KIBERNUM
IT Academy

Módulo 3 – Análisis Exploratorio y Programación Estadística

Regresiones Lineales Simples

Ciencia de Datos

Regresiones lineales

¿Existe una relación entre la altura y el peso?

En caso de existir, ¿será una relación lineal?

¿Se podrá predecir el peso de una persona a partir de su altura?

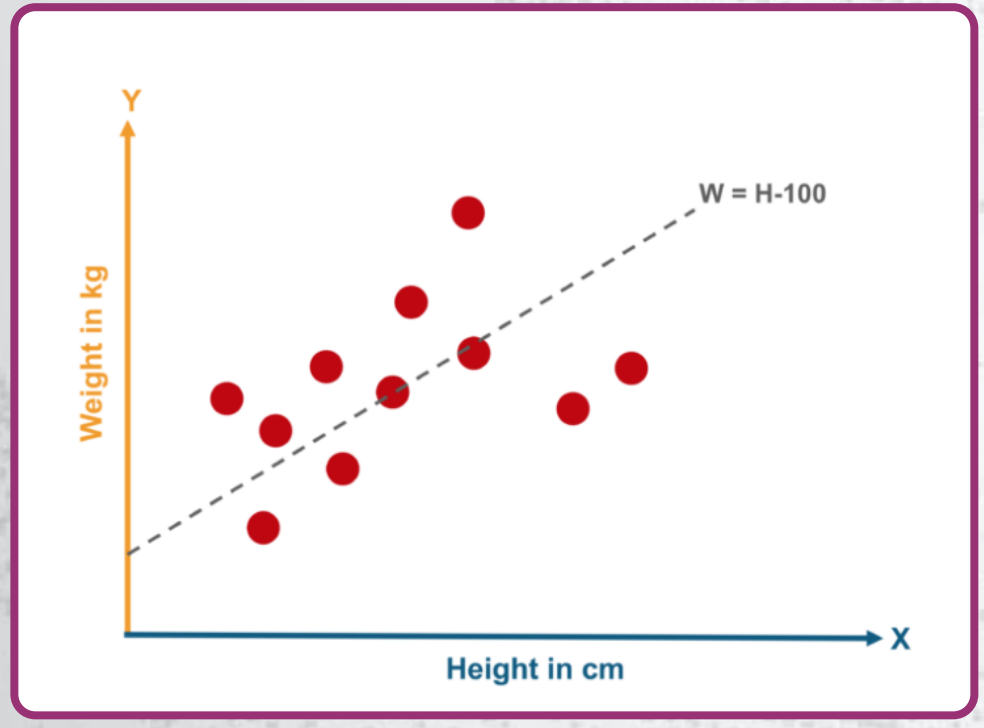
Dataset Estatura Peso

- Para responder las preguntas anteriores, tomaremos el dataset Height-Weight tomado de SOCR (Statistic Online Computational Resource), que contiene mediciones de 25 mil individuos.

Index	Height(Inches)	Weight(Pounds)
1	65.78331	112.9925
2	71.51521	136.4873
3	69.39874	153.0269
4	68.21660	142.3354
5	67.78781	144.2971
...	...	...
24996	69.50215	118.0312
24997	64.54826	120.1932
24998	64.69855	118.2655
24999	67.52918	132.2682
25000	68.87761	124.8742

http://socr.ucla.edu/docs/resources/SOCR_Data/SOCR_Data_Dinov_020108_HeightsWeights.html

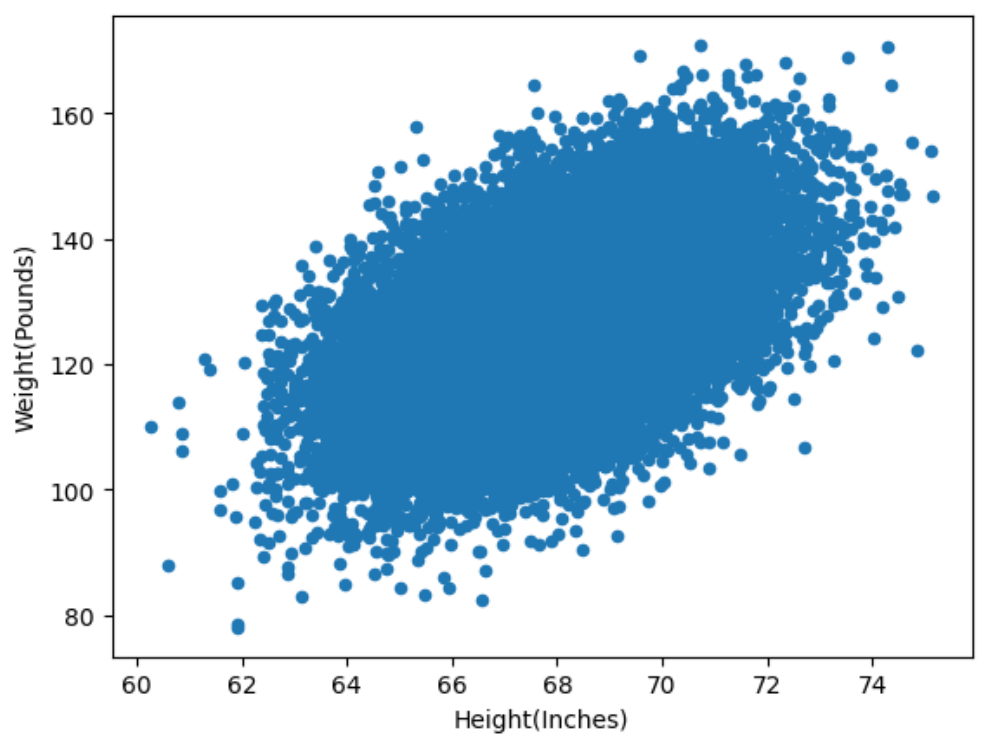
Dataset Estatura Peso



- Nuestra primera suposición, será que existe una relación lineal entre la altura de una persona y su peso. A simple vista, se podría deducir que el peso de una persona es directamente proporcional a su estatura, pero dejemos que la estadística nos ayude con esto.
- Nuestra segunda suposición, será que el peso de una persona depende de su altura. Es decir, hay una relación causal entre peso y estatura.
- Una Regresión Lineal es uno de los modelos estadísticos más simple que se pueden elaborar. Es usada para mostrar la relación lineal entre una variable dependiente y una o más variables independientes.

Dataset Estatura Peso

```
df.plot(kind='scatter', x='Height(Inches)', y='Weight(Pounds)')
```



En el diagrama de dispersión, se observa una correlación positiva entre estatura y peso. Es decir, a medida que aumenta la altura de un individuo, su peso también aumenta. También se aprecia que la relación entre ambas variables es lineal.

