Data Analytics: UE18CS312

Question Bank

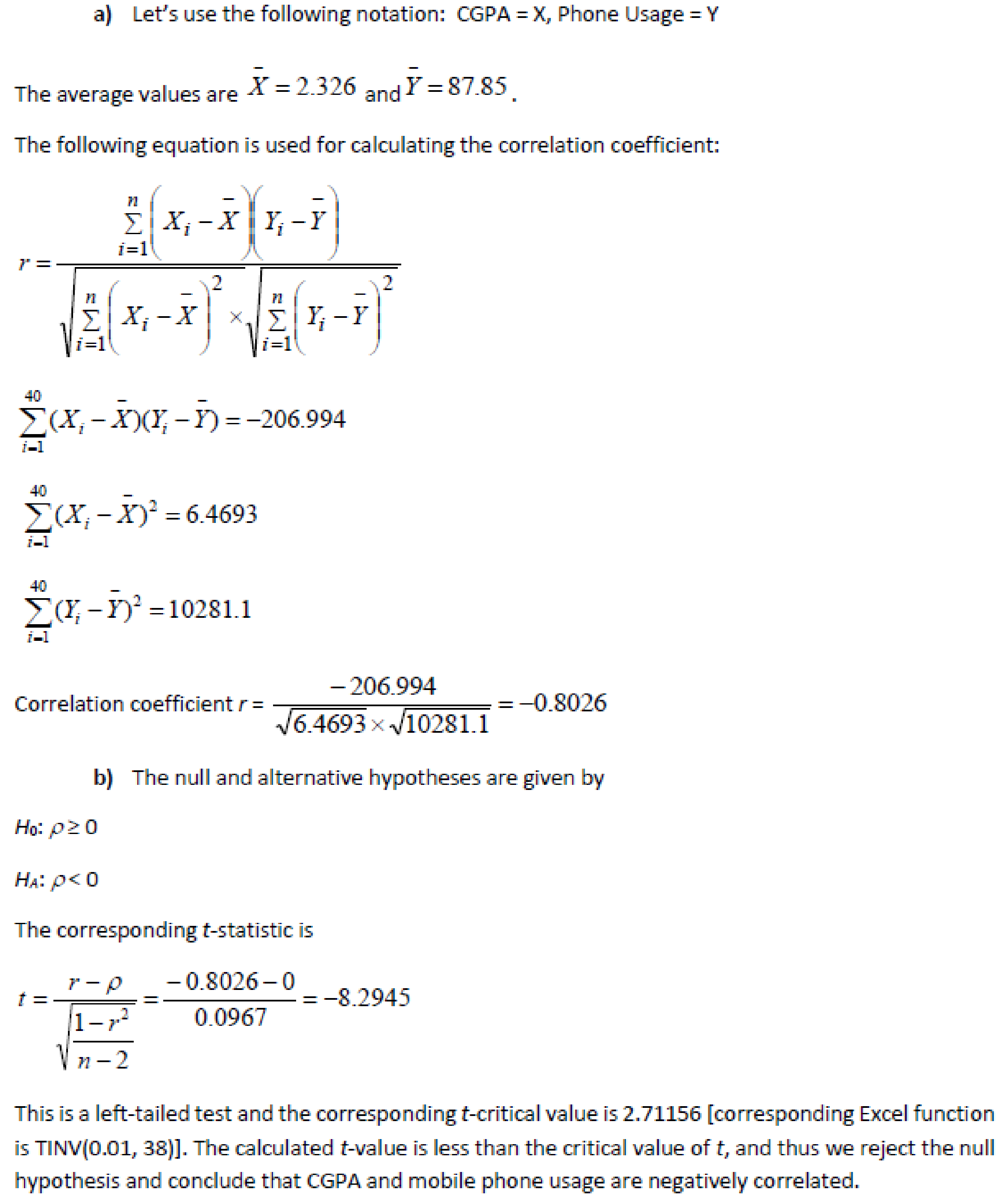
Unit -2 Regression Analysis

<table>
<tr><td>Sl. No</td><td>Questions and Answers</td></tr>
<tr><td>1</td><td>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)
<table>
<tr><td>CGPA</td><td>2.65</td><td>2.25</td><td>1.86</td><td>1.47</td><td>2.10</td><td>1.94</td><td>2.71</td><td>1.83</td><td>2.65</td><td>2.04</td></tr>
<tr><td>Phone Usage</td><td>75</td><td>89</td><td>65</td><td>136</td><td>95</td><td>103</td><td>74</td><td>109</td><td>7</td><td>98</td></tr>
<tr><td>CGPA</td><td>2.54</td><td>2.16</td><td>2.28</td><td>2.47</td><td>2.18</td><td>2.57</td><td>1.97</td><td>2.87</td><td>2.10</td><td>3.28</td></tr>
<tr><td>Phone Usage</td><td>60</td><td>93</td><td>88</td><td>81</td><td>92</td><td>78</td><td>102</td><td>70</td><td>95</td><td>89</td></tr>
<tr><td>CGPA</td><td>2.78</td><td>2.441.87</td><td>2.50</td><td>2.24</td><td>2.01</td><td>2.17</td><td>2.20</td><td>2.05</td><td>1.63</td><td></td></tr>
<tr><td>Phone Usage</td><td>72</td><td>82</td><td>107</td><td>80</td><td>89</td><td>100</td><td>92</td><td>91</td><td>98</td><td>123</td></tr>
<tr><td>CGPA</td><td>2.28</td><td>2.63</td><td>2.86</td><td>2.24</td><td>2.44</td><td>2.69</td><td>2.22</td><td>3.07</td><td>1.77</td><td>3.03</td></tr>
<tr><td>Phone Usage</td><td>88</td><td>76</td><td>70</td><td>89</td><td>82</td><td>74</td><td>90</td><td>65</td><td>113</td><td>66</td></tr>
</table>
</td></tr>
<tr><td>Soln</td><td></td></tr>
</table>

