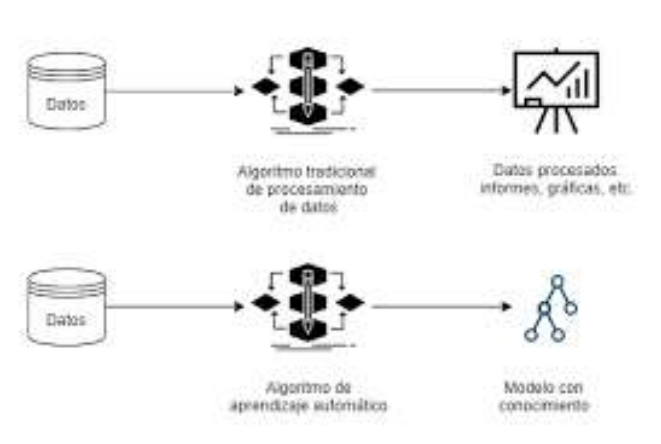


T301

Ing. Maximiliano Carsi Castrejón – Extracción y Conocimiento en Bases de Datos

DESCRIPCIÓN BREVE

Este documento trata sobre la Investigación de Comprensión de la Regresión Lineal Simple

Luis Eduardo Bahena Castillo
9°C IDyGS

INTRODUCCION

Tarea de Investigación: Comprensión de la Regresión Lineal Simple

Objetivo: Profundizar en la comprensión de la regresión lineal simple a través de la investigación de los conceptos teóricos fundamentales.

Instrucciones:

Parte 1: Investigación Teórica

1. **Conceptos Básicos de la Regresión Lineal Simple:**
   - Investigar y definir los siguientes términos:
     - Variable dependiente (respuesta).
     - Variable independiente (predictora).
     - Coeficiente de regresión (β0 y β1).
     - Error (ϵ).
   - Explicar la fórmula de la regresión lineal simple: $y=\beta0+\beta1x+\epsilon$
2. **Método de Mínimos Cuadrados:**
   - Investigar cómo se calculan los coeficientes de regresión (β0 y β1) utilizando el método de mínimos cuadrados.
   - Describir el proceso para minimizar la suma de los cuadrados de los errores (residuos).
3. **Evaluación del Modelo:**
   - Investigar las métricas utilizadas para evaluar el modelo de regresión lineal simple, tales como:
     - R-cuadrado (R^2).
     - Error Cuadrático Medio (MSE).
   - Explicar qué representan estas métricas y cómo se interpretan.

Entregables:

1. **Informe Teórico:**
   - Un documento en formato PDF que incluya la investigación teórica sobre los conceptos básicos, el método de mínimos cuadrados y la evaluación del modelo.
   - Debe incluir referencias a las fuentes utilizadas para la investigación.

Instrucciones para la Entrega:

- El reporte debe ser un archivo .pdf con la siguiente nomenclatura: `P(# de practica)_Apaterno_Amaterno_Nombre(s).pdf`.
- Cualquier archivo en otra extensión o con un nombre incorrecto no será evaluado.

Recursos Sugeridos:

1. **Libros y Artículos:**

- "An Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
- "Applied Linear Statistical Models" by John Neter, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman.

2. **Cursos y Tutoriales en Línea:**
   - Coursera: "Machine Learning" by Andrew Ng (Regresión lineal).
3. **Documentación de R:**
   - Documentación de la función `lm` en R
   - Documentación de la función `summary.lm` en R

DESARROLLO

La regresión lineal es un modelo matemático que describe la relación entre varias variables. Los modelos de regresión lineal son un procedimiento estadístico que ayuda a predecir el futuro. Se utiliza en los campos científicos y en los negocios, y en las últimas décadas se ha utilizado en el aprendizaje automático.

La regresión lineal simple se trata de establecer una relación entre una variable independiente y su correspondiente variable dependiente. Esta relación se expresa como una línea recta. No es posible trazar una línea recta que pase por todos los puntos de un gráfico si estos se encuentran ordenados de manera caótica. Por lo tanto, sólo se determina la ubicación óptima de esta línea mediante una regresión lineal.

