Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Sl. No	Questions

1

1. Professor Bell at Bellandur University, Bangalore believes that the cumulative grade point average (CGPA) of the students is negatively correlated with usage (measured in average minutes per day) of smart phones. Table 1 shows the CGPA and smart phone usage in minutes per day of 40 students.

(a) Calculate the Pearson correlation coefficient between CGPA and mobile phone usage of students.

(b) Conduct a hypothesis test at a = 0.01 to check whether CGPA and mobile phone usage are negatively correlated.

(c) Professor Bell believes that the correlation is less than −0.4. Conduct a hypothesis test at a = 0.1 to check whether the claim is correct.

Table.1: Data of CGPA and mobile phone usage (Average minutes per day)

CGPA	2.65	2.25	1.86	1.47	2.10	1.94	2.71	1.83	2.65	2.04
Phone Usage	75	89	65	136	95	103	74	109	7	98
CGPA	2.54	2.16	2.28	2.47	2.18	2.57	1.97	2.87	2.10	3.28
Phone Usage	60	93	88	81	92	78	102	70	95	89
CGPA	2.78	2.441.87	2.50	2.24	2.01	2.17	2.20	2.05	1.63	
Phone Usage	72	82	107	80	89	100	92	91	98	123
CGPA	2.28	2.63	2.86	2.24	2.44	2.69	2.22	3.07	1.77	3.03
Phone Usage	88	76	70	89	82	74	90	65	113	66

2

Mr Chellappa is the founder of Oho Productions that produces movies in different languages of India. Mr Chellappa believes that the length of the movie (measured in minutes) is not related to its box-office collection.

Table 2 shows length of the movie (in minutes) and the box-office collection (in millions of rupees). Use an appropriate hypothesis test to check whether there is a correlation between length of the movie and the box-office collection at a significance level of 0.05.

TABLE 2 Data on length of the movie and the box-office collection

Length of the movie	121	79	170	160	77	147	115	76	110	141
Box-office collection	1078	415	441	1192	258	1185	139	427	309	411

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Length of the movie	100	82	82	114	110	163	92	172	142	136
Box-office collection	506	441	595	1728	1507	518	1463	1356	1014	422
Length of the movie	143	108	154	140	177	97	106	163	142	115
Box-office collection	508	1262	1783	1281	1253	1178	1103	454	301	296

3 Table 3. provides ranking of Indian states based on corruption and Table 4. provides ranking based on literacy rate.

Calculate the Spearman rank correlation between the corruption rank and literacy rank.

TABLE 3 Rank based on corruption (1 implies high corruption)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	1	2	3	4	5	6	7	8
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	9	10	11	12	13	14	15	16

TABLE 4. Rank based on literacy rate (1 implies high literacy)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	16	12	10	11	7	15	4	9
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	2	5	13	1	8	14	6	3

Conduct a hypothesis test to check whether corruption and literacy rate are negatively correlated at a = 0.05.

4 .Harrison Seth, Dean of a Business School, believes that the outgoing salary of their MBA students may be correlated with their undergraduate specialization. Harrison believes that the students with engineering specialization at the undergraduate degree received more salary compared to other degrees.

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Table 5. shows the outgoing salary (in millions of rupees) of MBA graduates and their discipline in undergraduate (1 = engineering and 0 = non-engineering). Calculate the correlation between salary and engineering discipline,

TABLE 5. Salary (in millions of rupees) and undergraduate degree (1 = engineering and 0 = non-engineering)

Degree	0	1	0	1	0	0	1	0	0	1
Salary	3.3	2.22	1.82	2.55	1.84	2.53	2.87	2.39	2.32	2.79
Degree	1	1	0	1	0	0	1	1	0	0
Salary	2.22	2.31	2.05	2.04	1.7	2.28	2.56	3.13	2.26	2.56
Degree	0	0	0	0	1	0	0	0	1	1
Salary	2.03	1.45	1.62	.92	2.31	2.37	1.59	2.56	3.13	3

5 Tele power is a telephone service provider which collects data on customer churn and the number of mobile handsets used by the customer.

Table 6. shows the data in which Y denotes churn (Y = 1 implies churn and Y = 0 implies no churn) and variable X denotes the number of handsets used by the customer where X = 0 implies the customer uses single handset and X = 1 implies the customer uses more than one handset for making phone calls. Calculate the Phi-coefficient for the data shown in Table 6.

TABLE 6. Number of handsets (X) and customer churn (Y)

X	1	1	0	0	0	1	1	1	1	1
Y	1	1	1	1	0	0	1	0	1	1
X	0	1	1	1	1	0	0	1	1	1
Y	0	1	0	1	1	0	0	1	1	1
X	1	1	1	0	1	0	1	0	1	1
Y	0	1	1	0	1	0	0	1	1	1
X	1	1	1	1	0	1	1	0	1	1
Y	0	1	0	1	1	1	1	0	0	1
X	0	0	1	0	1	0	1	1	0	1
Y	0	0	1	1	1	0	0	1	1	1

6 For a simple linear regression, prove the following relationship between F-statistic and R^2: $F =(n-2) R^2 / (1- R^2)$. In a simple linear regression model, prove that the value of F-statistic is same as the square of t-statistic value (that is, $F = t^2$).

7 Price of a diamond is determined by 4Cs, namely, Carat, Cut, Clarity and Color. Carat is the weight of the diamond, and 1 carat is equivalent to 0.2 grams. Data on carat and price of 6000 diamonds are used for developing SLR models. The mean and the standard deviation of diamond price and carat are provided in Table 1.
TABLE 7. Descriptive statistics

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

	Carat	Price
Mean	1.33	11792
Standard Deviation	0.48	10184

A regression model (model 1) based on data of 6000 diamonds is developed using price as the dependent variable and carat as the independent variable.

