REGRESIÓN LINEAL

Kenny Cárdenas Parra

¿QUÉ ES REGRESIÓN?

Predecir un valor numérico (variable dependiente) basado en una serie de datos de entrada o variables independientes.

- Sentido estadístico: Predecir el valor esperado.
- Sentido causal: Predecir el valor numérico esperado.

Esto es la diferencia entre predicción y clasificación.

PREDICCIÓN VS CLASIFICACIÓN

- ¿Cuántas unidades vamos a vender?
- ¿Qué marca será el carro?
- ¿El cliente va a comprar el producto?
- ¿Cuánto serán las ventas totales?

PREDICCIÓN VS CLASIFICACIÓN

- ¿Cuántas unidades vamos a vender?
- ¿Qué marca será el carro?
- ¿El cliente va a comprar el producto?
- ¿Cuánto serán las ventas totales?

REGRESIÓN LINEAL

Es el algoritmo más básico de regresión lineal, donde el resultado esperado se supone que es la suma ponderada de todas las entradas y que el cambio que sufre la variable dependiente "Y" es proporcional al cambio que sufran las variables independientes "X".

Regresión Simple $$Y = \beta_0 + \beta_1 \cdot x_1$$

Regresión Múltiple $$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 \ldots + \beta_i \cdot x_i$$

VENTAJAS Y DESVENTAJAS

Ventajas

- Fácil de aplicar y de ajustar.
- Concisos (capacidad de almacenamiento)
- Menos propensos a sobreajustarse, esto quiere decir rendimiento sobre nuevos datos.
- Interpretables

VENTAJAS Y DESVENTAJAS

Desventajas

- No puede expresar relaciones complejas, solo relaciones aditivas y lineales.
- Colinealidad.
- Alta colinealidad
  - Coeficientes con valores no deseados
  - Modelos inestables.

VALORES PREDICHOS

Son los valores calculados a partir de la combinación lineal encontrada previamente.

$$\hat{Y} = X \cdot \beta$$

RESIDUALES

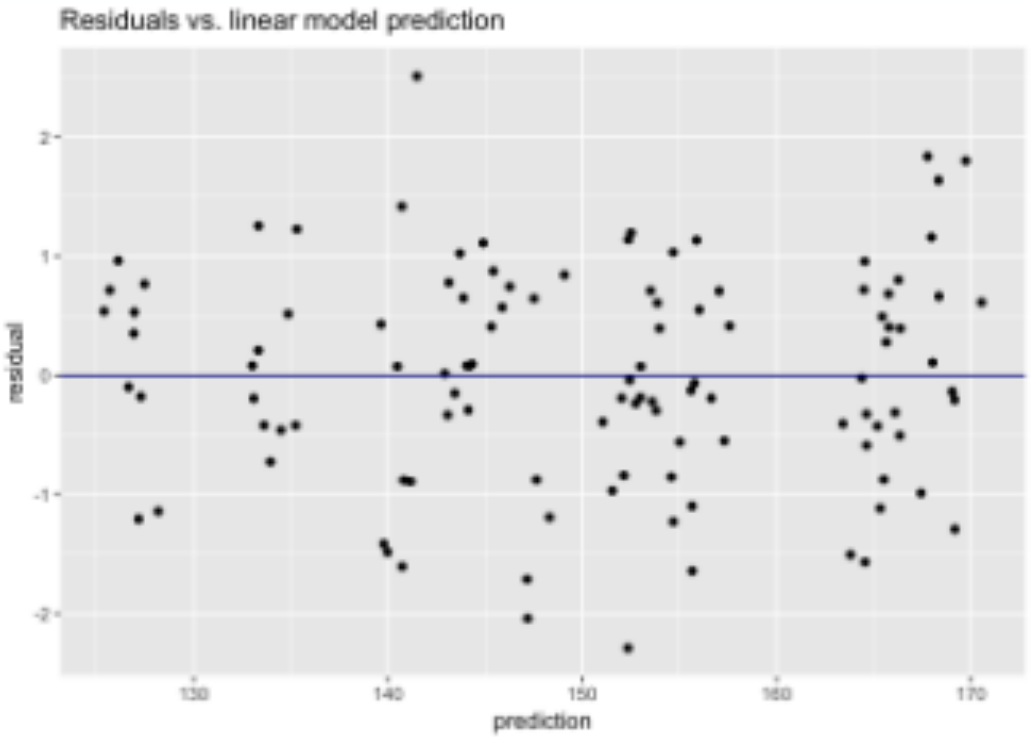
Es la diferencia entre el valor predicho y el valor observado.

