ITSRLL
INSTITUTO TECNOLÓGICO SUPERIOR DE LA REGIÓN DE LOS LLANOS

# Ingeniería Mecatrónica

## PROGRAMACIÓN AVANZADA

Enero – Junio 2022

M.C. Osbaldo Aragón Banderas

UNIDAD:

| 1 | **2** | 3 | 4 | 5 |
|---|---|---|---|---|

Actividad número: 4

Nombre de actividad:

NOOTEBOOK: Análisis de Datos Aplicables a Regresión Lineal Simple.

Actividad realizada por:

Melissa Gómez Rentería.

Guadalupe Victoria, Durango

Fecha de entrega:

| 09 | 03 | 2025 |
|---|---|---|

# REGRESIÓN LINEAL SIMPLE.

La **regresión lineal simple** es un método estadístico que nos ayuda a entender y predecir la relación entre dos variables. La idea principal es encontrar una línea recta que mejor represente cómo una variable depende de otra.

En este caso, trabajamos con:

- **Variable independiente (X)**: Es la que usamos como referencia, la que creemos que influye en la otra.
- **Variable dependiente (Y)**: Es la que queremos predecir o analizar con base en X.

Este tipo de modelo es útil cuando queremos responder preguntas como:

- ¿Cómo influye la experiencia laboral en el salario?
- ¿Cómo afecta la temperatura al consumo de electricidad?
- ¿Existe una relación entre la inversión en publicidad y las ventas de un producto?

Básicamente, lo que hace la regresión lineal simple es ajustar una recta sobre los datos disponibles para que nos ayude a hacer predicciones o entender tendencias.

## Ecuación de la Regresión Lineal Simple

El modelo de regresión lineal simple se representa con la ecuación:

$$Y = b_0 + b_1 X + \varepsilon$$

Donde:

- Y es la variable que queremos predecir.
- X es la variable que usamos para hacer la predicción.
- $b_0$ es el **intercepto**, el valor de Y cuando X=0.
- $b_1$ es la **pendiente**, que nos dice cuánto cambia Y por cada unidad de X.

- ε es el **error**, porque en la práctica nunca podemos predecir Y con total exactitud.

Por ejemplo, si encontramos que la ecuación del salario en función de la experiencia es:

Salario = 5000 + 2000 x Años de experiencia

Esto significaría que un empleado sin experiencia (0 años) ganaría $5000, y por cada año adicional de experiencia su salario aumentaría en $2000.

## Cómo se Encuentra la Mejor Línea de Ajuste (Método de Mínimos Cuadrados)

Para encontrar la mejor recta que se ajuste a los datos, usamos el **método de mínimos cuadrados**, que minimiza la diferencia entre los valores reales (Y) y los valores predichos.

El objetivo es que la suma de los errores al cuadrado sea lo más pequeña posible. Esto se logra encontrando los valores óptimos de $b_0$ y $b_1$ con las siguientes fórmulas:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

En términos simples, esta fórmula nos dice:

1. Si la pendiente $b_1$ es positiva, significa que a medida que X aumenta, Y también tiende a aumentar.
2. Si $b_1$ es negativa, entonces cuando X aumenta, Y disminuye.

Para visualizar esto, podemos graficar un **diagrama de dispersión** con los puntos de datos y superponer la línea de regresión para ver qué tan bien se ajusta.

## ¿Por qué es útil la regresión lineal simple?

Este modelo es una de las herramientas más básicas en el análisis de datos, pero sigue siendo muy útil. Se usa en diversas áreas como economía, ingeniería, medicina y marketing para analizar tendencias y hacer predicciones.

Por supuesto, este modelo tiene sus **limitaciones**. No siempre una relación entre dos variables es lineal, y factores externos pueden influir en los resultados. Aun así, cuando se usa correctamente, es una forma poderosa de entender relaciones y tomar decisiones basadas en datos.

## Búsqueda y selección de datos.

DATASET: **Predicción de Precios de Autos Usados**

Car Price Dataset

Este conjunto de datos es especialmente adecuado para la regresión lineal simple, ya que permite analizar cómo una variable independiente numérica, como el kilometraje, afecta al precio de un automóvil (variable dependiente continua). Esta relación es directa y fácilmente interpretable, lo que facilita el análisis y la comprensión de los resultados.