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

<table>
<tr>
<td></td>
<td>


a) Let's use the following notation: CGPA = X, Phone Usage = Y

The average values are $\bar{X} = 2.326$ and $\bar{Y} = 87.85$.

The following equation is used for calculating the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

$$\sum_{i=1}^{40}(X_i - \bar{X})(Y_i - \bar{Y}) = -206.994$$

$$\sum_{i=1}^{40}(X_i - \bar{X})^2 = 6.4693$$

$$\sum_{i=1}^{40}(Y_i - \bar{Y})^2 = 10281.1$$

Correlation coefficient $r = \frac{-206.994}{\sqrt{6.4693} \times \sqrt{10281.1}} = -0.8026$

b) The null and alternative hypotheses are given by

H_0: $\rho \geq 0$

H_A: $\rho < 0$

The corresponding *t*-statistic is

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{-0.8026 - 0}{0.0967} = -8.2945$$

This is a left-tailed test and the corresponding *t*-critical value is 2.71156 [corresponding Excel function is TINV(0.01, 38)]. The calculated *t*-value is less than the critical value of *t*, and thus we reject the null hypothesis and conclude that CGPA and mobile phone usage are negatively correlated.
</td>
</tr>
<tr>
<td>2</td>
<td>Table 3. provides ranking of Indian states based on corruption and Table 4. provides ranking based on literacy rate.</td>
</tr>
</table>

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

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.

Soln.

The Spearman rank correlation is given by

$$r = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6 \times 992}{16(16^2-1)} = -0.4588$$

