AccelerateAI

Data Science Global Bootcamp

Assignment 05

Simple Linear Regression

Q1: The Dean of a college wants to examine the effect of internship experience on marketability in the workplace. She takes a random sample of 4 students. For these 4, she finds out how many times each had an internship and how many job offers they received upon graduation. These data are presented in the table below.

Student	Internship(s)	Job Offer
1	1	4
2	2	6
3	1	3
4	0	1

1. What is the independent variable X?
 a) Internship
 b) Job Offers
 c) Marketability in the workplace
 d) None of the Above

2. Referring to the above data, the estimate of the slope is
 a) 0.4 b) 2 c) 2.50 d) 5

Solution:

Internship is the independent variable. **Job offer** is the dependent variable.

Context: Independent variable(s) is/are one that we change in order to look at (causal) effects on another variable which is the dependent variable.

Slope - In order to determine slope, we have to calculate the predicted value and check which one has the minimum square error summed for all datapoint.

Regression Equation: Job Offer = b * Internships, where b=slope.

	X	Y	Predicted Values				Squared Errors			
Student	Internship(s)	Job Offer	b = 0.4	b = 2	b = 2.5	b = 5	b = 0.4	b = 2	**b = 2.5**	b = 5
1	1	4	0.4	2	2.5	5	12.96	4	2.25	1
2	2	6	0.8	4	5	10	27.04	4	1	16
3	1	3	0.4	2	2.5	5	6.76	1	0.25	4
4	0	1	0	0	0	0	1	1	1	1
							47.76	10	4.5	22

For the given data, **slope=2.5** has the minimum sum of squared errors.

Q2: Score received on an exam measured in "percentage points" (Y) is regressed on "percentage attendance" (X) for 22 students in the course Statistics for Data Science. If the Y intercept, b0 = 39.39 and the slope, b1 = 0.341, which of the following statement is correct?

a) If attendance increases by 0.341%, the estimated average score received will increase by 1 percentage point.
b) If attendance increases by 1%, the estimated average score received will increase by 39.39 percentage points.
c) If attendance increases by 1%, the estimated average score received will increase by 0.341 percentage points.
d) If the score received increases by 39.39%, the estimated average attendance will go up by 1%.

Solution:

Regression Equation: Y = 39.39 + 0.341*X
where Y = percentage points, X = percentage attendance

So $\Delta Y = 0.341 * \Delta X$ Δ – change in the variable.

Unit change in X, will lead to 0.341 percentage point increase.

Answer: C) If attendance increases by 1%, the estimated average score received will increase by 0.341 percentage points.

Q3. A survey was conducted to see if a relation exists between expenditure on higher education (X), and Salary growth in percentage (Y). The result obtained is summarized in the table. Write down the regression equation.

	Mean	SD
Salary Growth (%)	178	63.15
Spend on Higher Education (1000)	47.8	22.9
Coefficient of Correlation	0.43	

Solution:

Regression equation is of the form Y = b0 + b1*X

We need to calculate b0 (the intercept) and b1 (the slope)

b1 = Slope = (Coefficient of Correlation) * ((Standard deviation of Y)/(Standard deviation of X))

b1 = (0.43) * ((63.15)/(22.9)) ~ 1.186

b0 = (mean of Y) - ((slope) * (mean of X))

= 178 - ((1.186*47.8)) ~ 121.309

Substituting for b0 and b1, we get: Y = 121.309 + (1.186) * X

Q4. A model was built to determine how crime rate in the neighbourhood impacts property prices in USA.
The incomplete coefficient table is shown below. Calculate the rate at which the property price changes for unit change in the Crime Rate (CRIM).

Table 1.4 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	24.033	.409		58.740	.000
	CRIM		.044		-9.460	.000

a. Dependent Variable: Price

Solution:

For the SLR Model Y(Price) = b0 + b1*(CRIM), we need to estimate b1.

As b1 is not provided in the coefficient table, we can find it from the T-value and Standard Error for b1.

T-estimated = (b1 – 0) / (Std. Error of b1) [T-test for parameter Beta]

For CRIM, the values are: T-estimated = -9.460, Std. Error = 0.044

Hence b1 = T-estimated * Std. Error = -9.460 * 0.044 = **-0.41624**

Q5. Ajishek Bacchan, researcher at GharDekho.com claims that for every unit increase in crime rate, the price will decrease by at least INR 30,000. Check whether Mr. Bacchan is correct at 95% confidence level.

He also claims that when CRIM = 0, the average price of the property will be 24.033. Is he correct? Explain your conclusion.

The regression equation is :

Price = $\beta 0 + \beta 1 * CRIM$ Price is in 100,000 USD, CRIM is in percentage.

Table 1.4 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	24.033	.409		58.740	.000
	CRIM		.044		-9.460	.000

Var	N	Minimum	Maximum	Mean	Std. Deviation
CRIM	506	1.26	88.98	3.61	8.60
Price	506	5.76	50.00	27.53	9.20

Solution

We have the following regression model:
Y(Price) = b0 + b1*(CRIM)

We calculated b1 = -0.41624 => this is the mean value
Now let's calculate the 95% CI for b1:
b1_lower = b1 - t(0.025, 504)* (Std Err)
b1_upper = b1 + t(0.025, 504)* (Std Err)

Here n = 506, so degrees of freedom = 504

b1_upper = -0.41624 + 1.964 * 0.044
= -0.3297

The price will decrease by 32,970 (at 95% CI). So Mr. Bacchan is correct.

Answer:

1) Yes, Mr Bacchan is correct on the first statement.

2) No. Though b0 = 24.033, since the minimum value of CRIM in datasets is 1.26, CRIM=0 is out of bounds, so we cannot conclude anything about model parameter when the independent variable is out of bound.

Q6. Tuck Maintenance:
A trucking company wants to predict its yearly maintenance expense for its trucks using miles driven. Its data is saved in the file: ***MLR_TruckMaintenence.xlsx***
a) Estimate the MLR equation for maintenance expense vs miles driven.
b) Interpret R-Squared for this dataset.

Solution File: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment05/AAI%20-%20DS%20-%20Assignment05%20-%20Solution%20Q6-7.ipynb

---

Q7. MLR – Overhead Cost
Benedrix, a machine tool company in interested in understand the impact of machine hours and production run, on its overhead cost.

The data on a monthly basis for 3 years is provided in the dataset:
MLR_FactoryOverhead.csv

Fit the regression equations:

- Overhead = F (machine hours)
- Overhead = F (production runs)

a) Is production run strongly correlated with machine hours?
b) Find the R-Squared in both the above cases?
c) Which variable is a better predictor of Overhead cost?

Solution file: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment05/AAI%20-%20DS%20-%20Assignment05%20-%20Solution%20Q6-7.ipynb

---