## Preprocesamiento de datos.

```
data=pd.read_csv('car_price_dataset.csv')
print(f'data.shape: {data.shape}')
data.head()
```

```
data.shape: (10000, 10)
```

| | Brand | Model | Year | Engine_Size | Fuel_Type | Transmission | Mileage | Doors | Owner_Count | Price |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Kia | Rio | 2020 | 4.2 | Diesel | Manual | 289944 | 3 | 5 | 8501 |
| **1** | Chevrolet | Malibu | 2012 | 2.0 | Hybrid | Automatic | 5356 | 2 | 3 | 12092 |
| **2** | Mercedes | GLA | 2020 | 4.2 | Diesel | Automatic | 231440 | 4 | 2 | 11171 |
| **3** | Audi | Q5 | 2023 | 2.0 | Electric | Manual | 160971 | 2 | 1 | 11780 |
| **4** | Volkswagen | Golf | 2003 | 2.6 | Hybrid | Semi-Automatic | 286618 | 3 | 3 | 2867 |

***Figura 1: Carga del dataset.***

```
df.describe(include=[np.number]).T
```

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Year** | 10000.0 | 2011.54370 | 6.897699 | 2000.0 | 2006.00 | 2012.0 | 2017.0 | 2023.0 |
| **Engine_Size** | 10000.0 | 3.00056 | 1.149324 | 1.0 | 2.00 | 3.0 | 4.0 | 5.0 |
| **Mileage** | 10000.0 | 149239.11180 | 86322.348957 | 25.0 | 74649.25 | 149587.0 | 223577.5 | 299947.0 |
| **Doors** | 10000.0 | 3.49710 | 1.110097 | 2.0 | 3.00 | 3.0 | 4.0 | 5.0 |
| **Owner_Count** | 10000.0 | 2.99110 | 1.422682 | 1.0 | 2.00 | 3.0 | 4.0 | 5.0 |
| **Price** | 10000.0 | 8852.96440 | 3112.596810 | 2000.0 | 6646.00 | 8858.5 | 11086.5 | 18301.0 |

```
df.describe(include=[object]).T
```

| | count | unique | top | freq |
|---|---|---|---|---|
| **Model** | 10000 | 30 | Accord | 365 |
| **Fuel_Type** | 10000 | 4 | Electric | 2625 |
| **Transmission** | 10000 | 3 | Manual | 3372 |

***Figura 2: Limpieza de datos.***

```
plt.figure(figsize=(8, 5))
sns.histplot(df["Price"], bins=30, kde=True, color="blue")
plt.xlabel("Precio del auto (en miles de dólares)")
plt.ylabel("Frecuencia")
plt.title("Distribución de los precios de los autos")
plt.show()
```