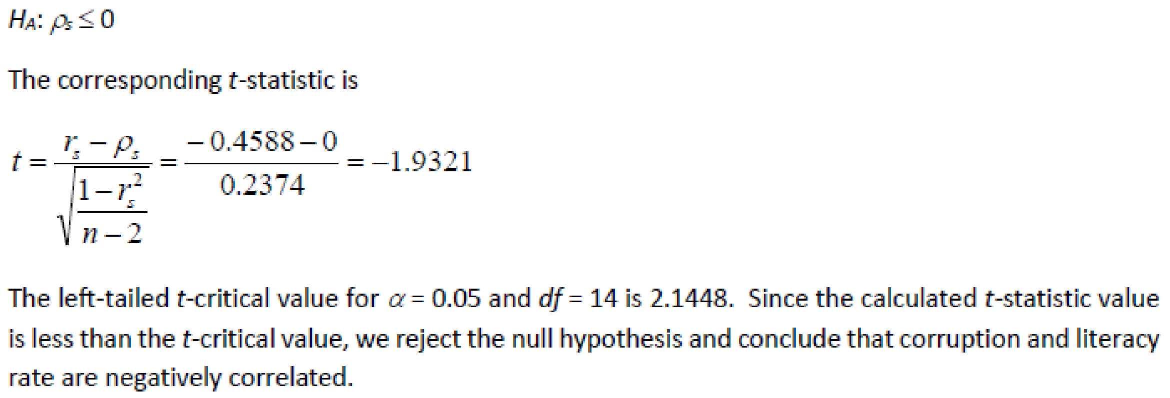
The null and alternative hypotheses are

H_0: $\rho_s > 0$

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis



H_A: $\rho_s \leq 0$

The corresponding *t*-statistic is

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1 - r_s^2}{n-2}}} = \frac{-0.4588 - 0}{0.2374} = -1.9321$$

The left-tailed *t*-critical value for $\alpha = 0.05$ and $df = 14$ is 2.1448. Since the calculated *t*-statistic value is less than the *t*-critical value, we reject the null hypothesis and conclude that corruption and literacy rate are negatively correlated.

3

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

Soln

The contingency table for the data is given below.

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Contingency Table

	Y			
		0	1	Total
X	0	10	6	16
	1	9	25	34
Total		19	31	50

From Contingency Table, we have

$N_{00} = 10$, $N_{01} = 6$, $N_{10} = 9$, $N_{11} = 25$, $N_{X0} = 16$, $N_{X1} = 34$, $N_{Y0} = 19$, and $N_{Y1} = 31$

The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}} = \frac{25\times 10 - 9\times 6}{\sqrt{16\times 34\times 19\times 31}} = 0.3462$$

4

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).

Soln

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

$$F = \frac{SSR\times(n-2)}{SSE}$$

We have to prove that

$$F = \frac{R^2}{(1-R^2)/(n-2)}$$

Lets start with

$$F = \frac{SSR\times(n-2)}{SSE} = \frac{(SST-SSE)\times(n-2)}{SSE} = \left(\frac{SST}{SSE} - 1\right)\times(n-2)$$

$$= \left(\frac{1}{(1-R^2)} - 1\right)\times(n-2) = \left(\frac{1-1+R^2}{(1-R^2)}\right)\times(n-2) = \left(\frac{R^2}{(1-R^2)}\right)\times(n-2) = \left(\frac{R^2}{(1-R^2)/(n-2)}\right)$$

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

5 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

	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: $Y = \beta 0 + \beta 1 \times \text{Carat}$

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

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

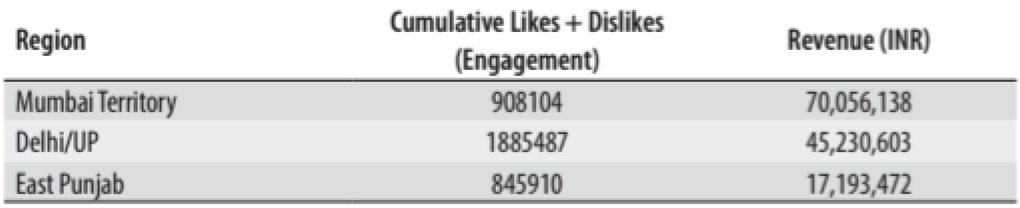
	a) Based on the values in Table 9.15 and 9.16 we calculate F-statistic value as F = $t^2 = (-63.439)^2 = 4024.5067$. N = 6000 Also we know $F = \frac{R^2}{(1-R^2)/(n-2)}$ $\frac{1}{R^2} = \frac{(n-2)}{F} + 1 = \frac{(6000-2)}{4024.5067} + 1 = 2.4904$ $R^2 = 0.4015$ That means the model is explaining 40.15% of the variation in the value of Y. Also, from the coefficient and standard error of Carat t-value = $\frac{18381.261}{141.733} = 129.6893$ and the corresponding p-value is 0.000. Hence we can say the beta coefficient for Carat is statistically significant. From fig 9.14 we could find a funnel shape (non-constant variance of residuals) thus indicating heteroscedasticity. So, this model is not significant. **b)** From the model 1, we have Y = -12738.581 + 18381.261 × Carat , for every one-carat increase in diamond weight the price of diamond increases by 18381.261. Hence, we can conclude that the price of the diamond increases by at least 10,000 for every one-carat increase in the diamond weight (at significant value 0.05). **c)** The regression model 2 is given by ln(Y) = 7.265 + 1.375 × Carat
6	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

