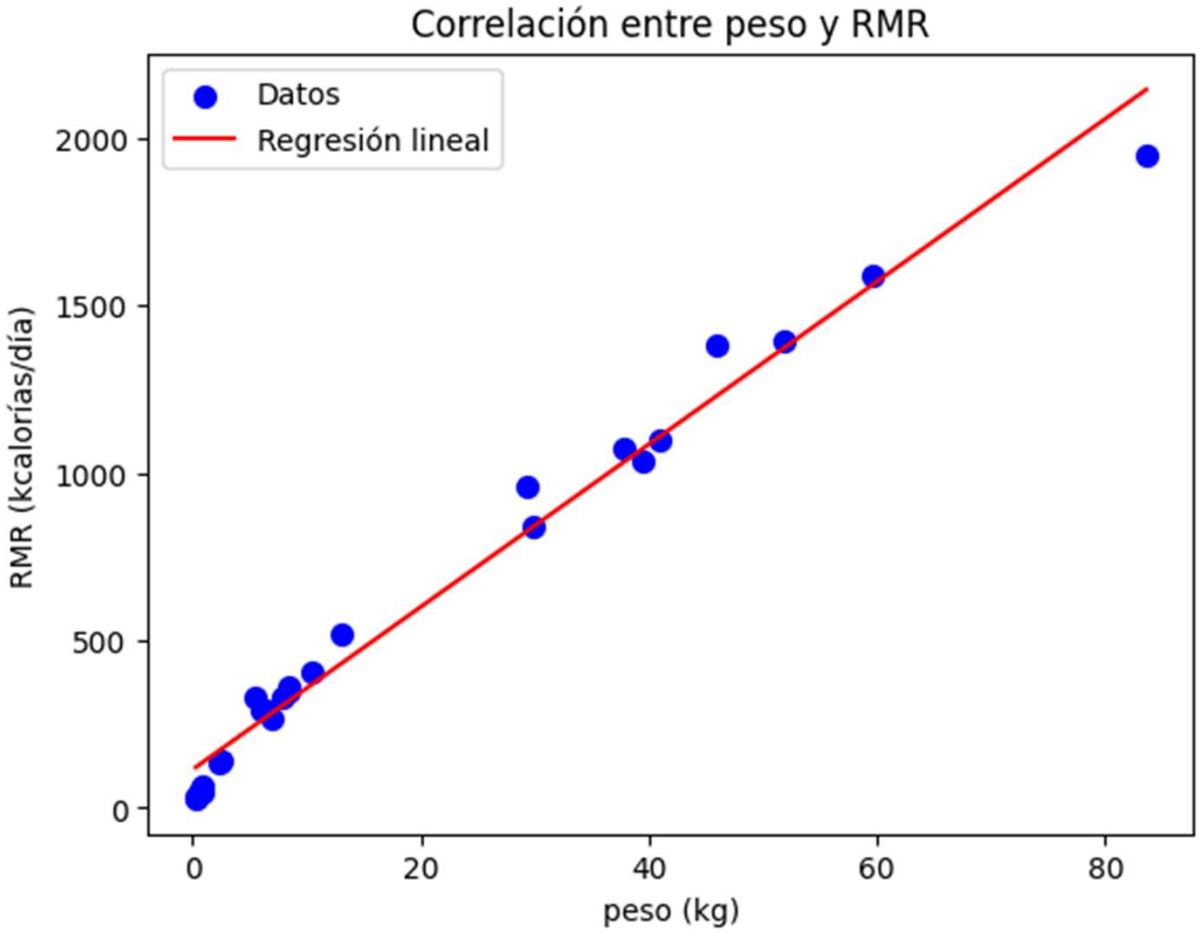
$$Y = Ax^{\beta} \tag{7}$$

donde la RMR se representa con la letra Y y el peso con la letra x.
β y A son constantes.

[5]Esto es, respiración, circulación, crecimiento celular y funciones de digestión.
[6]También se cumple en niños y adolescentes (https://www.fao.org/3/m2845e/m2845e00.htm)

Transformaciones

- Veamos entonces, los datos del ejemplo 21.
- Se muestra también una regresión lineal en rojo.
- Se ve que no es mala la aproximación lineal, pero podría ser mejor.
- Probemos entonces la ley de potencia de la Ec. (7).



Transformaciones

- Lo que se puede hacer en estos casos es aplicar logaritmo en ambos lados de la Ec. (7), pero para estudiar el problema de manera probabilística es conveniente agregar el error E:

$$ln(Y) = ln\left(Ax^{\beta}E\right)$$

nos queda la siguiente expresión:

$$ln(Y) = ln(A) + \beta ln(x) + ln(E) \tag{8}$$

- De esta forma si llamamos $W = ln(Y)$ y $z = ln(x)$ entonces tenemos una función lineal y podemos aplicar lo aprendido en esta parte de la Unidad III[7].

[7] ¿Se anima a continuar este ejercicio?

Conclusiones de la Parte II

- En las tres partes que conformaron esta parte hemos trabajado con modelos lineales en una dimensión.
- Estudiamos intervalos de confianza y tests de hipótesis sobre la ordenada al origen (β_0) y la pendiente (β_1) de una respuesta lineal con ruido aditivo Normal con media cero y varianza constante.
- Para poder evaluar lo anterior también le pedimos al ruido que sea independiente.
- Estudiamos los residuos y cómo evaluar si se cumplen las condiciones impuestas al modelo.
- Finalmente, definimos el coeficiente de correlación (ρ), coeficiente de correlación muestral (R) y un test de hipótesis sobre éste, para evaluar si hay o no correlación. Además, comentamos la interpretación del coeficiente de determinación (R^2).

Conclusiones de la Parte II

- Es posible realizar un test de hipótesis para contrastar el valor del coeficiente de correlación, es decir, $H_0 : \rho = \rho_0$ con $H_1 : \rho \neq \rho_0$. Aunque se debe pedir un tamaño de muestra grande (más que 30, para los interesados ver Referencia [1]).
- Vimos transformaciones útiles al momento de estabilizar la varianza (ejemplo de transformación $H(Y) = \sqrt{Y}$) o para cuando el fenómeno que estamos estudiando no es lineal (ejemplo de transformación de ley de potencia $H(Y) = ln(Y)$). Esta idea se puede utilizar aplicando diferentes funciones que pueden tener significado físico, para más datos ver referencia [2].

Bibliografía

1- D. C. Montgomery and G. C. Runger, “Applied Statistics and Probability for Engineers”, Third Edition, John Wiley & Sons, Inc.

2- D. C. Montgomery, E. A. Peck, G. G. Vining, “Introduction to Linear Regression Analysis”, John Wiley & Sons, Inc.

3- Peter Dalgaard, “Introductory Statistics with R”, Second Edition, Springer.