Conceptos Básicos de la Regresión Lineal Simple

1. **Variable Dependiente (respuesta):** También conocida como variable de respuesta, es una variable cuyo valor depende de cómo se cambia la variable independiente. Debe predecirse o explicarse mediante un modelo de regresión. y generalmente se expresa como **y** y su valor depende de la variable independiente.
2. **Variable Independiente (predictora):** También llamada variable predictora. Esta es una variable que representa la cantidad que cambió en un experimento y se utiliza para predecir o explicar cambios en la variable dependiente. Esto se conoce comúnmente como **x** y se utiliza para modelar la variación en la variable dependiente.
3. **Coeficiente de Regresión (β0 y β1):**
   - **β0 (Intercepto):** Parámetro de pendiente en la relación entre **x** e **y**: Este es el cambio en **y** multiplicado por el cambio en **x**. Este es un parámetro importante en la aplicación y representa la intersección de la línea de regresión y el eje **y**.
   - **β1 (Pendiente):** También conocida como "pendiente", indica cuánto aumenta y por cada unidad de aumento en **x**. Devuelve el cambio promedio en y por cada unidad adicional de **x**. Representa la pendiente de la recta de regresión.
4. **Error (ϵ):** El error, denotado **ϵ**, es la diferencia entre el valor observado y el valor predicho por el modelo. Esto refleja la variabilidad en y que no puede explicarse por **x**. Se supone que los errores tienen media cero y varianza constante.

Fórmula de la Regresión Líneal

$$y = \beta 0 + \beta 1x + \epsilon$$

Donde:

y es la variable dependiente: el valor que desea predecir en función de la variable independiente x

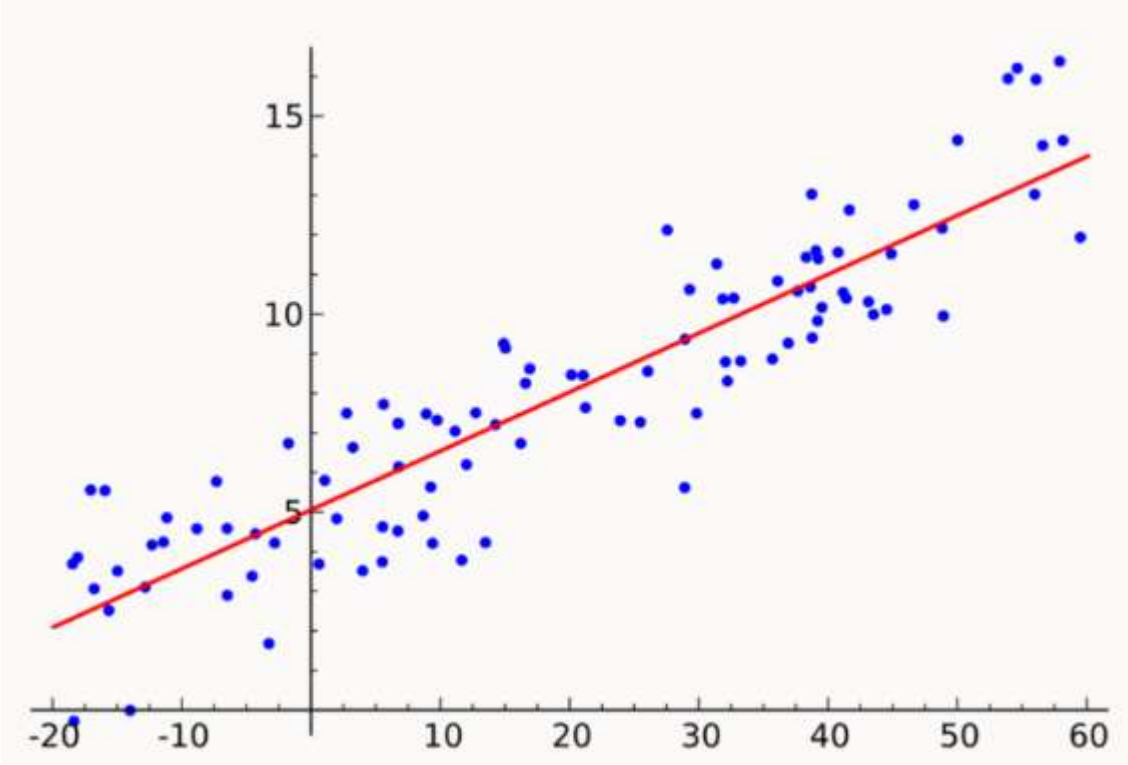
x es la variable independiente: variable utilizada para predecir el valor de y.

β0 es el intercepto: el valor de la y cuando x es igual a cero, intercepta en y.

β1 es la pendiente: cambio de y para cada unidad de cambio x representa la pendiente de la recta.

ϵ es el error: la diferencia entre los valores observados y predichos captura la variabilidad que el modelo no puede explicar.

A continuación, se muestra un ejemplo gráfico de un modelo de una regresión lineal simple:



Aplicación de la regresión lineal simple:

Para predecir la cosecha en función de la precipitación, con la precipitación como variable independiente y la cosecha como variable dependiente.

Para saber qué calificación obtendrán los alumnos en función del número de horas que estudien: aquí la cantidad de horas de estudio representa la variable independiente y las calificaciones, la dependiente.