Model 1: Y = β0 + β1 × Carat

The SPSS output for model 1 and the corresponding residual plot is shown in Table 7 and Figure 8, respectively.

TABLE 8. Regression co-efficient Model

8

Model		Unstandardized Coefficients		Standardized Coefficients		t-value	Sig.
		B	Std. Error	Beta			
1	(Constant)	−12738.581	200.801			−63.439	.000
	Carat	18381.261	141.733				

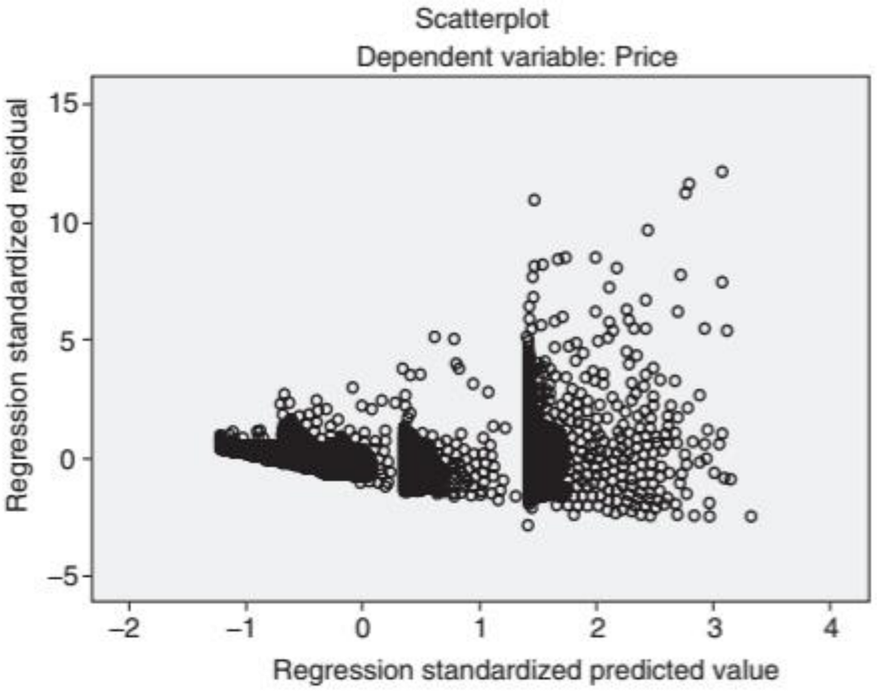


FIGURE 1. Plot between standardized predicted value versus standardized residual for model 1.

(c) What is the interpretation of the coefficient for the variable carat in model 2?

Data Analytics: UE18CS312

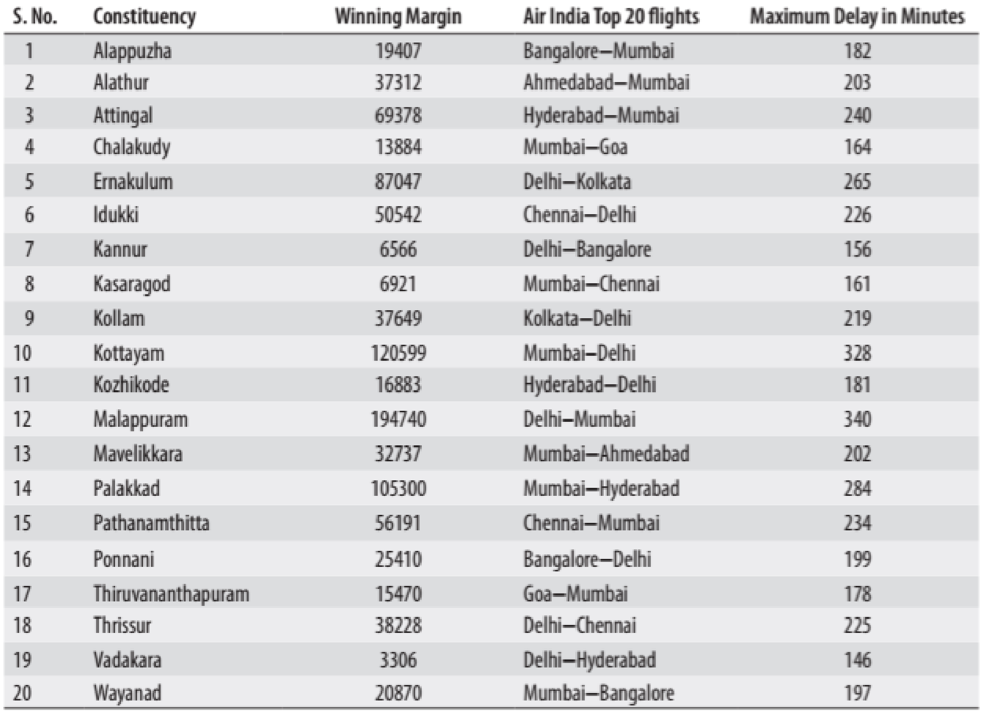
Question Bank

Unit -2 Regression Analysis

(d) Calculate the maximum possible price of a specific diamond whose weight is 0.4 grams at 95% confidence level using model 2.
(e) Between models 1 and 2, which model should be used to explain variation in the diamond price? State the reasons clearly.