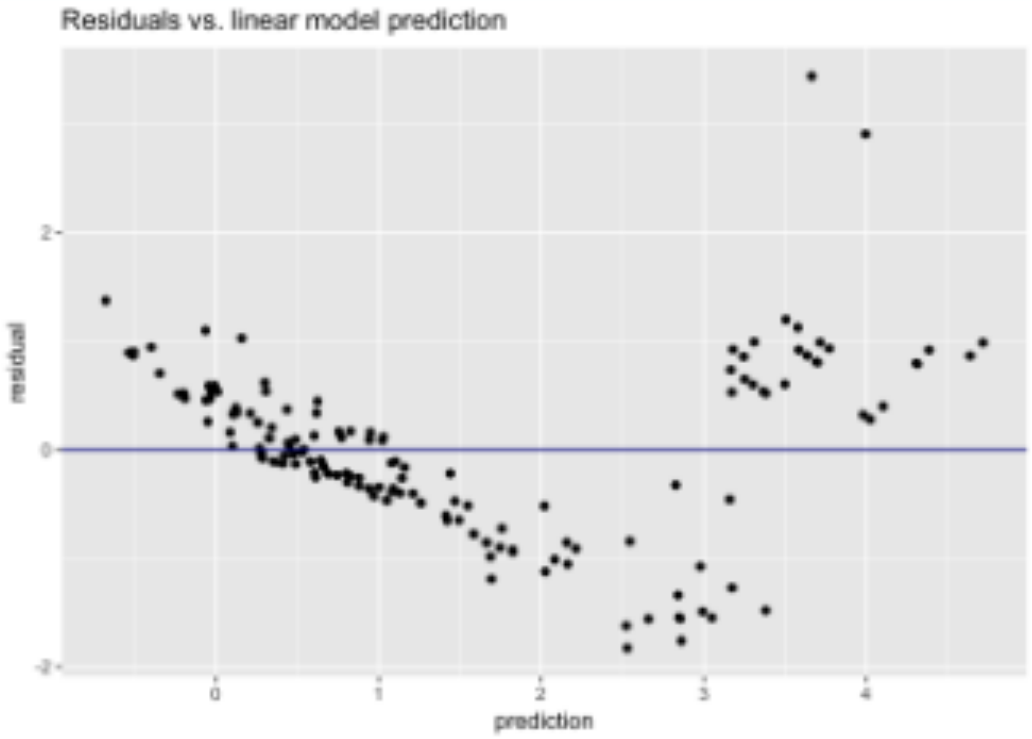
$$e = y - \hat{y}$$

Es la diferencia entre el valor predicho y el valor observado.

RESIDUALES

La diferencia entre el valor predicho y el valor observado.





RAÍZ DEL ERROR CUADRÁTICO MEDIO

Es el error de predicción típico, el cual se busca disminuir con el modelo de regresión lineal.

$$RMSE = \sqrt{\overline{(\widehat{y_i} - y_i)^2}}$$

¿ES EL RMSE ALTO O BAJO?

Para identificar si el RMSE es alto o bajo, se compara con la desviación estándar de la variable que queremos predecir.

Si este es menor, quiere decir que el modelo es mucho mejor que tomar para la predicción de los datos la media.

R^2 : COEFICIENTE DE DETERMINACIÓN

Es un estimador para identificar qué tan ajustado está el modelo a los datos.

Es un valor entre 0 y 1.

$$R^2 = 1 - \frac{RSS}{SS_{tot}}$$

EVALUACIÓN DE UN MODELO DE REGRESIÓN DE MANERA GRÁFICA