

P303

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

DESCRIPCIÓN BREVE

Este documento trata sobre solucionar un problema en lenguaje de programación R

Luis Eduardo Bahena Castillo

9°C IDyGS



INTRODUCCIÓN

Práctica: Regresión Lineal con el Dataset Iris

Objetivo de la Práctica:

Aprender a calcular una regresión lineal simple y verificar los resultados utilizando R. Esta práctica incluye el cálculo de medias, varianza, covarianza y los coeficientes de regresión (β_0 y β_1), así como la predicción de nuevos valores. Los estudiantes verificarán sus resultados usando R y evaluarán el modelo con R^2 .

Parte 1: Cálculo Digital en R

1. Selección de Variables

- **Variable Independiente (x):** `Sepal.Length`
- **Variable Dependiente (y):** `Petal.Length`

2. Pasos a Seguir en R

1. **Cargar el Dataset y Calcular Medias:**
 - Calcular la media de `Sepal.Length`
 - Calcular la media de `Petal.Length`
2. **Calcular la Varianza de x y la Covarianza de x y:**
 - Calcular la varianza de `Sepal.Length`
 - Calcular la covarianza entre `Sepal.Length` y `Petal.Length`
3. **Calcular los Coeficientes de Regresión (β_0 y β_1):**
 - Calcular β_1
 - Calcular β_0
4. **Predicción de un Nuevo Valor:**
 - Utilizar los coeficientes calculados para predecir el `Petal.Length` para un nuevo valor de `Sepal.Length` (por ejemplo, `xnuevo=5.5`)
5. **Evaluación del Modelo:**
 - Calcular el R^2
 - Graficar los datos y la línea de regresión

Parte 2: Reporte

Estructura del Reporte:

1. **Introducción:**
 - Explicación breve del objetivo de la práctica y la importancia de la regresión lineal.
2. **Cálculos y Resultados:**
 - **Medias:**
 - Media de `Sepal.Length`
 - Media de `Petal.Length`
 - **Varianza y Covarianza:**

- Varianza de `Sepal.Length`
 - Covarianza entre `Sepal.Length` y `Petal.Length`
 - **Coeficientes de Regresión:**
 - β_1 : β_1
 - β_0 : β_0
 - **Predicción:**
 - Predicción de `Petal.Length` para `Sepal.Length` = 5.5
 - **Evaluación del Modelo:**
 - R^2
3. **Gráficos:**
- Gráfico de dispersión con la línea de regresión.
4. **Conclusiones:**
- Resumen de los hallazgos.
 - Importancia de verificar los cálculos utilizando herramientas de software.

Entrega:

- **Reporte:** Subir un informe en formato PDF que incluya la introducción, cálculos y resultados, gráficos y conclusiones.
- **Código R:** Subir el código R utilizado para la verificación.

DESARROLLO

Introducción

El objetivo de esta práctica es aprender a calcular una regresión lineal simple y verificar los resultados utilizando R. En particular, usaremos la variable `Sepal.Length` como predictor (x) para predecir la variable `Petal.Length` (y) en el dataset `Iris`. Calcularemos las medias, varianza, covarianza y los coeficientes de regresión (β_0 y β_1), así como la predicción de nuevos valores. Evaluaremos el modelo con R^2 y visualizaremos los datos junto con la línea de regresión.

Cálculos y Resultados

Medias

- **Media de `Sepal.Length`:**

$$\bar{x} = \text{mean}(\text{iris}\$Sepal.Length)$$

- **Media de `Petal.Length`:**

$$\bar{y} = \text{mean}(\text{iris}\$Petal.Length)$$

Varianza y Covarianza

- **Varianza de Sepal.Length:**

$$Var(x) = var(iris\$Sepal.Length)$$

- **Covarianza entre Sepal.Length y Petal.Length:**

$$Cov(x, y) = cov(iris\$Sepal.Length, iris\$Petal.Length)$$

Coefficientes de Regresión

- **β_1 \beta_1:**

$$\beta_1 = \frac{Cov(x, y)}{Var(x)}$$

- **β_0 \beta_0:**

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

Predicción

- **Predicción de Petal.Length para Sepal.Length = 5.5:**

$$\hat{y} = \beta_0 + \beta_1 \cdot 5.5$$

Evaluación del Modelo

- **R^2 :**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- **Gráfico de Dispersión con Línea de Regresión:**

Graficar los datos y la línea de regresión ajustada.