Para prever el salario basado en la experiencia: la experiencia se convierte en la variable independiente y el salario en la variable dependiente.

Cálculo de los Coeficientes de Regresión (β0 y β1):

El método de mínimos cuadrados se utiliza para estimar los coeficientes beta 0 y beta 1 minimizando la suma de los cuadrados de los errores (residuos). Esto se hace para obtener la línea de mejor ajuste que minimiza la discrepancia entre los valores observados y los valores predichos por el modelo de regresión. El principal objetivo del método de mínimos cuadrados es encontrar los valores óptimos de los coeficientes intercepto y pendiente que hagan que la línea de regresión sea la más cercana posible a los datos observados. La cercanía se mide mediante la minimización de la suma de los cuadrados de las diferencias (errores) entre los valores observados y los valores predichos.

Pasos para el cálculo:

1. **Calcular la media de x y la media de y:**
   - $\bar{x} = \frac{1}{n} \quad \sum_{i=1}^{n} xi$
   - $\bar{y} = \frac{1}{n} \quad \sum_{i=1}^{n} yi$

   Donde n es el número de observaciones.

2. **Calcular la pendiente:**
   - $\beta 1 = \frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{\sum_{i=1}^{n}(xi-\bar{x})^2}$

   Este cálculo involucra determinar cómo varían y y x conjuntamente y cómo varía x por sí mismo.

3. **Calcular el intercepto (β0):**

$$\beta 0 = \bar{y} - \beta 1 \bar{x}$$

   El intercepto se calcula utilizando las medias de **x** y **y** la pendiente previamente calculada.

Proceso para Minimizar la Suma de los Cuadrados de los Errores:

El objetivo es minimizar la función de costo S:

$$S = \sum_{i=1}^{n} (yi - \hat{y}i)^2$$

Donde:

- yi es el valor observado de la variable dependiente.
- $\hat{y}i$ es el valor predicho por el modelo, calculado como: $\hat{y}i = \beta 0 + \beta 1 \text{xi}$

Minimizar S significa encontrar los valores de β0 y β1 que hagan que la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos sea la menor posible.

Evaluación del Modelo

Métricas Utilizadas para Evaluar el Modelo de Regresión Lineal Simple:

1. **R-cuadrado (R^2):** El coeficiente de determinación, mide la proporción de la variabilidad en y que se puede explicar mediante x. Se calcula como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (yi - \hat{y}i)^2}{\sum_{i=1}^{n} (yi - \bar{y})^2}$$

2. **Error Cuadrático Medio (MSE):** El Error Cuadrático Medio (Mean Squared Error) mide la media de los cuadrados de los errores. Se calcula como:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (yi - \hat{y}i)^2$$

Interpretación de las Métricas:

- **R^2:** Un valor alto de (cercano a 1) sugiere que el modelo explica bien la variabilidad de los datos. Sin embargo, un valor bajo no necesariamente indica un mal modelo, especialmente en contextos donde la variabilidad es inherentemente alta.
- **MSE:** Un valor bajo de MSE indica que los valores predichos están cerca de los valores observados. Sin embargo, es importante considerar la escala de los datos al interpretar el MSE. Una interpretación adecuada requiere comparar el MSE con la variabilidad inherente de los datos.

CONCLUSIÓN

La regresión lineal simple es una técnica estadística básica que se utiliza para modelar la relación entre una variable dependiente y una variable independiente. Esta relación se representa mediante una línea recta que busca ajustarse lo mejor posible a los datos observados. El objetivo es encontrar una línea que minimice la diferencia entre los valores observados y los valores predichos por el modelo.

Para determinar esta línea de mejor ajuste, se utilizan ciertos pasos y cálculos. Primero, se calcula la media de las variables involucradas. Luego, se estima la pendiente de la línea basándose en la relación entre la variación de las dos variables. Finalmente, se determina el punto de intersección de la línea con el eje de las ordenadas, utilizando las medias previamente calculadas.

Es fundamental evaluar la eficacia del modelo de regresión lineal simple para entender qué tan bien describe la relación entre las variables. Entre las principales métricas de evaluación se encuentran el coeficiente de determinación y el error cuadrático medio. El coeficiente de determinación indica qué proporción de la variación en la variable dependiente puede ser explicada por la variable independiente. Un valor cercano a uno sugiere que el modelo explica adecuadamente la variabilidad de los datos, mientras que un valor bajo sugiere lo contrario. Por otro lado, el error cuadrático medio mide la magnitud del error en las predicciones del modelo; un valor más bajo indica que las predicciones están más cerca de los valores observados.

Fuentes

- https://ebac.mx/blog/regreson-lineal
- https://datatab.es/tutorial/linear-regression