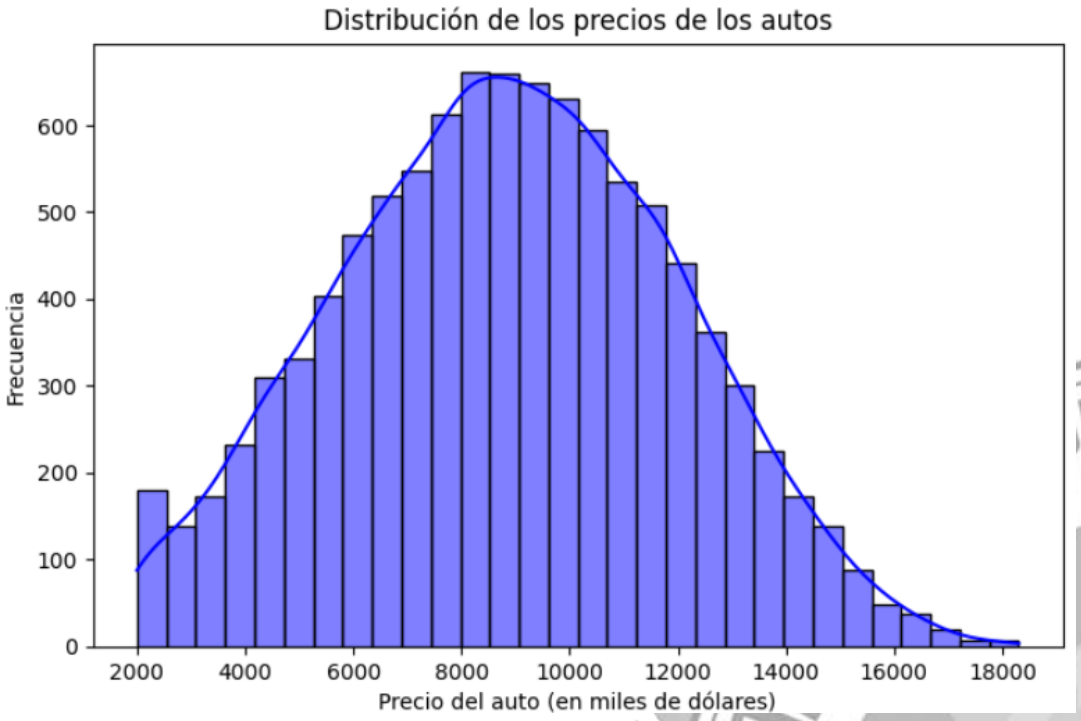


***Figura 3: Gráficos de distribución.***

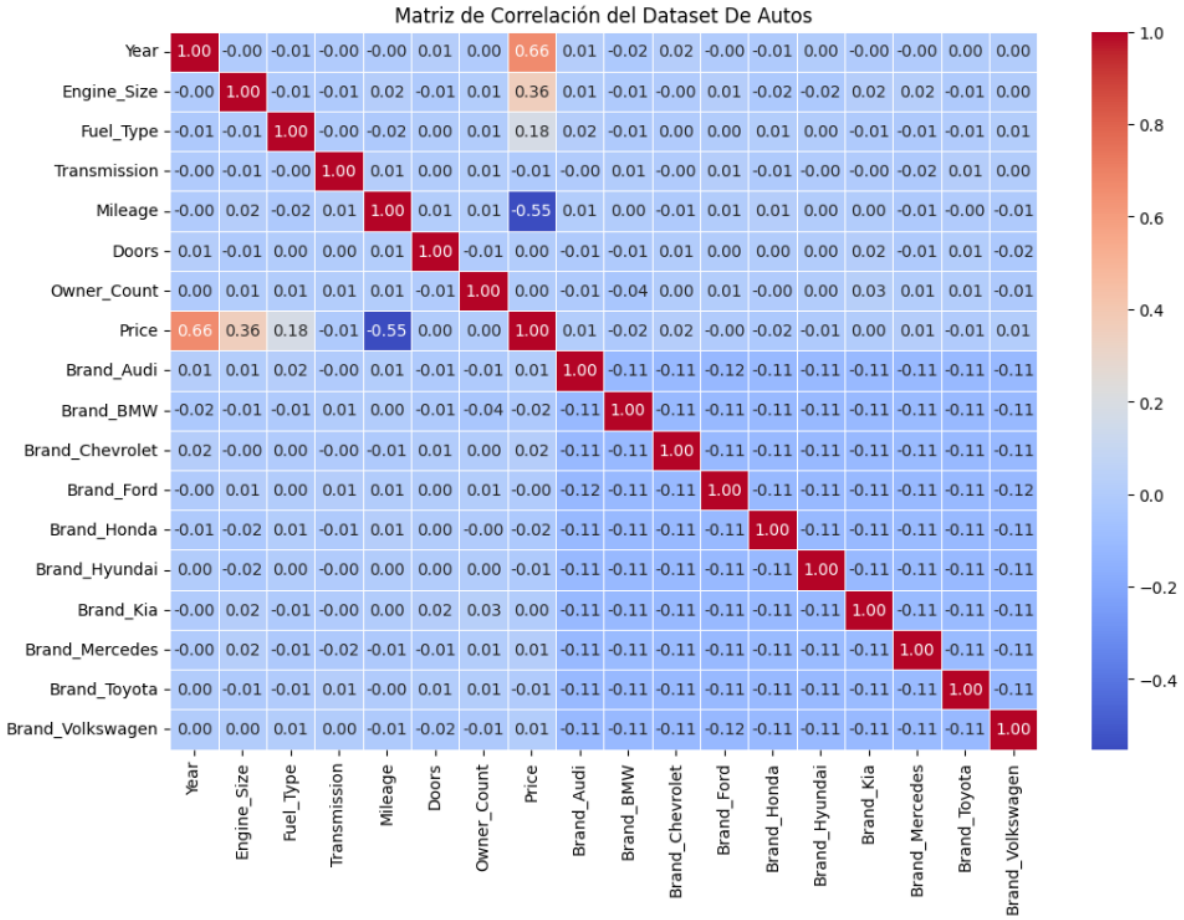


***Figura 4:  Gráficos de dispersión.***

## Aplicación y evaluación del Modelo de Regresión Lineal Simple.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Definir variables predictoras y objetivo
X = df.drop(columns=["Price"])  # Todas las columnas excepto MEDV
y = df["Price"]  # Variable objetivo

# Dividir los datos en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crear y entrenar el modelo de regresión lineal
model = LinearRegression()
model.fit(X_train, y_train)

# Realizar predicciones
y_pred = model.predict(X_test)

# Evaluar el modelo
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

# Mostrar métricas de evaluación
metrics = pd.DataFrame({
    "Métrica": ["Error Absoluto Medio (MAE)", "Error Cuadrático Medio (MSE)",
                "Raíz del Error Cuadrático Medio (RMSE)", "Coeficiente de Determinación (R²)"],
    "Valor": [mae, mse, rmse, r2]
})

print("Métricas de Evaluación del Modelo")
print(metrics)
```

```
Métricas de Evaluación del Modelo
                              Métrica          Valor
0              Error Absoluto Medio (MAE)     824.478745
1            Error Cuadrático Medio (MSE)  866700.544077
2  Raíz del Error Cuadrático Medio (RMSE)     930.967531
3       Coeficiente de Determinación (R²)       0.905668
```

***Figura 5: Aplicación y evaluación del modelo.***