Capturas

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save Run Source

```

1 # Cargar el dataset iris
2 data(iris)
3
4 # Selección de variables
5 x <- iris$Sepal.Length
6 y <- iris$Petal.Length
7
8 # Calcular medias
9 mean_x <- mean(x)
10 mean_y <- mean(y)
11
12 # Calcular varianzas de x y covarianza entre x e y
13 var_x <- var(x)
14 cov_xy <- cov(x, y)
15
16 # Calcular coeficientes de regresión
17 beta0 <- cov_xy / var_x
18 beta1 <- mean(y) - beta0 * mean(x)
19
20 # Predicción para un nuevo valor de Sepal.Length = 5.5
21 x_nuevo <- 5.5
22 y_pred <- beta0 * x_nuevo + beta1
23
24 # Calcular R^2
25 y_hat <- beta0 * x_nuevo + beta1
26 SST <- sum((y - mean(y))^2)
27
28 # Calcular R^2
29 R2 <- 1 - (sum((y_hat - y)^2) / SST)
30
31 # Mostrar resultados
32 print(beta0, beta1, y_hat, R2)
  
```

Además: Warning message: In file(file, "rt", "w") : no fue posible abrir el archivo 'documentos/resultados/regresion.csv': No such file or directory

> # Guardar la tabla de resultados en un archivo CSV

> write.csv(resultado, "resultados/regresion.csv", row.names = FALSE)

> # Cargar el dataset iris

> data(iris)

> # Selección de variables

> x <- iris\$Sepal.Length

> y <- iris\$Petal.Length

> # Calcular medias

> mean_x <- mean(x)

> mean_y <- mean(y)

> # Calcular varianzas de x y covarianza entre x e y

> var_x <- var(x)

> cov_xy <- cov(x, y)

> # Calcular coeficientes de regresión

> beta0 <- cov_xy / var_x

> beta1 <- mean(y) - beta0 * mean(x)

> # Predicción para un nuevo valor de Sepal.Length = 5.5

> x_nuevo <- 5.5

> y_pred <- beta0 * x_nuevo + beta1

> # Calcular R^2

> y_hat <- beta0 * x_nuevo + beta1

> SST <- sum((y - mean(y))^2)

> R2 <- 1 - (sum((y_hat - y)^2) / SST)

> # Mostrar resultados

> print(beta0, beta1, y_hat, R2)

Environment History Connections Tutorial

R Global Environment

iris 150 obs. of 5 variables

values

mean_x 5.84333333333333

mean_y 3.758

x num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.6 4.9 ...

y num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.3 1.4 1.3 ...

Files Plots Packages Help Viewer Presentation

Martes, 25 de junio, 11:08:29

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save Run Source

```

1 # Cargar el dataset iris
2 data(iris)
3
4 # Selección de variables
5 x <- iris$Sepal.Length
6 y <- iris$Petal.Length
7
8 # Calcular medias
9 mean_x <- mean(x)
10 mean_y <- mean(y)
11
12 # Calcular varianzas de x y covarianza entre x e y
13 var_x <- var(x)
14 cov_xy <- cov(x, y)
15
16 # Calcular coeficientes de regresión
17 beta0 <- cov_xy / var_x
18 beta1 <- mean(y) - beta0 * mean(x)
19
20 # Predicción para un nuevo valor de Sepal.Length = 5.5
21 x_nuevo <- 5.5
22 y_pred <- beta0 * x_nuevo + beta1
23
24 # Calcular R^2
25 y_hat <- beta0 * x_nuevo + beta1
26 SST <- sum((y - mean(y))^2)
27
28 # Calcular R^2
29 R2 <- 1 - (sum((y_hat - y)^2) / SST)
30
31 # Mostrar resultados
32 print(beta0, beta1, y_hat, R2)
  
```

Además: Warning message: In file(file, "rt", "w") : no fue posible abrir el archivo 'documentos/resultados/regresion.csv': No such file or directory

> # Guardar la tabla de resultados en un archivo CSV

> write.csv(resultado, "resultados/regresion.csv", row.names = FALSE)

> # Cargar el dataset iris

> data(iris)

> # Selección de variables

> x <- iris\$Sepal.Length

> y <- iris\$Petal.Length

> # Calcular medias

> mean_x <- mean(x)

> mean_y <- mean(y)

> # Calcular varianzas de x y covarianza entre x e y

> var_x <- var(x)

> cov_xy <- cov(x, y)

> # Calcular coeficientes de regresión

> beta0 <- cov_xy / var_x

> beta1 <- mean(y) - beta0 * mean(x)

> # Predicción para un nuevo valor de Sepal.Length = 5.5

> x_nuevo <- 5.5

> y_pred <- beta0 * x_nuevo + beta1

> # Calcular R^2

> y_hat <- beta0 * x_nuevo + beta1

> SST <- sum((y - mean(y))^2)

