**Vargas Carrillo Raul Paulino**

**Regresión Lineal Simple**

Obtener la información utilizando la librería lm de R deberá adjuntar tanto el código utilizado como los resultados obtenidos. Para los conjuntos de datos que se describen en cada uno de los retos deberá obtener la siguiente información:

* Crear un diagrama de dispersión que nos muestre la relación entre las variables.
* Definir el modelo que se va a ajustar.
* Obtener las estimaciones de los parámetros del modelo.
* Interpretación de los parámetros obtenidos.
* Crear una gráfica de dispersión con la recta de regresión que representa el modelo ajustado.

**1. Reto**

**Gráfico, Gráfico de dispersión

Descripción generada automáticamente**El archivo de Excel precipitación contienen la información de una extensa región de pastizales donde se ubicaron 4 áreas con diferentes valores de Precipitación pluvial anual conocidos de antemano (respectivamente 300 mm,600 mm, 900 mm y 1200 mm). Dentro de cada área se eligieron al azar 4 parcelas de 100 m2 cada una y en cada parcela medir la productividad primaria neta del pastizal (PPN).

|  |  |  |  |
| --- | --- | --- | --- |
| Precipitación | ppn | Precipitación | ppn |
| 300 | 80 | 900 | 560 |
| 300 | 150 | 900 | 590 |
| 300 | 125 | 900 | 400 |
| 300 | 170 | 900 | 380 |
| 600 | 390 | 1200 | 790 |
| 600 | 470 | 1200 | 610 |
| 600 | 225 | 1200 | 640 |
| 600 | 350 | 1200 | 680 |

precipitacion %>%

ggplot(aes(Precipitacion, ppn)) +

geom\_point()

* **Definir el modelo que se va a ajustar.**

#ppn = B0 +B1 \* Precipitación

* **Obtener las estimaciones de los parámetros del modelo.**

modppn <- lm(ppn~Precipitacion, precipitacion)

#ppn = -29.37 + 0.59 \* Precipitacion

predppn = -29.37 + 0.59 \* 300

residuom = 80 – predppn

> modppn

Call:

lm(formula = ppn ~ Precipitacion, data = precipitacion)

Coefficients:

(Intercept) Precipitacion

-29.37 0.59

> predppn = -29.37 + 0.59 \* 300

> predppn

[1] 147.63

> residuom = 80 - predppn

> residuom

[1] -67.63

> # Predicciones

> predict(modppn)

1 2 3 4 5

147.625 147.625 147.625 147.625 324.625

6 7 8 9 10

324.625 324.625 324.625 501.625 501.625

11 12 13 14 15

501.625 501.625 678.625 678.625 678.625

16

678.625

* **Interpretación de los parámetros obtenidos.**

Por cada unidad que aumenta de precipitación, la productividad aumenta en 0.59. Respecto al valor real (80) de precipitación con el valor predicho (147.63), hay un error o residuo de -67.63.

* **Crear una gráfica de dispersión con la recta de regresión que representa el modelo ajustado.**

#Ajustar el modelo / Línea del modelo ajustado

precipitacion %>%

ggplot(aes(Precipitacion, ppn)) +

geom\_point() +

Gráfico, Gráfico de dispersión

Descripción generada automáticamente geom\_smooth(method = "lm", se = F)

summary(modppn)

Call:

lm(formula = ppn ~ Precipitacion, data = precipitacion)

Residuals:

Min 1Q Median 3Q Max

-121.625 -67.875 1.875 60.125 145.375

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -29.37500 50.84476 -0.578 0.573

Precipitacion 0.59000 0.06189 9.534 1.68e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Según los datos del summary, existe una relación estadísticamente significativa entre la variable dependiente y la independiente.

**2. Reto**

El archivo de Exel de ventas contiene la información de una muestra de 15 supermercados para las variables número de empleados y ventas semanales en cientos de miles de pesos. El interés del gerente es predecir la venta a partir del número de empleado.

|  |  |  |  |
| --- | --- | --- | --- |
| NumEmpleado | Ventas | NumEmpleado | Ventas |
| 10 | 4 | 32 | 11 |
| 17 | 6 | 32 | 13 |
| 17 | 6 | 36 | 15 |
| 17 | 5 | 40 | 16 |
| 19 | 9 | 41 | 16 |
| 24 | 8 | 43 | 17 |
| 24 | 12 | 48 | 20 |
| 30 | 12 |  |  |

* **Crear un diagrama de dispersión que nos muestre la relación entre las variables.**

> ventas %>%

+ ggplot(aes(NumEmpleado, Ventas)) +

+ geom\_point()

**Gráfico, Gráfico de dispersión

Descripción generada automáticamente**

* **Definir el modelo que se va a ajustar.**

#Ventas = B0 + B1\*NumEmpleado

* **Obtener las estimaciones de los parámetros del modelo.**

> modVentas <- lm(Ventas~NumEmpleado, ventas)

> modVentas

Call:

lm(formula = Ventas ~ NumEmpleado, data = ventas)

Coefficients:

(Intercept) NumEmpleado

-0.5894 0.4159

> # Ventas = -0.5894 + 0.4159\*NumEmpleado

> predVentas <- -0.5894 + 0.4159\*10

> # Ventas = -0.5894 + 0.4159\*NumEmpleado

> predVentas <- -0.5894 + 0.4159\*10

> predVentas

[1] 3.5696

> residuo <- 4 - predVentas

> residuo

[1] 0.4304

> predict(modVentas)

1 2 3 4 5 6 7 8 9

3.569712 6.481070 6.481070 6.481070 7.312887 9.392428 9.392428 11.887878 12.719694

10 11 12 13 14 15

12.719694 14.383327 16.046960 16.462869 17.294685 19.374226

> summary(modVentas)

Call:

lm(formula = Ventas ~ NumEmpleado, data = ventas)

Residuals:

Min 1Q Median 3Q Max

-1.71969 -0.48107 -0.04696 0.52348 2.60757

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.58937 0.85757 -0.687 0.504

NumEmpleado 0.41591 0.02791 14.900 1.5e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.195 on 13 degrees of freedom

Multiple R-squared: 0.9447, Adjusted R-squared: 0.9404

F-statistic: 222 on 1 and 13 DF, p-value: 1.499e-09

* **Interpretación de los parámetros obtenidos.**

Por cada unidad que aumente en empleados, las ventas incrementan en 0.41. Además, del valor predicho al valor real, hay un error o un residuo de 0.4304.

Entre la variable independiente y la variable dependiente existe una relación estadísticamente significativa.

* **Crear una gráfica de dispersión con la recta de regresión que representa el modelo ajustado.**

> ventas %>%

+ ggplot(aes(NumEmpleado, Ventas)) +

+ geom\_point() +

+ geom\_smooth(method = "lm", se = F)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente`geom\_smooth()` using formula = 'y ~ x'