9 Table 9. provides the winning margin of all 20 Lok Sabha constituencies of Kerala in 2014 parliament elections of India and maximum delay of top 20 flights (origin−destination) of Air India between 15 July 2014 and 15 September 2014.

TABLE 9 Data on Lok Sabha election winning margin of Kerala constituencies and maximum delay of top 20 Air India



S. No.	Constituency	Winning Margin	Air India Top 20 flights	Maximum Delay in Minutes
1	Alappuzha	19407	Bangalore—Mumbai	182
2	Alathur	37312	Ahmedabad—Mumbai	203
3	Attingal	69378	Hyderabad—Mumbai	240
4	Chalakudy	13884	Mumbai—Goa	164
5	Ernakulum	87047	Delhi—Kolkata	265
6	Idukki	50542	Chennai—Delhi	226
7	Kannur	6566	Delhi—Bangalore	156
8	Kasaragod	6921	Mumbai—Chennai	161
9	Kollam	37649	Kolkata—Delhi	219
10	Kottayam	120599	Mumbai—Delhi	328
11	Kozhikode	16883	Hyderabad—Delhi	181
12	Malappuram	194740	Delhi—Mumbai	340
13	Mavelikkara	32737	Mumbai—Ahmedabad	202
14	Palakkad	105300	Mumbai—Hyderabad	284
15	Pathanamthitta	56191	Chennai—Mumbai	234
16	Ponnani	25410	Bangalore—Delhi	199
17	Thiruvananthapuram	15470	Goa—Mumbai	178
18	Thrissur	38228	Delhi—Chennai	225
19	Vadakara	3306	Delhi—Hyderabad	146
20	Wayanad	20870	Mumbai—Bangalore	197

(a) Develop a simple linear regression model between winning margin (Y) and maximum flight delay (X) and calculate the regression coefficients.
(b) What is the value of R2?
(c) Is the model statistically significant, what can you infer from the regression model?

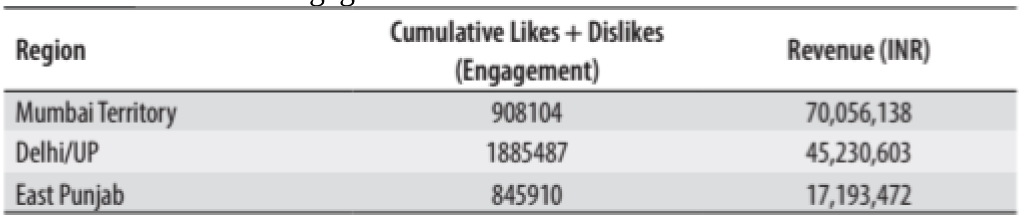
10 The box-office collection of a Bollywood movie across different regions and the corresponding social media engagement (likes + dislikes) is provided in Table

Data Analytics: UE18CS312

Question Bank

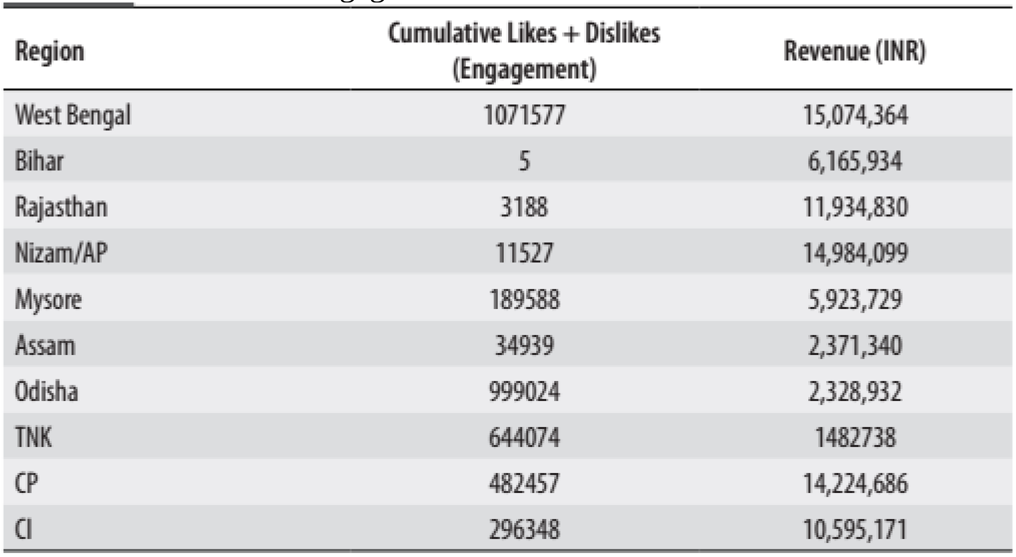
Unit -2 Regression Analysis

Table 10. Social media engagement versus box-office collection.



Region	Cumulative Likes + Dislikes (Engagement)	Revenue (INR)
Mumbai Territory	908104	70,056,138
Delhi/UP	1885487	45,230,603
East Punjab	845910	17,193,472

11 TABLE 11. Social media engagement versus box-office collection—Continued



Region	Cumulative Likes + Dislikes (Engagement)	Revenue (INR)
West Bengal	1071577	15,074,364
Bihar	5	6,165,934
Rajasthan	3188	11,934,830
Nizam/AP	11527	14,984,099
Mysore	189588	5,923,729
Assam	34939	2,371,340
Odisha	999024	2,328,932
TNK	644074	1482738
CP	482457	14,224,686
CI	296348	10,595,171