> R2 <- 1 - (sum((y_hat - y)^2) / SST)

> # Mostrar resultados

> print(beta0, beta1, y_hat, R2)

Environment History Connections Tutorial

R Global Environment

iris 150 obs. of 5 variables

values

beta0 -7.39144336908245

beta1 1.85843297825404

cov_xy 1.27431543624161

mean_x 5.84333333333333

mean_y 3.758

var_x 6.665693512304251

x num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.6 4.9 ...

x_nuevo 5.5

y num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.3 1.4 1.3 ...

y_pred 3.11893881070917

Files Plots Packages Help Viewer Presentation

Martes, 25 de junio, 11:08:39

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save | Run | Source | Addins

```

20 y_pred = betab + x_nuevo
21
22 # Calcular R^2
23 khat = betab + betab * x
24 SST = sum((y - mean_y)^2)
25 SSE = sum((y_hat - mean_y)^2)
26 SSE = sum((y - y_hat)^2)
27 R2 = 1 - SSE / SST
28
29 # Resultados
30 cat("Media de Sepal.Length:", mean_x, "\n")
31 cat("Media de Petal.Length:", mean_y, "\n")
32 cat("Varianza de Sepal.Length:", var_x, "\n")
33 cat("Covarianza entre Sepal.Length y Petal.Length:", cov_xy, "\n")
34 cat("Coeficiente de Regresión B1:", betab, "\n")
35 cat("Coeficiente de Regresión B0:", betab, "\n")
36 cat("Predicción de Petal.Length para Sepal.Length = 5.5:", y_pred, "\n")
37 cat("R^2 del modelo:", R2, "\n")
38
39 # (Fin del programa)
  
```

R 4.1.2

```

> mean_y = mean(y)
> # Calcular varianza de x y covarianza entre x y y
> var_x = var(x)
> cov_xy = cov(x, y)
> betab = cov_xy / var_x
> betab0 = mean_y - betab * mean_x
> x_nuevo = 5.5
> y_pred = betab0 + betab * x_nuevo
> y_hat = betab0 + betab * x
> SST = sum((y - mean_y)^2)
> SSE = sum((y_hat - mean_y)^2)
> SSE = sum((y - y_hat)^2)
> R2 = 1 - SSE / SST
>
  
```

Environment History Connections Tutorial

R - Global Environment

Beta

iris 158 obs. of 5 variables

Values

Variable	Value
betab	-7.39244336900240
betab0	1.05843297825404
cov_xy	1.27431543624161
mean_x	5.84333333333333
mean_y	3.758
R2	0.75955665772515
SSE	111.459155119019
SST	352.860244000181
SST	464.3254
var_x	0.685693112304251
x	num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.4 5 4.4 4.8 ...
x_nuevo	5.5
y	num [1:150] 1.6 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
y_hat	num [1:150] 2.36 2 1.63 1.40 2.19 ...
y_pred	3.1199181079917

Files Plots Packages Help Viewer Presentation

Files Plots Packages Help Viewer Presentation

Martes, 25 de Junio, 21:07:17

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save | Run | Source | Addins

```

26 SST = sum((y - mean_y)^2)
27 SSE = sum((y_hat - mean_y)^2)
28 SSE = sum((y - y_hat)^2)
29 R2 = 1 - SSE / SST
30
31 # Resultados
32 cat("Media de Sepal.Length:", mean_x, "\n")
33 cat("Media de Petal.Length:", mean_y, "\n")
34 cat("Varianza de Sepal.Length:", var_x, "\n")
35 cat("Covarianza entre Sepal.Length y Petal.Length:", cov_xy, "\n")
36 cat("Coeficiente de Regresión B1:", betab, "\n")
37 cat("Coeficiente de Regresión B0:", betab, "\n")
38 cat("Predicción de Petal.Length para Sepal.Length = 5.5:", y_pred, "\n")
39 cat("R^2 del modelo:", R2, "\n")
40
41 # (Fin del programa)
  