```python
# Crear la figura
plt.figure(figsize=(8, 6))
# Graficar los valores reales vs. predichos
plt.scatter(y_test, y_pred, alpha=0.5, color="blue", label="Predicciones")
# Graficar la línea ideal (y = x)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], '--', color='red', label="Ideal")
# Etiquetas y título
plt.xlabel("Valor Real (Price)")
plt.ylabel("Valor Predicho (Price)")
plt.title("Comparación de Valores Reales vs. Predichos")
plt.legend()
# Mostrar la gráfica
plt.show()
```

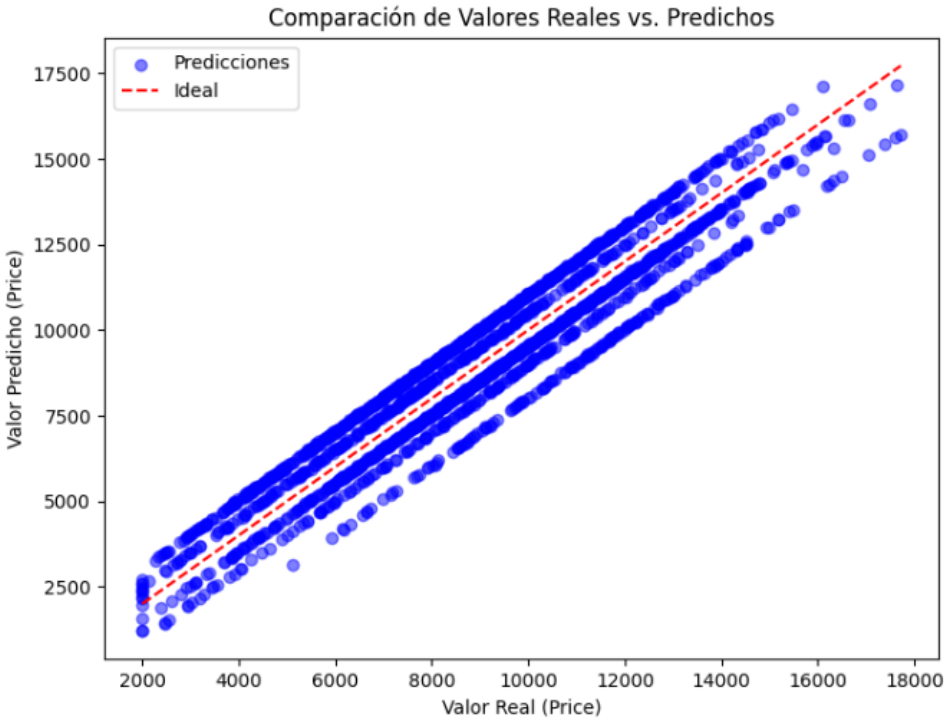


***Figura 6: Gráfico de la línea de regresión.***

## Análisis de resultados.

### 1. Interpretación del valor de los coeficientes

El modelo de regresión lineal simple encontró un coeficiente para la variable "Año del auto" ($X$), lo que indica cómo cambia el precio del auto ($Y$) por cada unidad de incremento en el año de fabricación.

Si el coeficiente es negativo, significa que los autos más antiguos tienden a valer menos. Si es positivo, indica que los autos más nuevos suelen tener un precio más alto.

Ejemplo:

Si el coeficiente es −1000, significa que por cada año que el auto es más viejo, su precio disminuye en $1000.

**2. Explicación del significado de $R^2$.**

El coeficiente de determinación indica qué porcentaje de la variabilidad del precio de los autos puede ser explicada por el año de fabricación.

Si $R^2$=0.8, significa que el 80% de la variabilidad en los precios se explica por el año del auto.

Si $R^2$ es bajo significa que el modelo no explica bien los cambios en el precio, lo que indica que otros factores (marca, kilometraje, estado del auto, etc.) influyen más.

**3. Evaluación de la fuerza de la relación entre variables**

Relación fuerte: Si $R^2$ >0.7, significa que el modelo es bueno para predecir el precio basado en el año.

Relación moderada: Si 0.4< $R^2$ < 0.7, la variable año tiene influencia, pero hay otros factores importantes.

Relación débil: Si $R^2$ < 0.4, el modelo no es muy útil y hay que considerar más variables.

**4. Posibles mejoras o ajustes al modelo**

Si el $R^2$ es bajo o los errores son altos, se pueden hacer mejoras como:

- Incluir más variables: En lugar de solo el año, agregar características como kilometraje, marca, tipo de combustible, transmisión, potencia del motor, etc.
- Transformaciones de datos: A veces, los datos pueden necesitar transformaciones como escalado o normalización.
- Eliminar valores atípicos: Si hay autos con precios extremadamente altos o bajos, podrían distorsionar el modelo.