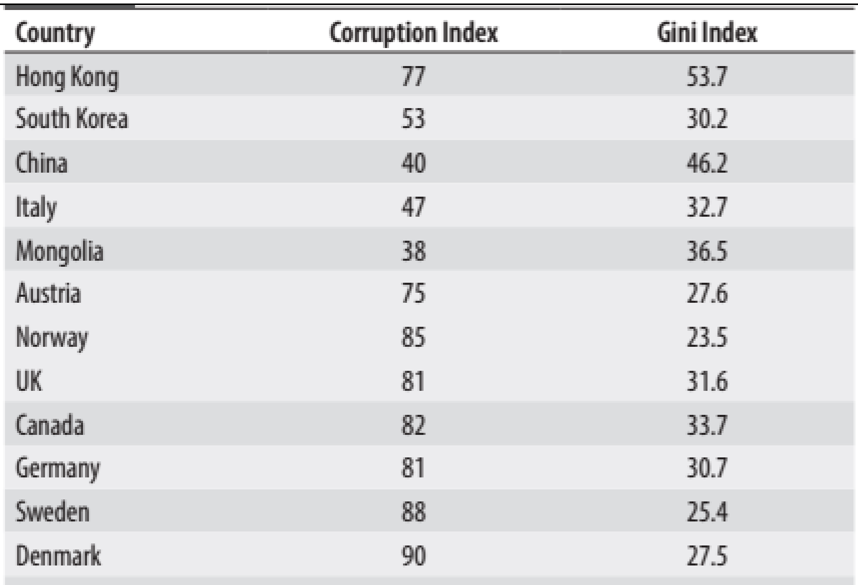
(a) Develop a simple linear regression model for the data shown in Table 11. Is there any evidence that the box-office collection (Y) of the movie has statistically significant relationship with the social media engagement (X)?
(b) What is the 95% confidence interval for the average box-office collection for a movie with 20,000 likes and dislikes?
(c) Should Bollywood movie producers invest more to promote their movies through social media?

12 Corruption perception index (source: Transparency International) and Gini Index (Source: Wikipedia) of 20 countries is shown in Table 11 Corruption perception index close to 100 indicates low corruption and close to 0 indicates high corruption. Gini index is a measure of income distribution among citizens of a country (high Gini indicates high inequality).
TABLE 12. Corruption Index and Gini Index

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis



Country	Corruption Index	Gini Index
Hong Kong	77	53.7
South Korea	53	30.2
China	40	46.2
Italy	47	32.7
Mongolia	38	36.5
Austria	75	27.6
Norway	85	23.5
UK	81	31.6
Canada	82	33.7
Germany	81	30.7
Sweden	88	25.4
Denmark	90	27.5

13 (a) Develop a simple linear regression model (Y = b0 + b1 X) between corruption perception index (Y) and Gini index (X). What is the change in the corruption perception index for every one-unit increase in Gini index?

(b) What proportion of the variation in corruption perception index is explained by Gini index?

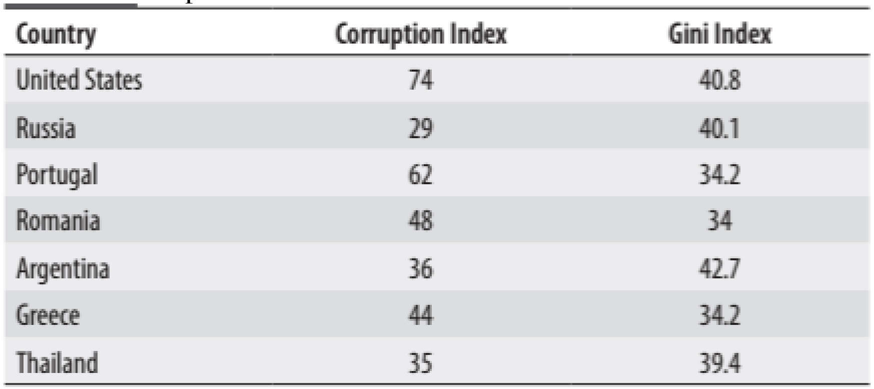
(c) Is there a statistically significant relationship between corruption perception index and Gini index at a = 0.1?

(d) Calculate the 95% confidence interval for the regression coefficient b1.

(e) Is it possible to conclude that the corruption perception index will decrease by at least 1 unit for every one-unit increase in Gini index? Conduct an appropriate hypothesis test at a = 0.05.

(f) Calculate 95% confidence interval for the expected value of corruption perception index for Gini index value = 30.

Table 13. Corruption Index and Gini Index—Continued



Country	Corruption Index	Gini Index
United States	74	40.8
Russia	29	40.1
Portugal	62	34.2
Romania	48	34
Argentina	36	42.7
Greece	44	34.2
Thailand	35	39.4

14 7. A regression model is developed between corruption perception index and per capita income (in US dollars) based on data on 20 countries. Regression model output obtained through Microsoft Excel is shown in Table 14. Note that Table 14 shows only partial output of the model developed.