```

R 4.1.2

```

> cat("Media de Sepal.Length:", mean_x, "\n")
Media de Sepal.Length: 5.843333
> cat("Media de Petal.Length:", mean_y, "\n")
Media de Petal.Length: 3.758
> cat("Varianza de Sepal.Length:", var_x, "\n")
Varianza de Sepal.Length: 0.6856931
> cat("Covarianza entre Sepal.Length y Petal.Length:", cov_xy, "\n")
Covarianza entre Sepal.Length y Petal.Length: 1.274315
> cat("Coeficiente de Regresión B1:", betab, "\n")
Coeficiente de Regresión B1: -7.392443
> cat("Coeficiente de Regresión B0:", betab0, "\n")
Coeficiente de Regresión B0: 1.058433
> cat("Predicción de Petal.Length para Sepal.Length = 5.5:", y_pred, "\n")
Predicción de Petal.Length para Sepal.Length = 5.5: 3.119918
> cat("R^2 del modelo:", R2, "\n")
R^2 del modelo: 0.7595566
>
  
```

Environment History Connections Tutorial

R - Global Environment

Beta

iris 158 obs. of 5 variables

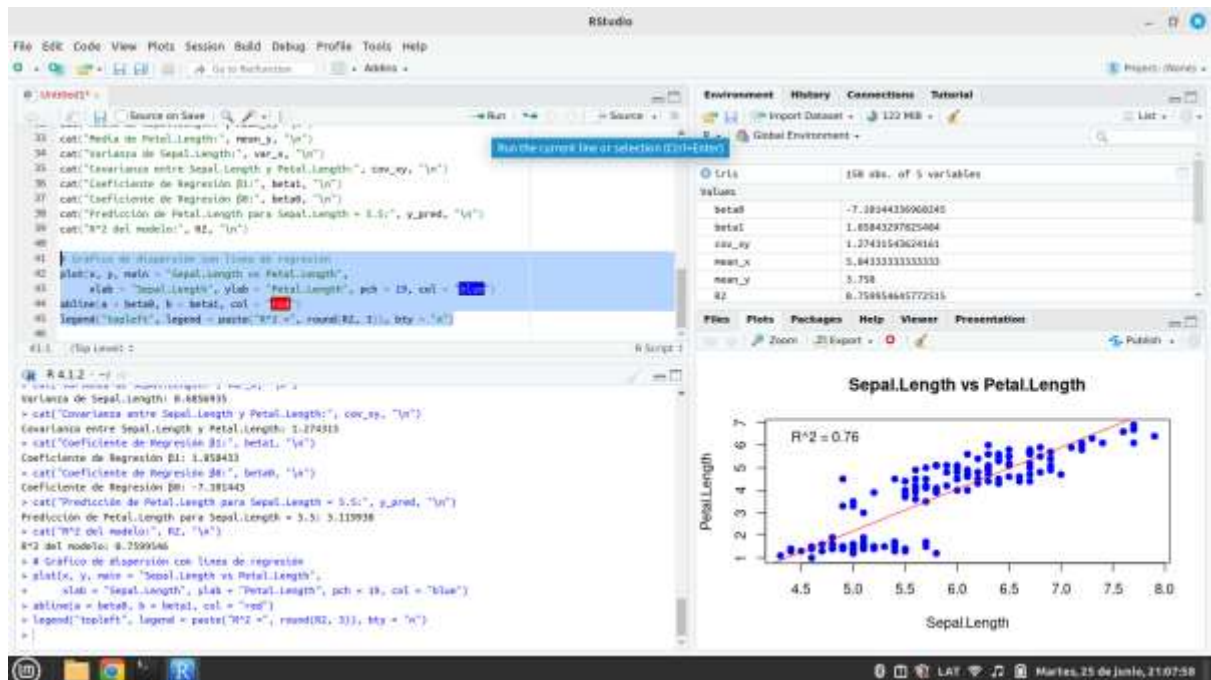
Values

Variable	Value
betab	-7.39244336900240
betab0	1.05843297825404
cov_xy	1.27431543624161
mean_x	5.84333333333333
mean_y	3.758
R2	0.75955665772515
SSE	111.459155119019
SST	352.860244000181
SST	464.3254
var_x	0.685693112304251
x	num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.4 5 4.4 4.8 ...
x_nuevo	5.5
y	num [1:150] 1.6 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
y_hat	num [1:150] 2.36 2 1.63 1.40 2.19 ...
y_pred	3.1199181079917

Files Plots Packages Help Viewer Presentation

Files Plots Packages Help Viewer Presentation

Martes, 25 de Junio, 21:07:40



Conclusiones

Este análisis muestra cómo se puede utilizar la regresión lineal para entender la relación entre dos variables. Los coeficientes de regresión obtenidos nos permiten predecir nuevos valores de Petal.Length basados en Sepal.Length. La evaluación del modelo con R^2 y la visualización gráfica ayuda a verificar la precisión del modelo. Es importante utilizar herramientas de software como R para realizar y verificar estos cálculos de manera eficiente y precisa.