La cuantía de la correlación es de aproximadamente 0.5, es decir, una correlación moderada.

```
df.corr()
```

	Height(Inches)	Weight(Pounds)
Height(Inches)	1.000000	0.502859
Weight(Pounds)	0.502859	1.000000

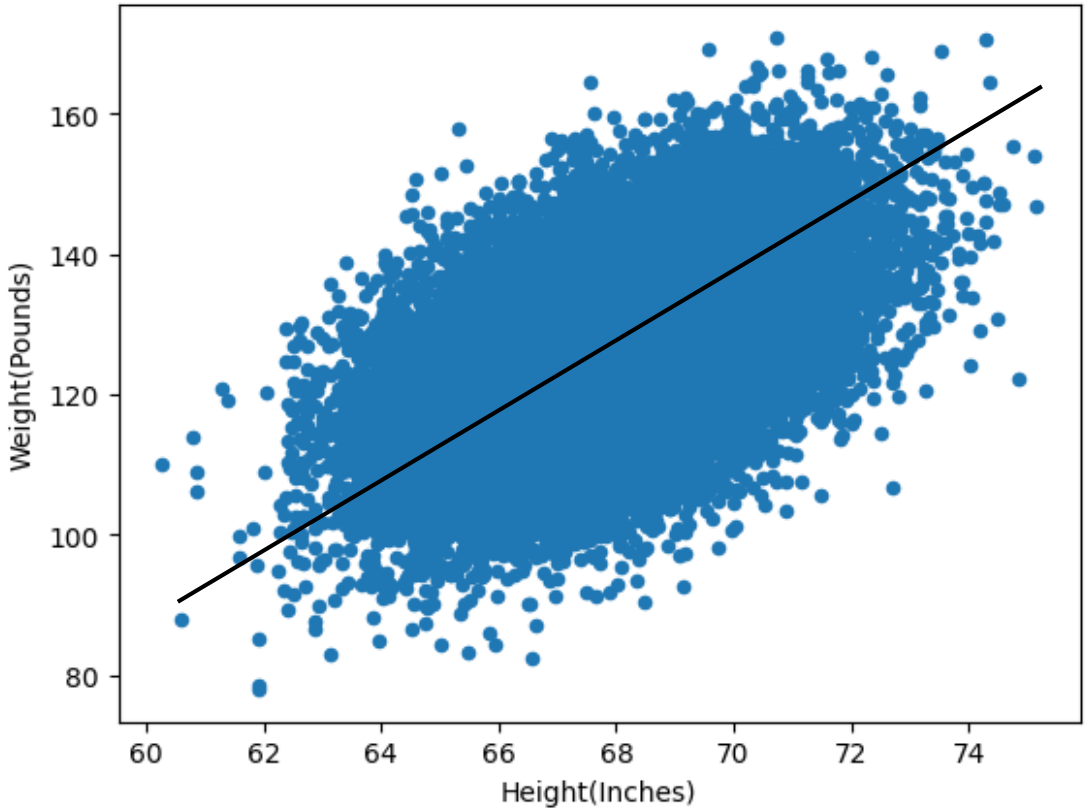
KIBERNUM

¿Qué es un Modelo?

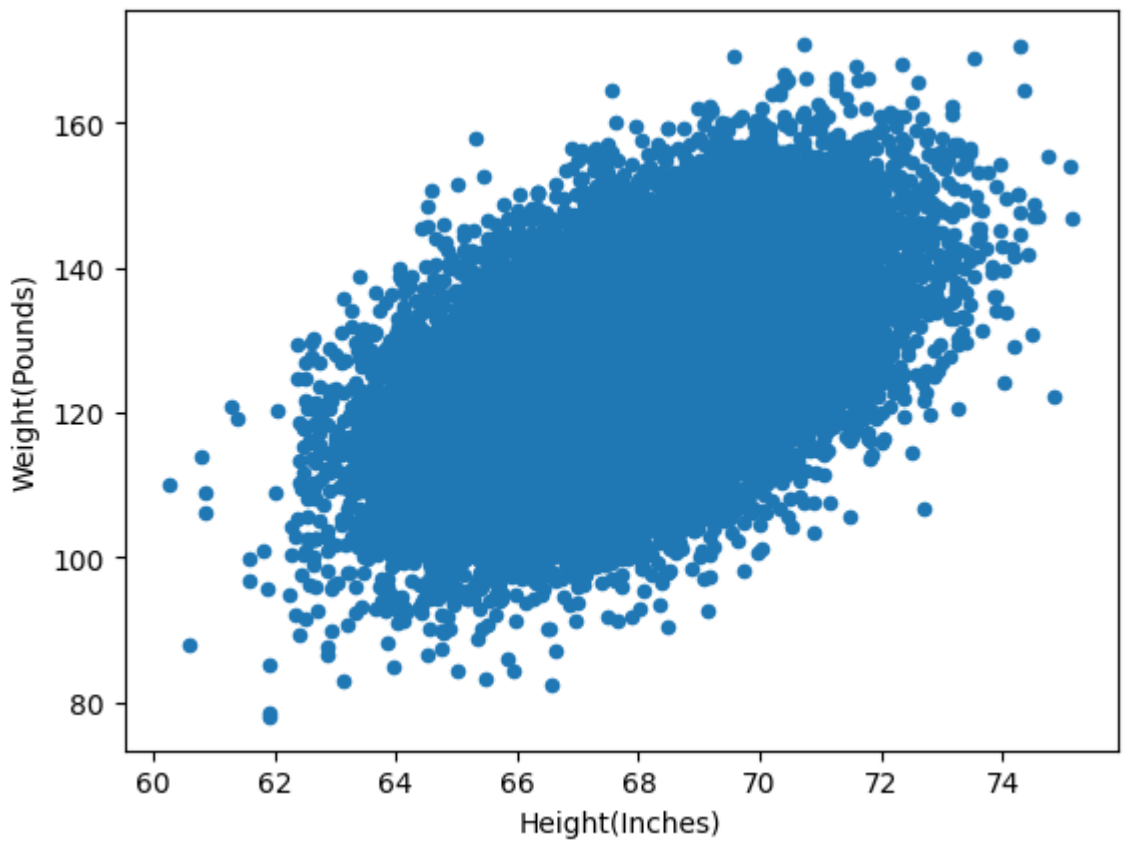
Un modelo, es una **simplificación de la realidad**, que permite **describir un fenómeno de forma general**.

Un modelo lineal, es un tipo de modelo matemático que se usa para predecir la relación entre una variable dependiente y una o más variables independiente. Un modelo lineal, simplifica la realidad del fenómeno y lo lleva a una ecuación lineal, lo que significa que la relación entre las variables es lineal o proporcional.

Hacer esta simplificación, no siempre es la alternativa más adecuada para representar la realidad, sin embargo, hay muchos fenómenos que tienen un comportamiento lineal y que este tipo de modelo puede capturar su esencia.



¿Qué es un Modelo Lineal?