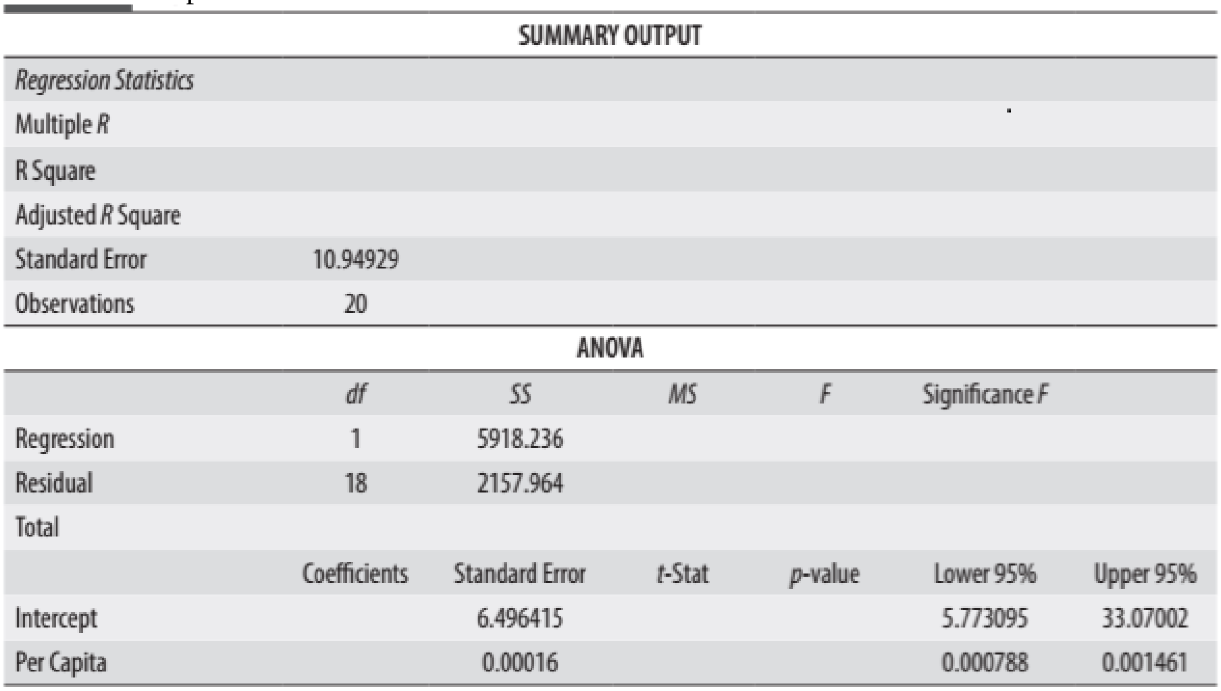
TABLE 14. Regression between corruption perception index (Y) and per capita (X)

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Table 14. Corruption Index and Gini Index—Continued



SUMMARY OUTPUT

Regression Statistics	
Multiple *R*	
R Square	
Adjusted *R* Square	
Standard Error	10.94929
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	Significance *F*	
Regression	1	5918.236				
Residual	18	2157.964				
Total						
	Coefficients	Standard Error	*t*-Stat	*p*-value	Lower 95%	Upper 95%
Intercept		6.496415			5.773095	33.07002
Per Capita		0.00016			0.000788	0.001461

(a) What proportion of the corruption perception index is explained by per capita?

(b) What is change in the value of corruption perception index for every one-dollar increase in per capita?

(c) Is there a statistically significant relationship between corruption perception index and per capita at a = 0.01?

(d) What is the average corruption perception index when per capita is $30,000. What is the corresponding 95% confidence interval?

(e) Per capita of a country is $30,000. What is the probability that the corruption perception index of this country is less than 50?

(f) Which of the following statements are true based on the model shown in Table 13?

(i) Corruption perception index and per capita are positively correlated.

(ii) Corruption perception index and per capita are negatively correlated.

(iii) There is no correlation between corruption perception index and per capita.

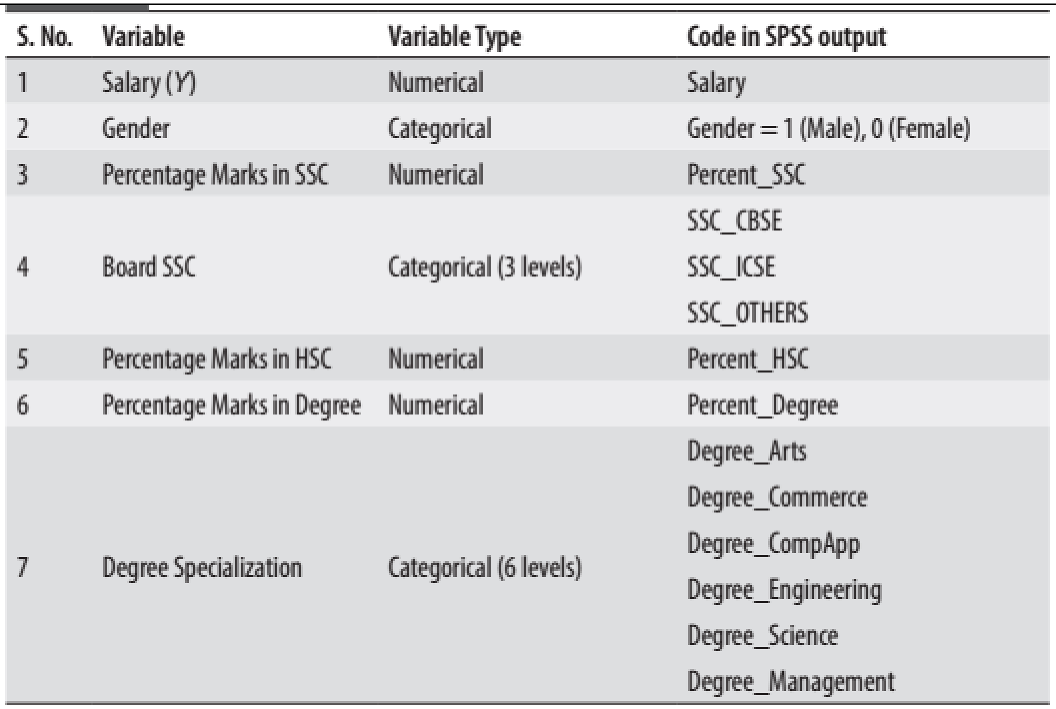
13 Data for Questions 1−6: The dean of a business school has collected data on their recent placement. To attract good students, it is important for the school to ensure that the students are placed with good salary package. The dean of the school believed that the salary earned by a student at placement depended on several variables. The data collected by the dean is listed in Table 15.