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Soln

a) The model outputs for the regression equation are provided below

Regression Statistics	
Multiple R	0.533850998
R Square	0.284996888
Adjusted R Square	0.219996605
Standard Error	17297831.51
Observations	13

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6245081.969	6936674.619	0.900299	0.387247
Engagement	18.49915255	8.834651015	2.093931	0.06023

$Y = 6245081.969 + 18.4991 \times$ social media engagement

The box office collection (Y) does not have a statistically significant relationship with the social media engagement (X). As from the Microsoft Excel output we can see that t-statistic value is 2.0930 for which the p-value is 0.06 which is not less than alpha = 0.05

Note: There is no strict rule for selecting the cut off for alpha, it depends on the context of the business problem. But as in this model we have chosen the value of alpha as 0.05 any p-value that is greater than alpha we will reject.

b) The 95% confidence interval for the average value of the response variable is given by

$$\hat{Y}_i \pm t_{\alpha/2, n-2} \times S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$\hat{Y}_i = 6245081.969 + 18.4991 \times 20000 = 6615065.02$; $t_{\alpha/2,\,11} = 2.5931$, $S_e = 17297831.51$

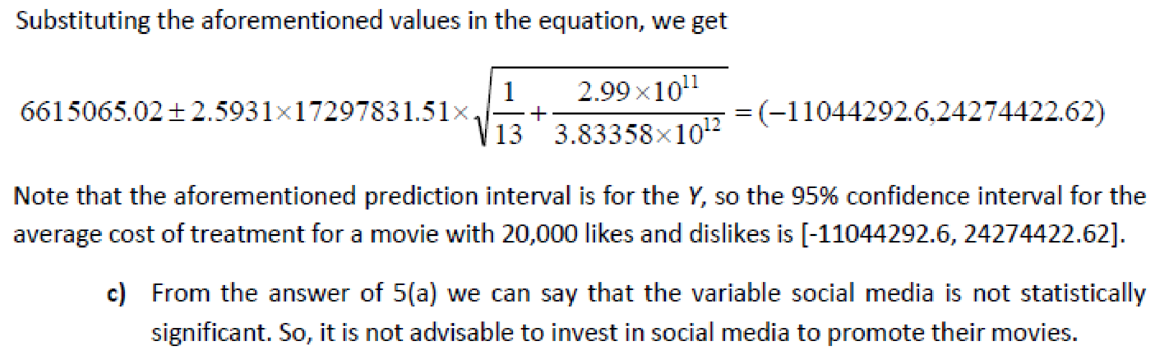
$\bar{X} = 567094.4615$, $(X_i - \bar{X})^2 = (20000 - 567094.4615)^2 = 2.99 \times 10^{11}$,

$\sum_{i=1}^{13}(X_i - \bar{X})^2 = 3.83358 \times 10^{12}$

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis



Substituting the aforementioned values in the equation, we get

$$6615065.02 \pm 2.5931\times 17297831.51\times\sqrt{\frac{1}{13}+\frac{2.99\times 10^{11}}{3.83358\times 10^{12}}} = (-11044292.6, 24274422.62)$$

Note that the aforementioned prediction interval is for the *Y*, so the 95% confidence interval for the average cost of treatment for a movie with 20,000 likes and dislikes is [-11044292.6, 24274422.62].