- En este ejemplo, si se calcula la pendiente de la recta, se obtiene el coeficiente o la razón que hay entre altura y peso.
- $\frac{\Delta y}{\Delta x} = \frac{140 - 100}{70{,}4 - 62{,}4} = \frac{40}{8} = 5\ {}^{lb}/_{in}$
- Como se puede observar, el modelo lineal dice que por cada pulgada que se incremente la altura, el peso se incrementaría en 5 libras.

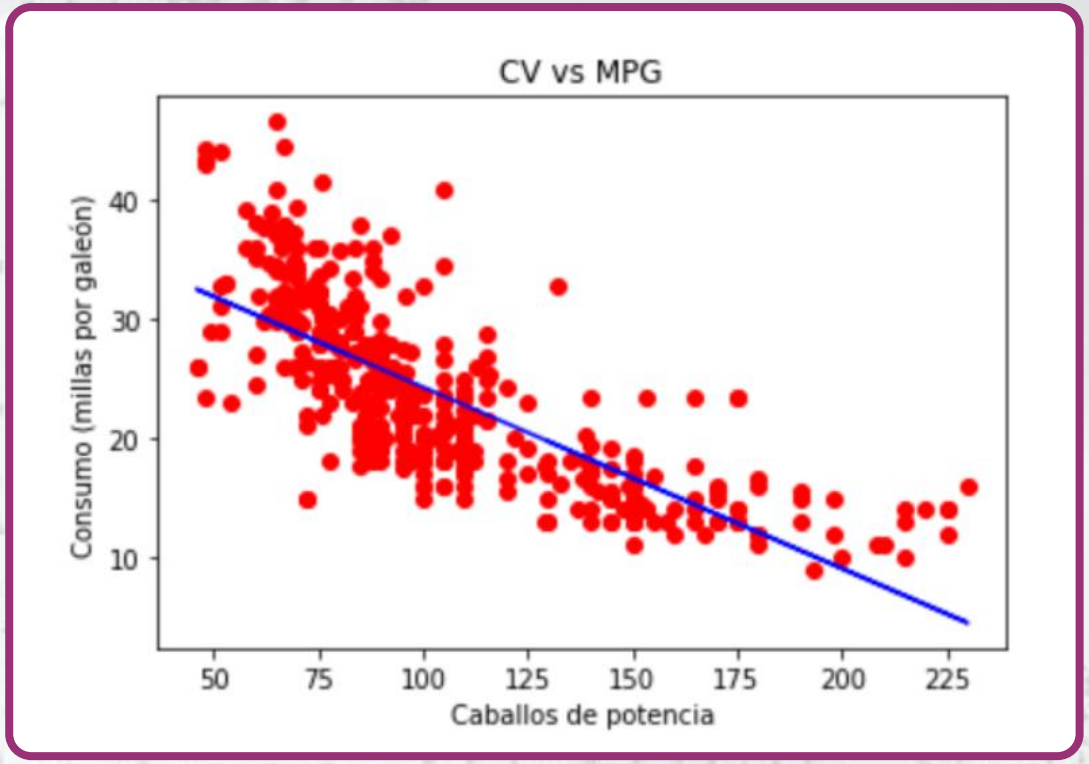
¿Qué es un Modelo de Regresión?

- Un modelo lineal, es un tipo de modelo regresivo, ya que la regresión lineal es un método de modelado que se utiliza para establecer una relación lineal entre una variable dependiente y una o más variables independientes. Por lo tanto, es posible decir que los modelos lineales son un subconjunto de los modelos regresivos.

- Sin embargo, hay otros tipos de modelos regresivos que no son lineales, como los modelos de regresión no lineal, los cuales permiten establecer una relación más compleja entre la variable dependiente y las variables independientes.

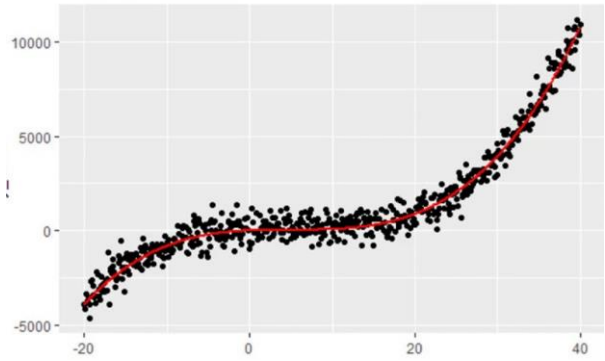
¿Qué es un Modelo de Regresión?

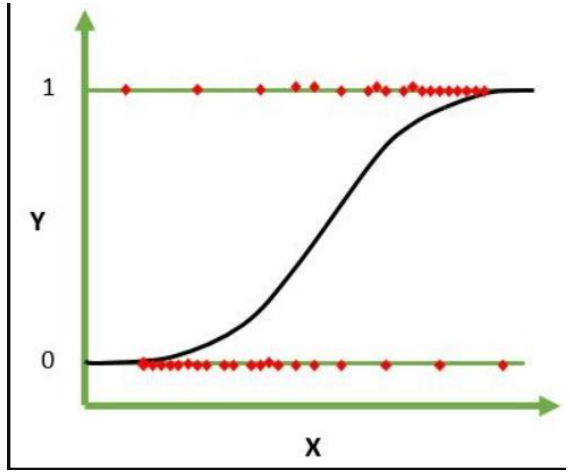
- En el siguiente ejemplo, se observa que la relación entre los caballos de potencia de un vehículo y el consumo no disminuye de forma lineal. En este caso, utilizar un modelamiento lineal no sería lo más indicado para capturar la esencia de este fenómeno, dado que un modelo lineal estimaría que un vehículo con más de 250 caballos de potencia tendría un consumo negativo, lo cual es imposible.