Table 15. Data Description.

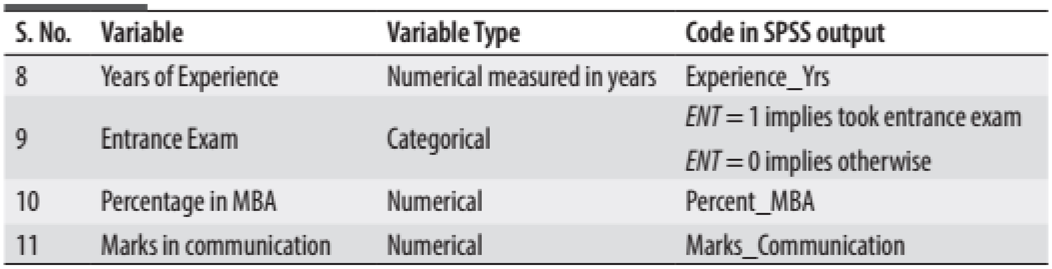
Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis



S. No.	Variable	Variable Type	Code in SPSS output
1	Salary (*Y*)	Numerical	Salary
2	Gender	Categorical	Gender = 1 (Male), 0 (Female)
3	Percentage Marks in SSC	Numerical	Percent_SSC
4	Board SSC	Categorical (3 levels)	SSC_CBSE SSC_ICSE SSC_OTHERS
5	Percentage Marks in HSC	Numerical	Percent_HSC
6	Percentage Marks in Degree	Numerical	Percent_Degree
7	Degree Specialization	Categorical (6 levels)	Degree_Arts Degree_Commerce Degree_CompApp Degree_Engineering Degree_Science Degree_Management



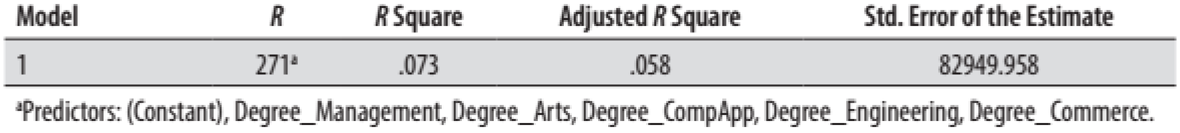
S. No.	Variable	Variable Type	Code in SPSS output
8	Years of Experience	Numerical measured in years	Experience_Yrs
9	Entrance Exam	Categorical	*ENT* = 1 implies took entrance exam *ENT* = 0 implies otherwise
10	Percentage in MBA	Numerical	Percent_MBA
11	Marks in communication	Numerical	Marks_Communication

The first regression model is built using degree of specialization as the explanatory variable.

$Y = \beta_0 + \beta_1$ Degree_Arts $+ \beta_2$ Degree_Commerce $+ \beta_3$ Degree_CompApp $+ \beta_4$ Degree_Engineering
$+ \beta_5$ Degree_Management.

The model 1 SPSS outputs are shown in Tables 16 - 18

Table. 16. Model Summary



Model	*R*	*R* Square	Adjusted *R* Square	Std. Error of the Estimate
1	271[a]	.073	.058	82949.958

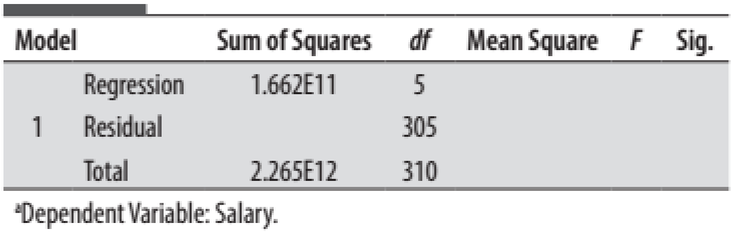
[a]Predictors: (Constant), Degree_Management, Degree_Arts, Degree_CompApp, Degree_Engineering, Degree_Commerce.

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

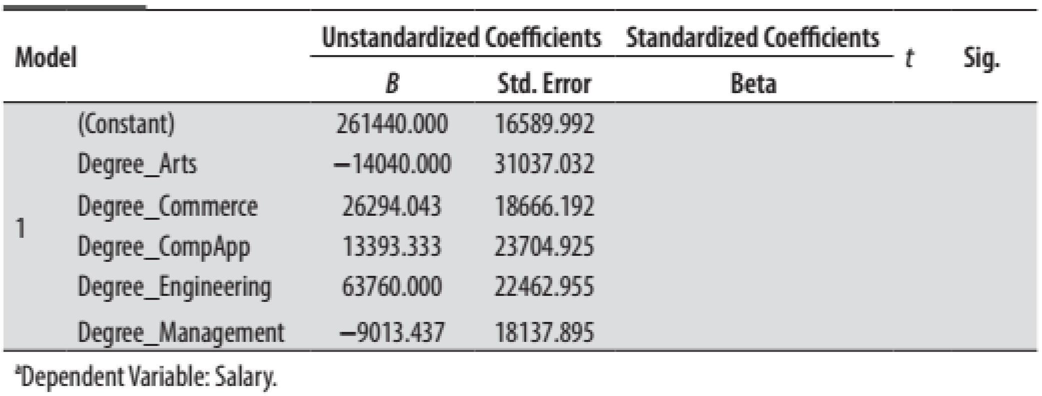
Table. 16. ANOVA



Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.662E11	5			
	Residual		305			
	Total	2.265E12	310			

ᵃDependent Variable: Salary.