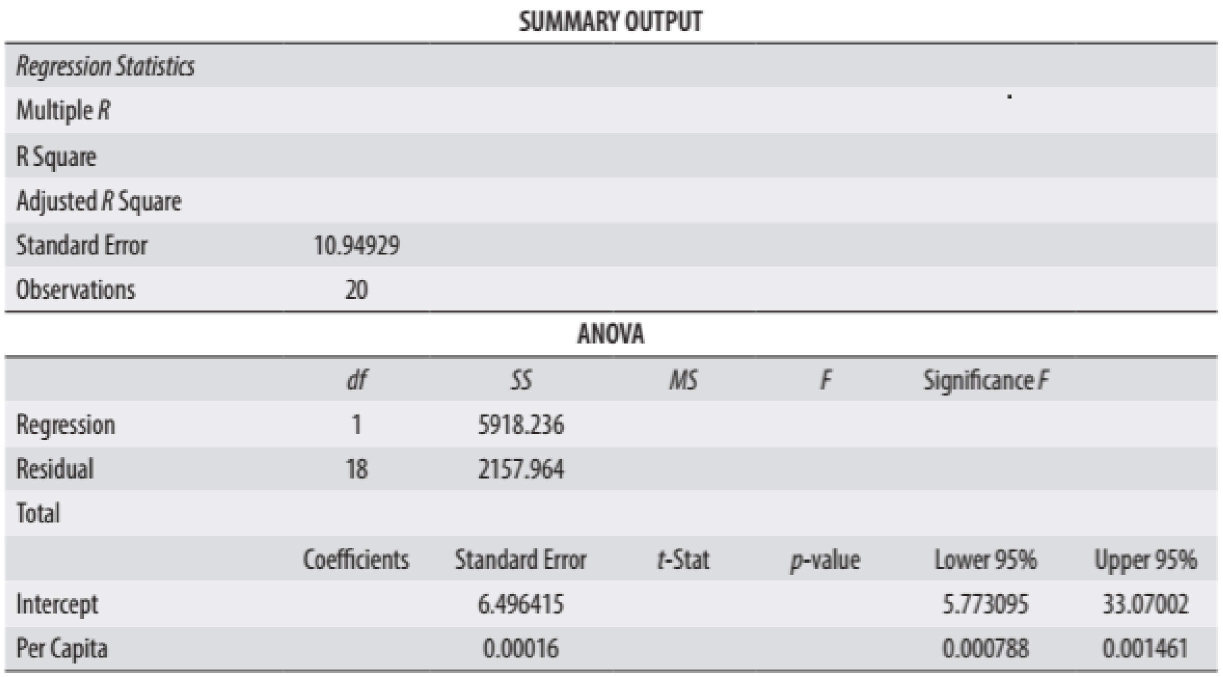
c) From the answer of 5(a) we can say that the variable social media is not statistically significant. So, it is not advisable to invest in social media to promote their movies.

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)

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.

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Soln	**a)** $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\left(\hat{Y}_i - Y_i\right)^2}{\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{2157.964}{(5918.236 + 2157.964)} = 0.7382$ Hence from the R^2 value we can say that 73.82% of the corruption index is explained by per capita. **b)** As we know $F = \frac{MSR}{MSE} = \frac{MSR}{SSE/(n-2)} = 49.3651 = t^2$ Hence, t = 7.0260 and the corresponding p-value is 5.44E-07 Also, $t = \frac{\hat{\beta}}{S_e(\hat{\beta})} = 7.0260$ From table 9.22, S_e = 0.00016 hence $\hat{\beta} = 7.0260 \times 0.00016 = 0.001124$ So we can say that for every one dollar increase in per capita the value of corruption perception index will increase by a factor of 0.001124