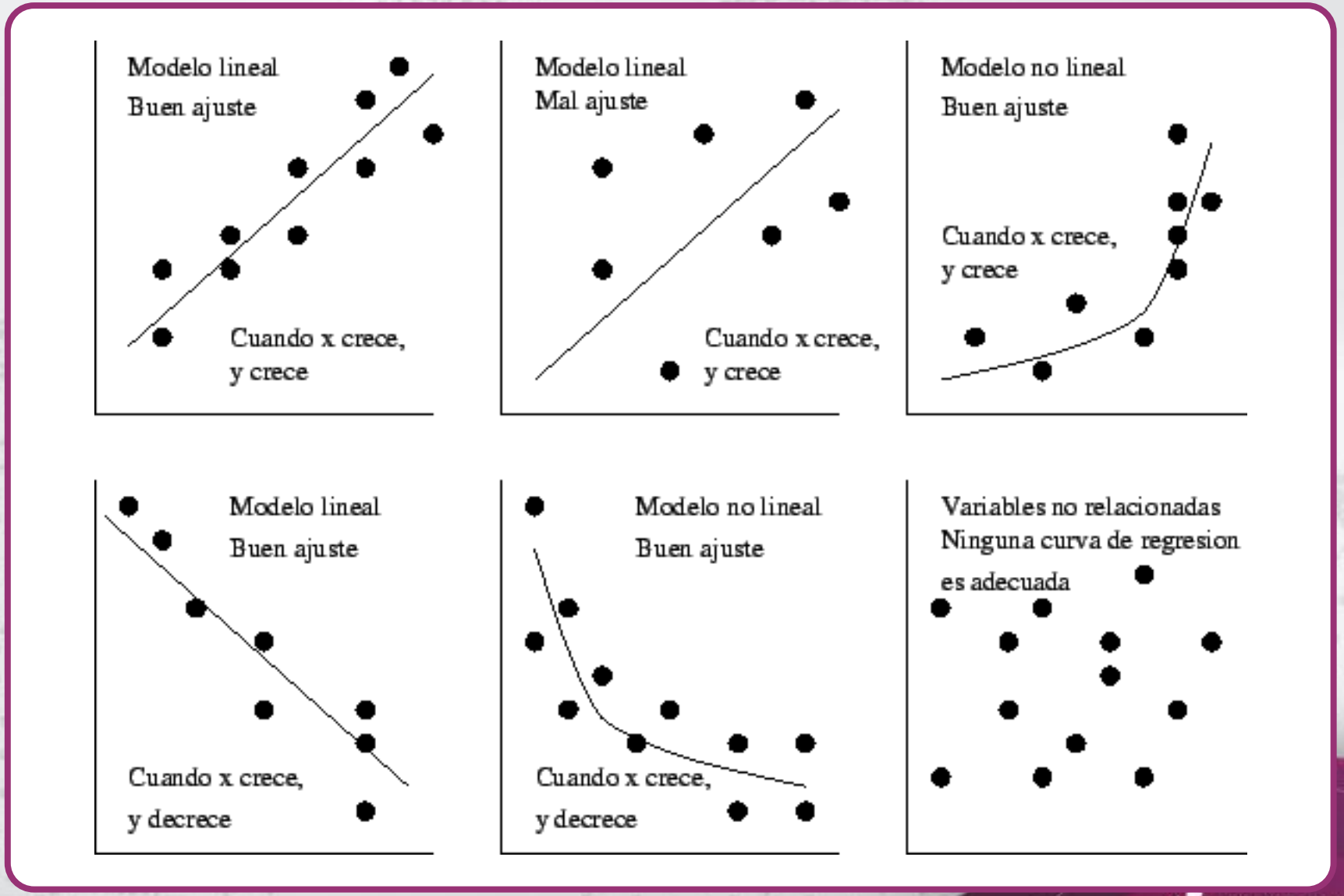
¿Qué otros tipos de regresión existen?

Otros ejemplos de modelos regresivos que no son lineales.

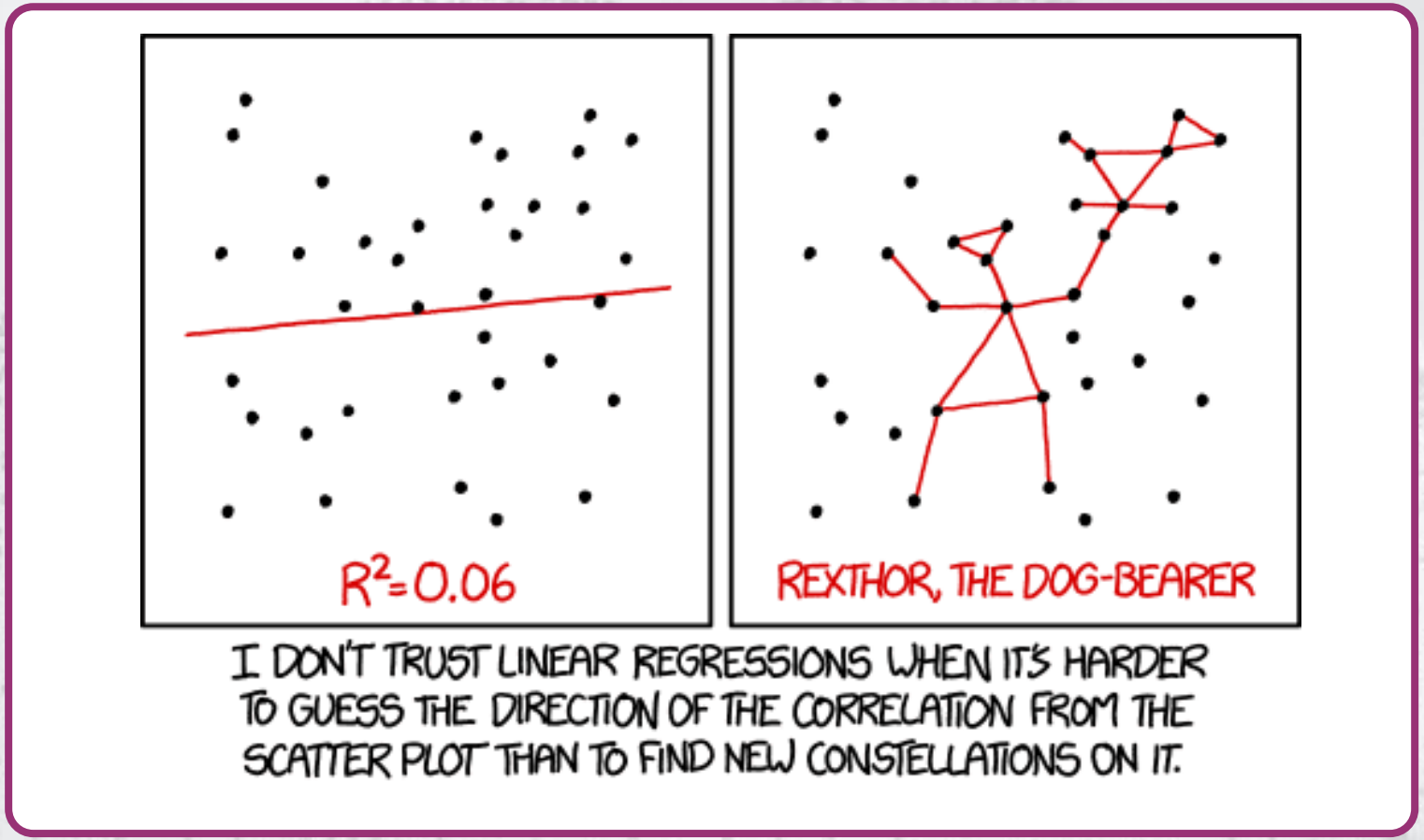




¿Cómo validar el modelo de regresión?