Table. 16. Coefficients



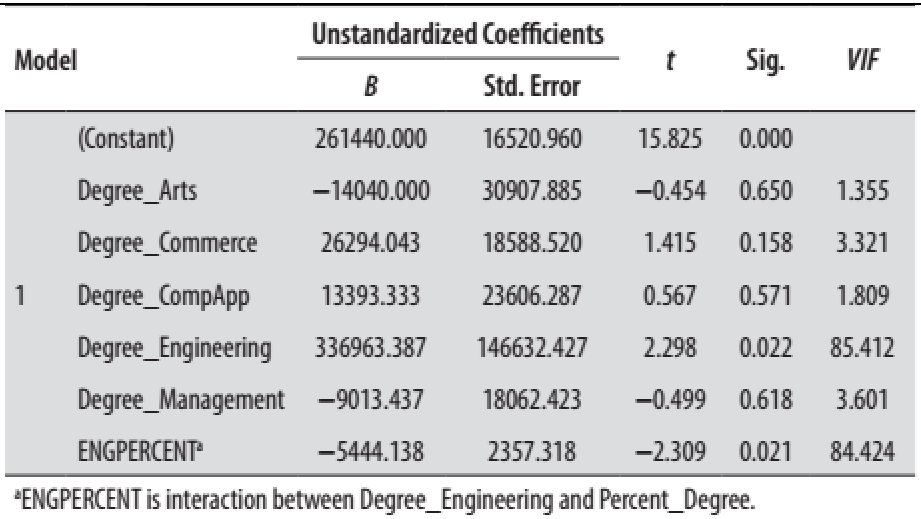
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	261440.000	16589.992			
	Degree_Arts	−14040.000	31037.032			
	Degree_Commerce	26294.043	18666.192			
	Degree_CompApp	13393.333	23704.925			
	Degree_Engineering	63760.000	22462.955			
	Degree_Management	−9013.437	18137.895			

ᵃDependent Variable: Salary.

14 Assuming that the salary package is important for the school, should the dean give more importance to certain degree disciplines while admitting the students to their MBA programme? Support your answers with precise arguments. 2. Is there a significant difference between the average salary earned by a student with science degree and commerce degree? Clearly state your arguments. 3. The dean of the school believes that the engineering students earn on average at least INR 25,000 more than the science students. Check whether his belief is true at 5% significance level by conducting an appropriate hypothesis tests. A new variable, which is the interaction between degree discipline engineering and the percentage marks in degree, is added to model 1 and the corresponding output is shown in Table 19.

Table. 17. Coefficients

Unit -2 Regression Analysis

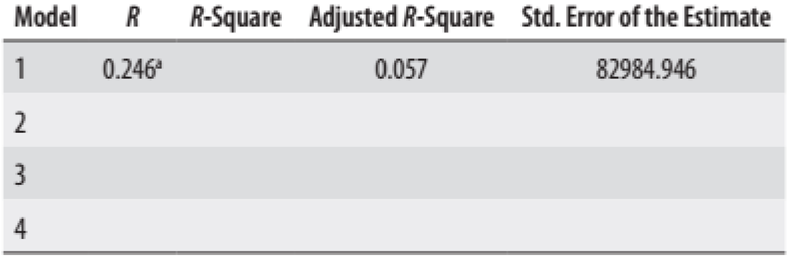


Model		Unstandardized Coefficients		t	Sig.	VIF
		B	Std. Error			
1	(Constant)	261440.000	16520.960	15.825	0.000	
	Degree_Arts	−14040.000	30907.885	−0.454	0.650	1.355
	Degree_Commerce	26294.043	18588.520	1.415	0.158	3.321
	Degree_CompApp	13393.333	23606.287	0.567	0.571	1.809
	Degree_Engineering	336963.387	146632.427	2.298	0.022	85.412
	Degree_Management	−9013.437	18062.423	−0.499	0.618	3.601
	ENGPERCENT[a]	−5444.138	2357.318	−2.309	0.021	84.424

[a]ENGPERCENT is interaction between Degree_Engineering and Percent_Degree.

15 Interpret the coefficient value for the interaction value ENGPERCENT (Degree_Engineering × Percent_Degree). Explain possible reason for the salary of engineering students decreasing as the percentage marks in degree increases. Clearly state your arguments. A stepwise regression is carried out using SPSS and the results of stepwise regression are shown in Tables 20 and 21.

