Day 8: Least Square Regression Line

24/27 challenges solved

Points: 24

10 Days of Statistics

Problem | Submissions | Leaderboard | Editorial | Tutorial

Regression Line

If our data shows a linear relationship between X and Y, then the straight line which best describes the relationship is the regression line. The regression line is given by $\hat{Y} = a + bX$.

Finding the Value of b

The value of b can be calculated using either of the following formulae:

- $b = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{n\sum(x_i^2) - (\sum x_i)^2}$
- $b = \rho \cdot \frac{\sigma_Y}{\sigma_X}$, where ρ is the Pearson correlation coefficient, σ_X is the standard deviation of X and σ_Y is the standard deviation of Y.

Finding the Value of a

$a = \bar{y} - b \cdot \bar{x}$, where $\bar{x}$ is the mean of X and $\bar{y}$ is the mean of Y.

Sums of Squares

- `Total Sums of Squares`: $SST = \sum(y_i - \bar{y})^2$
- `Regression Sums of Squares`: $SSR = \sum(\hat{y}_i - \bar{y})^2$
- `Error Sums of Squares`: $SSE = \sum(\hat{y}_i - y_i)^2$

If SSE is small, we can assume that our fit is good.

Coefficient of Determination (R-squared)

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

R^2 multiplied by 100 gives the percent of variation attributed to the linear regression between Y and X.

Example

Let's consider following data sets:

- $X = \{1, 2, 3, 4, 5\}$
- $Y = \{2, 1, 4, 3, 5\}$

So,

- $n = 5$
- $\sum x_i = 15$
- $\bar{x} = \frac{\sum x_i}{n} = 3$
- $\sum y_i = 15$
- $\bar{y} = \frac{\sum y_i}{n} = 3$
- $X^2 = \{1, 4, 9, 16, 25\} \Rightarrow \sum(x_i^2) = 55$
- $XY = \{2, 2, 12, 12, 25\} \Rightarrow \sum(x_i y_i) = 53$

Now we can compute the values of a and b:

$$b = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{n\sum(x_i^2) - (\sum x_i)^2} = \frac{5 \times 53 - 15 \times 15}{5 \times 55 - 15^2} = \frac{40}{50} = 0.8$$

And,

$$a = \bar{y} - b\bar{x} = 3 - 0.8 \times 3 = 0.6$$

So, the regression line is $\hat{Y} = 0.6 + 0.8X$.

Linear Regression in R

We can use the lm function to fit a linear model.

```
x = c(1, 2, 3, 4, 5)
y = c(2, 1, 4, 3, 5)

m = lm(y ~ x)
summary(m)
```

Running the above code produces the following output:

```
Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4    5 
 0.6 -1.2  1.0 -0.8  0.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6000     1.1489   0.522    0.638
x             0.8000     0.3464   2.309    0.104

Residual standard error: 1.095 on 3 degrees of freedom
Multiple R-squared:   0.64,	Adjusted R-squared:   0.52 
F-statistic: 5.333 on 1 and 3 DF,  p-value: 0.1041
```

If we want information for coefficients only, we can use the following code:

```
x = c(1, 2, 3, 4, 5)
y = c(2, 1, 4, 3, 5)

lm(y ~ x)
```

Running the above code produces the following output:

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x  
        0.6          0.8  
```

Linear Regression in Python

We can use the fit function in the sklearn.linear_model.LinearRegression class.

```
from sklearn import linear_model
import numpy as np
xl = [1, 2, 3, 4, 5]
x = np.asarray(xl).reshape(-1, 1)
y = [2, 1, 4, 3, 5]
lm = linear_model.LinearRegression()
lm.fit(x, y)
print(lm.intercept_)
print(lm.coef_[0])
```

Running the above code produces the following output:

```
0.6
0.8
```