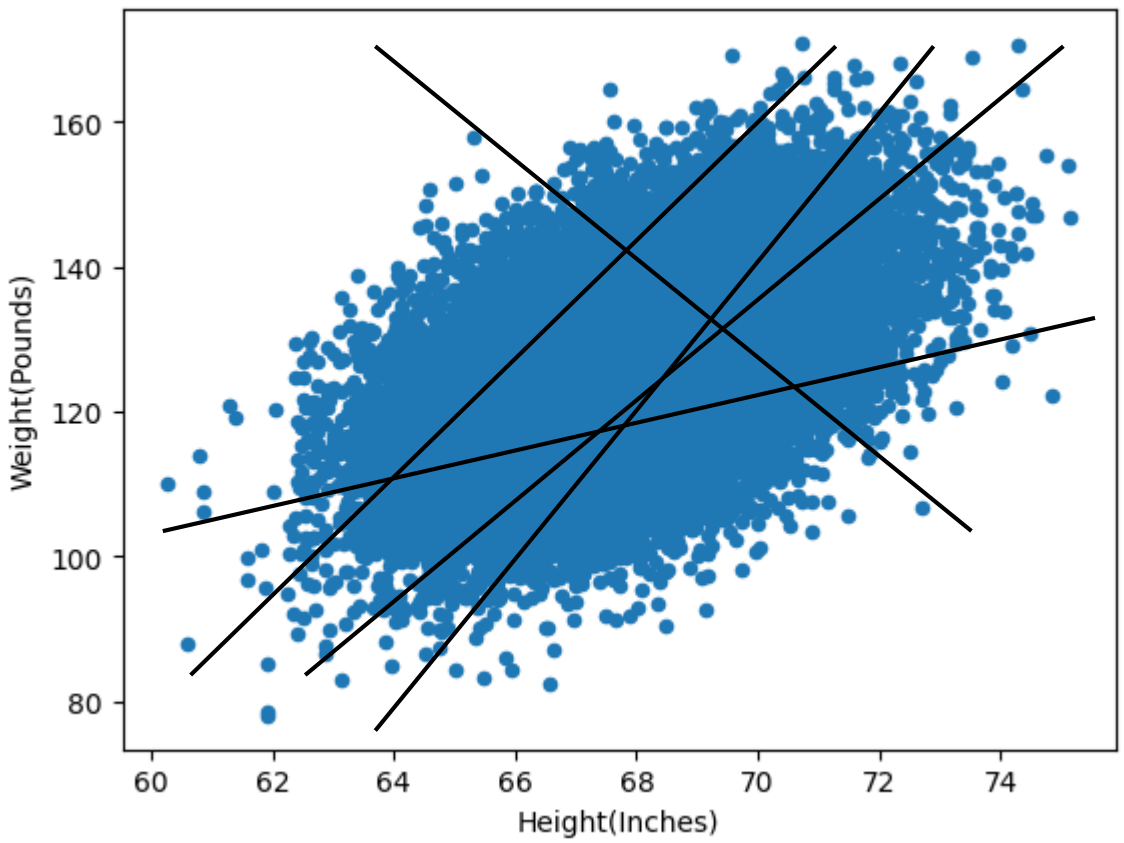
¿Cómo validar el modelo de regresión?



"No confío en las regresiones lineales cuando es más difícil adivinar la dirección de la correlación a partir del gráfico de dispersión que encontrar nuevas constelaciones en él."

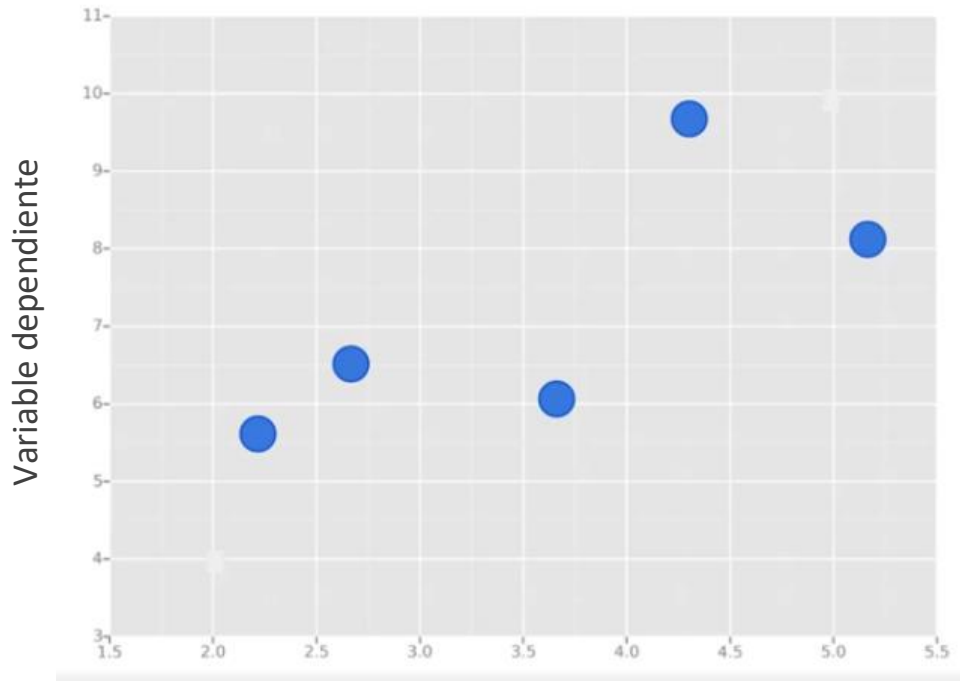
¿Qué línea utilizamos?

- Como hemos mencionado, un modelo regresivo lineal simplifica el fenómeno y lo representa con una línea recta, pero ¿qué línea utilizamos?
- No todas las líneas que podamos trazar nos van a servir para este propósito. Sin embargo, la línea que trazaremos será la mejor línea posible. Pero ¿cuál es la mejor línea posible?
- La mejor línea posible, es la que tenga el menor error.