Table. 17. Model summary



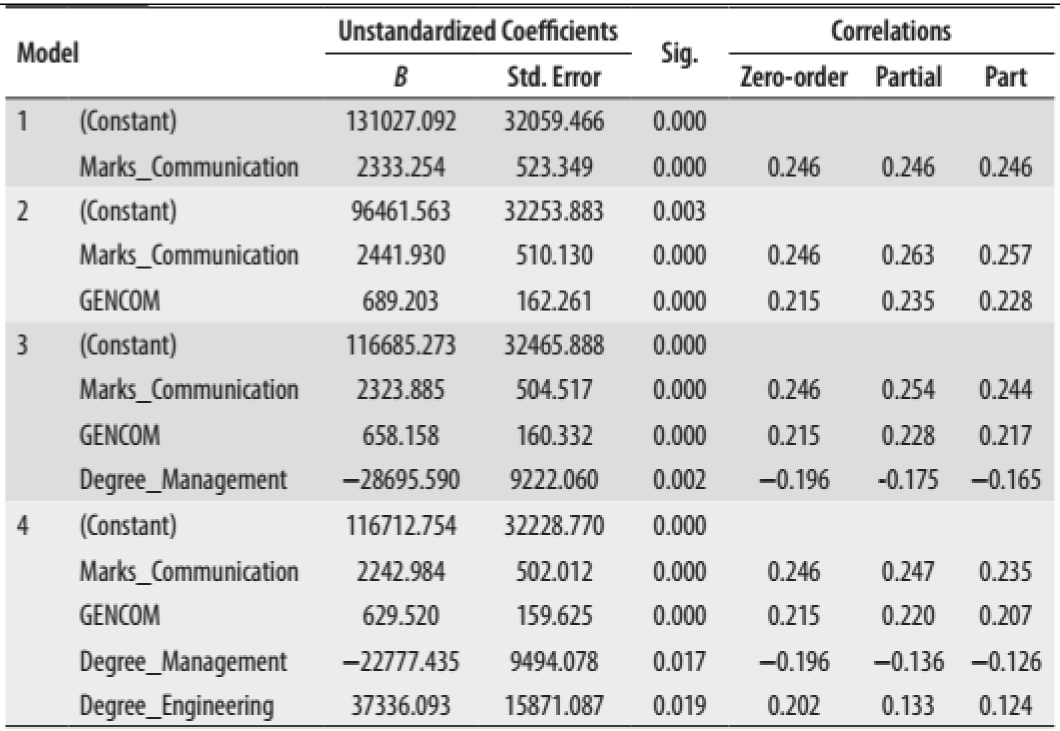
Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.246[a]		0.057	82984.946
2				
3				
4				

Table. 17. Coefficient Values

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis



Model		Unstandardized Coefficients		Sig.	Correlations		
		B	Std. Error		Zero-order	Partial	Part
1	(Constant)	131027.092	32059.466	0.000			
	Marks_Communication	2333.254	523.349	0.000	0.246	0.246	0.246
2	(Constant)	96461.563	32253.883	0.003			
	Marks_Communication	2441.930	510.130	0.000	0.246	0.263	0.257
	GENCOM	689.203	162.261	0.000	0.215	0.235	0.228
3	(Constant)	116685.273	32465.888	0.000			
	Marks_Communication	2323.885	504.517	0.000	0.246	0.254	0.244
	GENCOM	658.158	160.332	0.000	0.215	0.228	0.217
	Degree_Management	−28695.590	9222.060	0.002	−0.196	-0.175	−0.165
4	(Constant)	116712.754	32228.770	0.000			
	Marks_Communication	2242.984	502.012	0.000	0.246	0.247	0.235
	GENCOM	629.520	159.625	0.000	0.215	0.220	0.207
	Degree_Management	−22777.435	9494.078	0.017	−0.196	−0.136	−0.126
	Degree_Engineering	37336.093	15871.087	0.019	0.202	0.133	0.124

16 What is the R-square value at step 2 of the stepwise regression?

17 In Table , GENCOM is the interaction variable between gender and marks in communication. Which of the following statements is true? Clearly state your arguments.

(a) Salary is more sensitive to marks in communication for females than males.

(b) Salary is more sensitive to marks in communication for males than females.

(c) There is no difference between males and females with respect to marks in communication.

(d) Can't say.

18 Data for Questions 7−12 (Courtesy: Professor Trilochan Sastry, IIM Bangalore): An Agro Insurance company wanted to come up with a model and see how the total production of paddy depends on the rainfall. The complication is that the productivity also depends on various factors such as the total acreage under irrigation. The following variables were used to develop the regression models: PROD The total production in thousands of tons (dependent variable) IRR Total irrigated area in thousands of hectares (independent variable) NON Total non-irrigated area

Data Analytics: UE18CS312

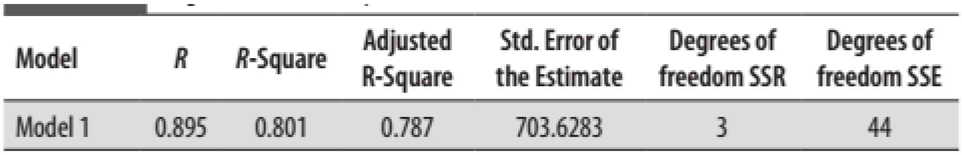
Question Bank

Unit -2 Regression Analysis

<table>
<tr><td></td><td>in thousands of hectares (independent variable) RAIN Total rainfall in millimetres (independent variable)

The SPSS regression model output is given Table 22.

Table 23. Regression Model output


<table><tr><th>Model</th><th><i>R</i></th><th><i>R</i>-Square</th><th>Adjusted R-Square</th><th>Std. Error of the Estimate</th><th>Degrees of freedom SSR</th><th>Degrees of freedom SSE</th></tr><tr><td>Model 1</td><td>0.895</td><td>0.801</td><td>0.787</td><td>703.6283</td><td>3</td><td>44</td></tr></table></td></tr>
<tr><td>19</td><td>If stepwise regression was used to arrive at Table 23, how many variables did SPSS consider? Give reasons.</td></tr>
<tr><td>20</td><td>How many observations were included in the regression? ANOVA corresponding to the MLR model developed is shown in Table 24.</td></tr>
</table>

The Question Bank questions are from the prescribed Text Book

Text Book:

1. “Business Analytics, The Science of Data-Driven Decision Making”, U. Dinesh Kumar, Wiley 2017