Data Analytics: UE18CS312

Question Bank

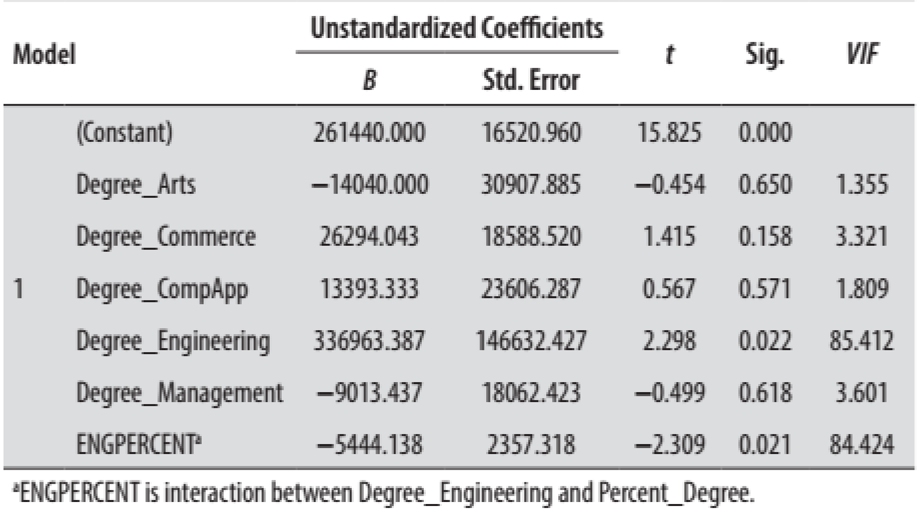
Unit -2 Regression Analysis

	c) The t-critical value for alpha = 0.01 is 2.8609, and as t-critical value is less than the t-statistic value, we can conclude that there is a statistically significant relationship between corruption perception index and per capita at alpha = 0.01 **d)**The 95% confidence interval for the average value of the response variable is given by $\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$ **e)** The co-efficient of per-capita (β) = 0.001124 Given, per-capita of country is $30,000. $H_0 : Y \geq 50$ $H_a : Y < 50$ $Z = \frac{33.72 - 50}{10.94929} = -1.4868$ $Z_{critical}$ = 2.861 (from table for α =0.01) Since $Z_{stat} < Z_{critical}$, we reject null hypothesis. So, we can conclude that the corruption perception index of this country is less than 50 when per-capita is $30,000. **f)** The correlation coefficient for the model shown in Table 9.21 is given as -0.4639. Hence we can conclude that corruption perception index and per capita are negatively correlated.
8	1. 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**

Data Analytics: UE18CS312

Question Bank

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.

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Soln

1. We calculate the t value from the corresponding beta coefficients and standard error in the following table using the formula $t = \left(\frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)}\right)$ and the corresponding p-values:

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	*t*	Sig.
1	(Constant)	261440.000	16589.992		15.75889	2.23E-41
	Degree_Arts	−14040.000	31037.032		-0.45236	0.651328
	Degree_Commerce	26294.043	18666.192		1.40864	0.159955
	Degree_CompApp	13393.333	23704.925		0.56500	0.572486
	Degree_Engineering	63760.000	22462.955		2.83845	0.004836
	Degree_Management	−9013.437	18137.895		-0.49694	0.619588

[a] Dependent Variable: Salary

From the t- value and its corresponding p-value we conclude that beta coefficients only for Degree_Engineering and Degree_science are statistically significant.

Hence we can say that as only these two degrees have statistically significant relationship on Salary, the dean should give more importance to these two degree disciplines over the others.

2. The variable Degree_Commerce is not significant as the p-value for the variable is 0.1599 which is more than 0.05. On the other hand Degree_Science is a significant variable with a significance level of 2.23E-41. This implies that there is no significant average difference in the average salary earned by a student with science degree and commerce degree.
3. H0: Salary of Engineering students – Salary of Science students <= 25,000

   H1: Salary of engineering students – Salary of Science students > 25,000

Using the coefficients from the model, we can compute if the claim made is true or not.

Salary of engineering students,

Y = 261440 + 63760 * 1

Y = 325200

Salary of science students,

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Y = 261440 + 63760 * 0

Y = 261440

Average difference in salary of engineering students and science students

= 325200 – 261440 = 63760

Std. error of estimate = 22462.955

Alpha = 0.05

Df = (n-1) = 306

$T_{critical}$= 1.9677

t-statistic = (63760 – 25000)/22462.955 = 1.7255

For-right tailed test, the $t_{critical}$ value is more than the t-statistic hence we retain the null hypotheses and conclude that the average difference in salary of engineering students and science students is not more than INR 25,000.

4. As from the Table 10.44 we can see that VIF for ENGPERCENT is 84.424 which is too high. The threshold value for VIF is 4 (a few authors suggest 10). Hence, in this model we need to assess the impact of multi-collinearity. Because impact of multi-collinearity is that it can change the sign of the regression coefficient (for example, instead of positive, the model may have negative regression coefficient for the predictor or vice versa, so that can be one explanation in this case.)

5. R^2 at step-2 = R^2 at step -2 + (part-correlation)^2

= $(0.246)^2 + (0.228)^2 = 0.1125$

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

The regression equation for model 2 is given by

Y = 96461.563 + 2241.930×Marks_communication + 689.203×GENCOM

Above equation can be written as

For Female (Gender = 0)

Y =96461.563 + 2241.930×Marks_communication

For Male (Gender = 1)

Y = 96461.563 + (2241.930+689.203)×Marks_communication

That is, the change in salary for female when Marks_Communication increases by one unit is 2241.930 and for male is 2931.133. Hence we conclude that Salary is more sensitive to marks in communication for males than females.

The Question Bank questions are from the prescribed Text Book

Data Analytics: UE18CS312

Question Bank

Unit -2 Regression Analysis

Text Book:

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