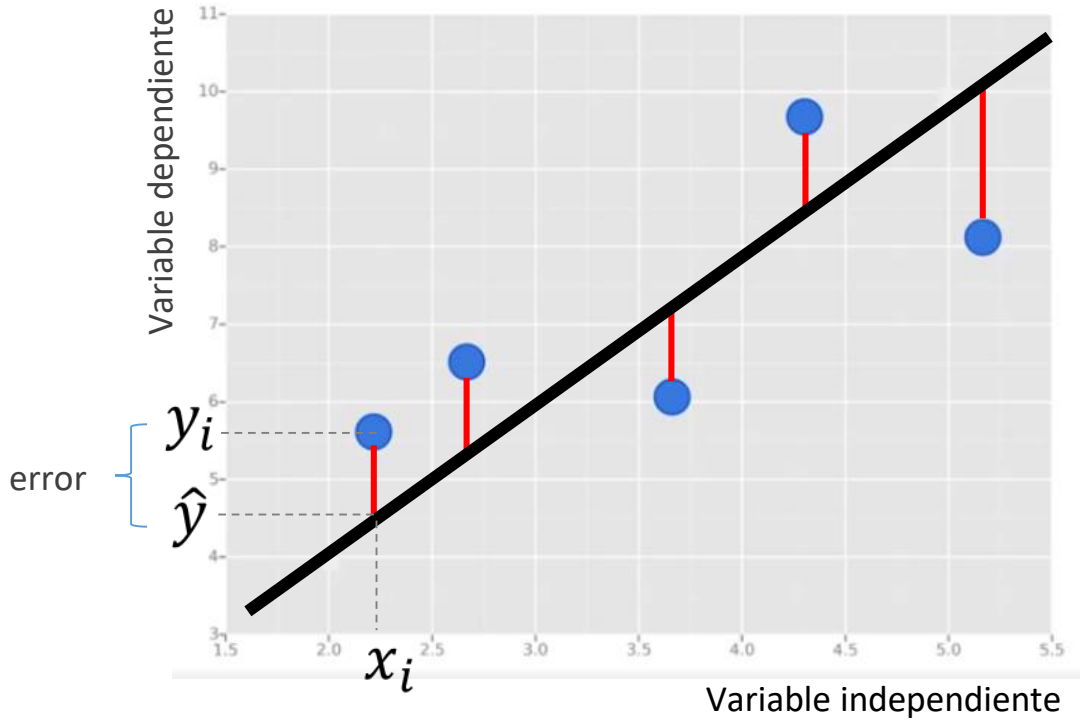


Método de los mínimos cuadrados

Mínimos Cuadrados es un método estadístico que se utiliza para determinar la línea de mejor ajuste o la línea de regresión, minimizando la suma de cuadrados creada por una función matemática.



Método de los mínimos cuadrados



- Es decir, tomaremos la línea recta que minimice la sumatoria total de los errores al cuadrado.

$$min \sum_{i=0}^{m} (y_i - \hat{y}_i)^2$$

- Por lo tanto, la tarea consiste en determinar los valores de b0 y de b1 que minimicen el error.

$$y = b_0 + b_1 \, x_1 + err$$

Determinando los coeficientes de la regresión lineal

El objetivo es encontrar los valores estimados para α y β que pueda proveer el mejor ajuste para los datos. En otras palabras, los valores que minimizan el siguiente problema de minimización:

$$\text{Find } \min_{\alpha,\beta} Q(\alpha,\beta), \quad \text{for } Q(\alpha,\beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

KIBERNUM

Determinando los coeficientes de la regresión lineal

Las siguientes ecuaciones permiten determinar los valores estimados de α y β:

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\,\bar{x},$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{s_{x,y}}{s_x^2}$$

$$= r_{xy}\frac{s_y}{s_x}.$$

KIBERNUM

Hipótesis de Trabajo

- En el contexto de la investigación científica, la hipótesis es una idea o propuesta que se formula como respuesta tentativa a una pregunta de investigación. Esta idea puede ser refutada o corroborada mediante la recolección de datos y la realización de experimentos y análisis estadísticos.
- Las hipótesis son importantes en la investigación científica porque permiten a los investigadores diseñar estudios y experimentos específicos para probar su validez. Si la hipótesis se confirma, se considera que se ha avanzado en el conocimiento del fenómeno estudiado. Si la hipótesis se refuta, se deben formular nuevas ideas o preguntas para continuar la investigación.
- En nuestro caso, formularemos la siguiente hipótesis de trabajo:

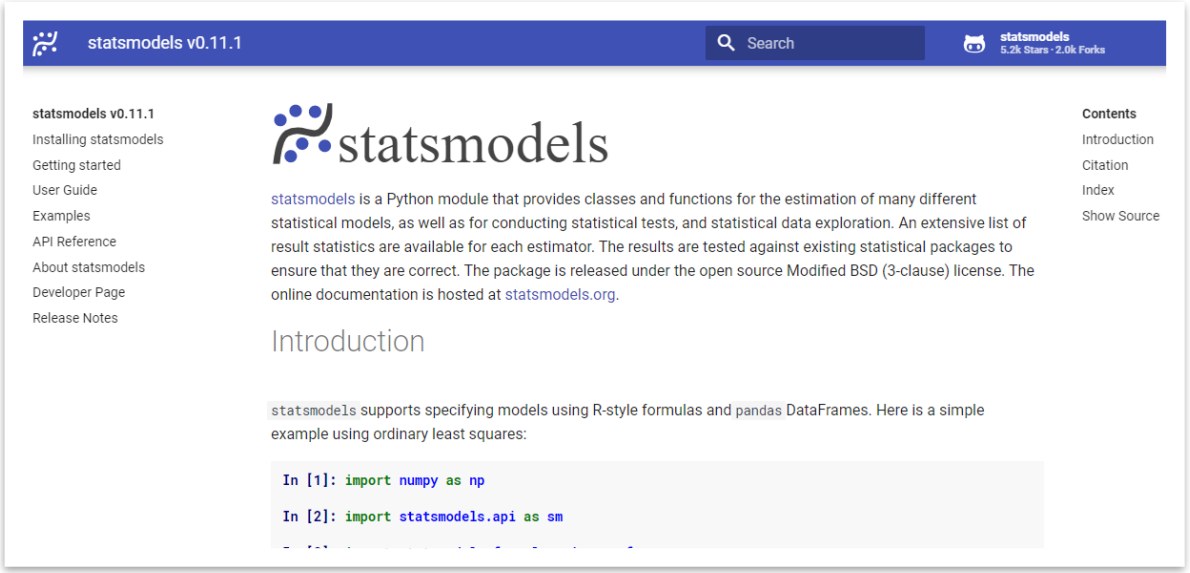
"El peso de un individuo depende de su altura, y tiene una relación lineal"

$$peso = b_0 + b_1\ altura$$

Librería StatsModels

Statsmodels es una librería de Python que se utiliza para realizar análisis estadísticos y econométricos. Proporciona una amplia gama de herramientas para el modelado estadístico, la estimación de parámetros y la realización de pruebas de hipótesis.

Statsmodels es especialmente útil para el análisis de datos de series temporales, análisis de regresión, análisis de datos de panel y análisis multivariado. La librería ofrece una amplia variedad de modelos estadísticos, incluyendo modelos lineales, modelos de series de tiempo, modelos de regresión logística, modelos de probabilidad, modelos de análisis de varianza, entre otros.



https://www.statsmodels.org/

KIBERNUM

Determinación de coeficientes

Importamos la librería **statsmodels**.

```
[7]  import statsmodels.formula.api as smf
```

Formulamos el modelo y lo ajustamos.

```
[13] lm = smf.ols(data=df, formula="Weight ~ Height").fit()
```

Variable dependiente (outcome)

Variable independiente (predictor)

Método que realiza el ajuste del modelo

KIBERNUM

Definición de la Fórmula

Para definir la fórmula que se modelará en un modelo regresivo, se puede utilizar el estilo de definición del lenguaje R. A continuación, algunos ejemplos:

Fórmula que incorpora 3 predictores (Literacy, Wealth y Region)

```
mod = smf.ols(formula='Lottery ~ Literacy + Wealth + Region', data=df)
```

Fórmula que incorpora la variable Region como variable categórica

```
res = smf.ols(formula='Lottery ~ Literacy + Wealth + C(Region)', data=df).fit()
```

Fórmula que remueve el intercepto del modelo

```
res = smf.ols(formula='Lottery ~ Literacy + Wealth + C(Region) -1 ', data=df).fit()
```

Fórmula que aplica una función a un predictor (función vectorizada)

```
res = smf.ols(formula='Lottery ~ np.log(Literacy)', data=df).fit()
```

Interpretando el modelo

- Una vez entrenado el algoritmo, se puede obtener del modelo los coeficientes b0 y b1.

$$y = b_0 + b_1 X$$

```
[14] lm.params

     Intercept   -82.575743
     Height        3.083476
     dtype: float64

[15] lm.params[0]

     -82.57574306454089

[16] lm.params[1]

     3.0834764454029653
```

Interpretando el modelo

También podemos desplegar el sumario del modelo ajustado.

```
lm.summary()
```

OLS Regression Results

Dep. Variable:	Weight	**R-squared:**	0.253
Model:	OLS	**Adj. R-squared:**	0.253
Method:	Least Squares	**F-statistic:**	8461.
Date:	Fri, 31 Mar 2023	**Prob (F-statistic):**	0.00
Time:	20:58:07	**Log-Likelihood:**	-93235.
No. Observations:	25000	**AIC:**	1.865e+05
Df Residuals:	24998	**BIC:**	1.865e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-82.5757	2.280	-36.214	0.000	-87.045	-78.106
Height	3.0835	0.034	91.981	0.000	3.018	3.149

Omnibus:	1.022	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.600	**Jarque-Bera (JB):**	1.027
Skew:	-0.016	**Prob(JB):**	0.598
Kurtosis:	2.996	**Cond. No.**	2.43e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.43e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

KIBERNUM

Interpretando el modelo

Con lo cual, nuestro modelo final sería el siguiente:

$$Peso = -82.5757 + 3.085 * \text{Altura}$$

Realizando predicciones con el modelo

Ahora, realizaremos predicciones con algunos valores. Para esto, utilizaremos el método **predict()**.

```
[20] x = pd.DataFrame([60, 65, 70, 75, 80], columns=['Height'])
     x
```

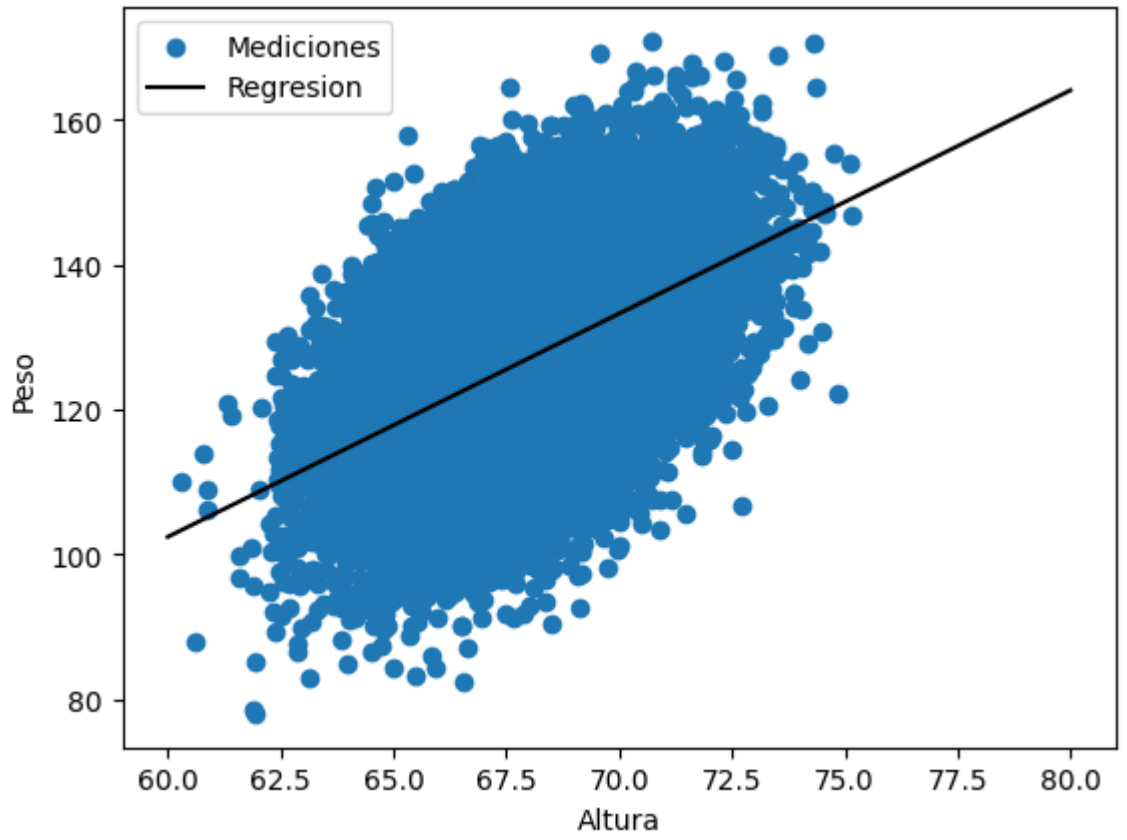
	Height
0	60
1	65
2	70
3	75
4	80

```
[21] lm.predict(x)
```

```
0    102.432844
1    117.850226
2    133.267608
3    148.684990
4    164.102373
dtype: float64
```

Visualizando el resultado

```
[24] plt.scatter(df['Height'], df['Weight'], label='Mediciones')
     plt.plot(x,lm.predict(x), c='black', label='Regresion')
     plt.xlabel('Altura')
     plt.ylabel('Peso')
     plt.legend(loc=0)
```

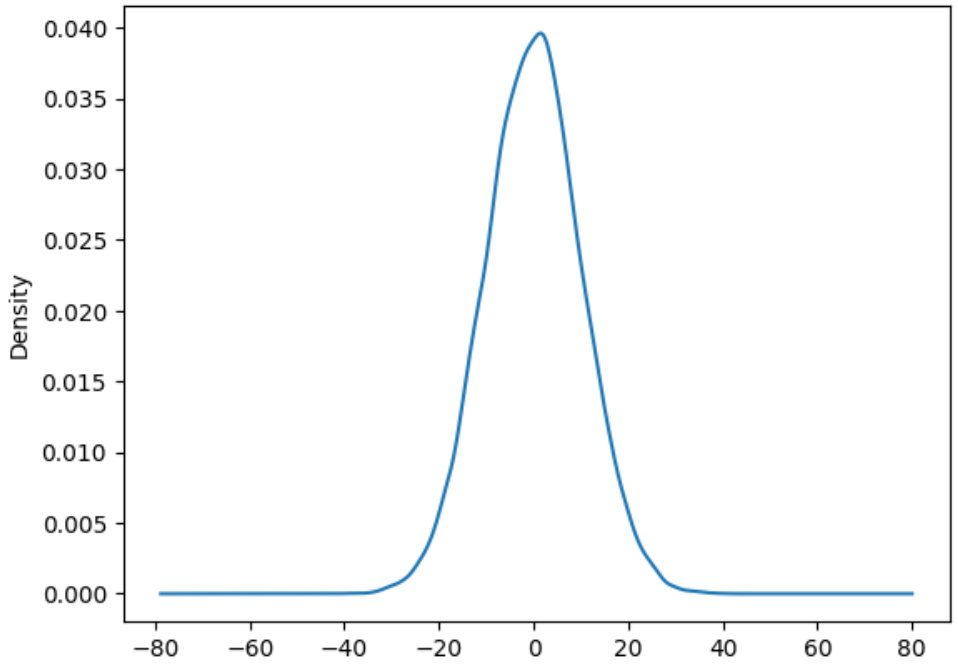


Análisis de Residuales

En general, se espera que los residuos se distribuyan normalmente con una media de cero y una varianza constante en todas las combinaciones de valores de las variables explicativas. Los patrones en los residuos que violan estas suposiciones pueden indicar que el modelo no es adecuado para los datos.

```
[25] y_pred = lm.predict(df['Height'])
     y_true = df['Weight']
```

```
[27] (y_pred - y_true).plot(kind='kde')
```



Métricas de evaluación de regresiones

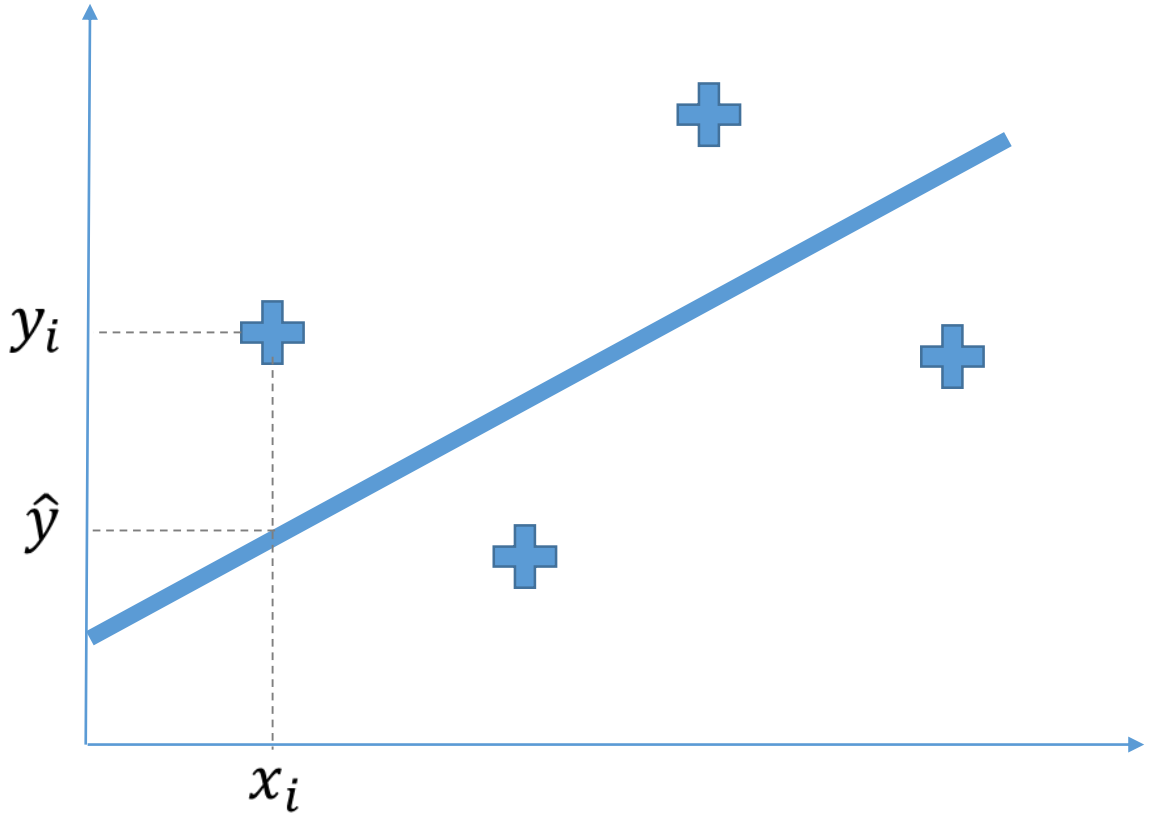
En una regresión lineal, las métricas de evaluación están enfocadas a cuantificar el error del modelo respecto a los valores reales en el set de datos donde realizamos la validación. Existen varias métricas de error, pero las más utilizadas son las siguientes:

- Mean Absolute Error (MAE).
- Mean Squared Error (MSE).
- Root Mean Squared Error (RMSE).

(MAE) Mean Absolute Error

Es la métrica más sencilla de todas para medir el error de un modelo regresivo, corresponde a la suma de las diferencias absolutas entre el valor predicho y el valor real dividido por la cantidad de mediciones.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$



Métricas de evaluación de regresiones

Para utilizar las funciones de métricas de evaluación de la regresión, podemos importar el módulo **eval_measures** de **statsmodels**.

```
[28] import statsmodels.tools.eval_measures as metrics
```

```
[29] # Mean Absolute Error
     metrics.meanabs(y_true,y_pred)
```

```
     8.037502348939364
```

```
[30] # Mean Squared Error
     metrics.mse(y_true,y_pred)
```

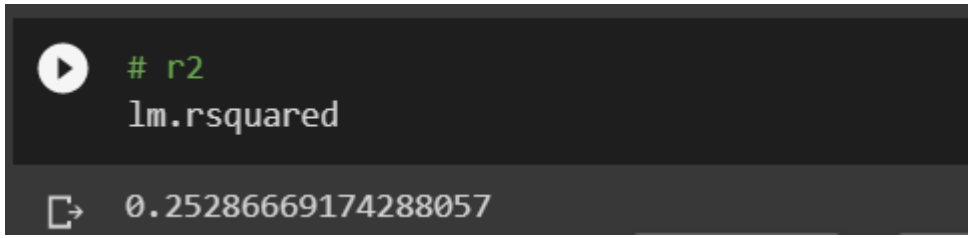
```
     101.58853248632849
```

```
[31] # Root Medium Squared Error
     metrics.rmse(y_true,y_pred)
```

```
     10.079113675632819
```

Coeficiente de determinación (R-cuadrado)

Mide cuán bien se ajusta nuestro modelo a los datos, es decir, mide la proporción de varianza de la variable objetivo que el modelo es capaz de explicar. Este coeficiente oscila entre 0 y 1, correspondiendo a 1 el ajuste perfecto del modelo y a 0 cuando la variable de salida no reacciona a nuestra variable predictora.



En este ejemplo, el modelo sólo es capaz de explicar el 25,3% de la varianza de la variable de salida.

Dudas y consultas

KIBERNUM

Fin presentación

KIBERNUM