Advanced Social Statistics

Homework Assignment 1 SOC 586

5

A researcher interested in understanding factors predicting math performance by students estimated a regression model and reported her results below.

Regression of student's current math score, N=407, R2 =0.5501

Variable	(Parameter) Estimate	Standard Error	t-ratio	*, **, or***
Constant	-0.0258	0.0900		
Male	0.0693	0.0615		
Drugs	-0.9463	0.4372		
Enrollment	0.0726	0.0773		
Math score 4 years ago	0.6374	0.0372		
Socioeconomic status	0.1348	0.0393		
Urban	0.0005	0.0009		

1. Using the information given above, fill in the blanks in the last two columns. Use *, **, and *** to indicate statistical significance at the 5%, 1% and 0.1%, two-tailed test, respectively.
2. Assess the overall model fit of the regression analysis above by an F-test of the R2, and report and interpret the false-rejection probability.
3. For an unrelated regression analysis, the researcher obtained:

$$\hat{y} = 8 + 2A - \frac{1}{4}B - \frac{1}{4}AB$$

Where A is a dummy grouping variable (0,1), B is a continuous variable (ranging from 0 to 20), and AB is their interaction term. Calculate the predicted values of y for the two groups (i.e., A=0 and A=1) for B=0, 4, 8, 12, 16 and 20, respectively. Present your results in a table or graph.

Solution to Question No. 1

From the information provided, we can see that there are six factors or ind ndent variables that the researcher is interested to use in her stu in order to predict math performance by students out of a total sample size of N = 407.

∵ the estimate of the parameter, i.e; the slope(b) of the regression line for each of the independent variables is already provided in the table including their corresponding standard error (S_b), so let us go ahead and calculate the t-statistic for each of the variables using the formula: $t = \frac{b - \beta}{S_b}$ where: b = Sample regression slope coefficient [of $b_1, b_2, b_3 \cdots b_6$]

β = Hypothesized slope [$\beta_1, \beta_2, \beta_3 \cdots$]

S_b - Estimator of the standard error of the slope

In the given formula, the hypothesized slope is the parameter which we don't know in reality. However, according to our assumption of the Null hypothesis (H_0), which is $H_0 : \beta = 0$, applicable to each of the variables given, the t-statistic formula will be simplified to : $t = \frac{b}{S_b}$ $[\because \beta = 0]$

Now, let us determine the degrees of freedom (df) that will be used to find the t-critical values from the standardized t-table

In this case, $df = N - k - 1$, where N is the total number of samples taken in our experiment (which is 407) and k is the number of explanatory or predictor variables in the model (which is 6).

$\therefore$ $df = 407 - 6 - 1 = 407 - 7 = 400$

Another logic for ascertaining the df in t-test during regression analysis is to understand that df come from the denominator, which is based on the residual Sum of Squares (RSS) - Since, there are $N-k-1$ degrees of freedom in the RSS.

So, Let's go ahead now and complete the table, so that we could compare the t-ratio (or t-statistic) for each of 6 independent variables as well as the intercept (i.e. Constant) with their corresponding t-critical values (derived from the t-table) using a two-tailed test at 5% (.05), 1% (.01) and 0.1% (.001) respectively.

The results are provided in the next page. For simplicity, all the independent variables are labelled as $x_1, x_2, x_3 \ldots$ upto x_6.

Table : 1

Variable	(Parameter) Estimate (b)	Standard error (s_b)	t-ratio	* t-critical 5%	** t-critical 1%	*** t-critical 0.1%
Constant	-0.0258 (b_0)	0.0900	-0.2867	±1.9659	±2.5882	±3.3150
Male (x_1)	0.0693 (b_1)	0.0615	1.1268	±1.9659	±2.5882	±3.3150
Drugs (x_2)	-0.9463 (b_2)	0.4372	-2.1645	*±1.9659	±2.5882	±3.3150
Enrollment (x_3)	0.0726 (b_3)	0.0773	0.9392	±1.9659	±2.5882	±3.3150
Math score 4 years ago (x_4)	0.6374 (b_4)	0.0372	17.1344	*±1.9659	*±2.5882	*±3.3150
Socio economic status (x_5)	0.1348 (b_5)	0.0393	3.4300	*±1.9659	*±2.5882	*±3.3150
Urban (x_6)	0.0005 (b_6)	0.0009	0.5556	±1.9659	±2.5882	±3.3150

Please note in the Table : 1 above that the t-critical value at 5% level of significance is the same for all the variables as obtained from the t-table. The same is true at 1% and 0.1% significance level respectively. The cells in Table 1 that are highlighted in green are the ones that are found statistically significant for that variable.

Now, based on the t-statistic (t-ratio) and t-critical, let us test the Statistical Significance for each of the variables ($x_1, x_2, x_3 \cdots$) in the study.

The hypothesis statements to test the significance of a particular regression coefficient, β_{x_i}, are :

$H_0 : \beta_{x_i} = 0$

$H_1 : \beta_{x_i} \neq 0$ [being two-tailed]

where, x_i represents $x_1, x_2, x_3 \cdots x_6$ and β_{x_i} is the slope parameter for each of the variables $x_1, x_2, x_3 \cdots$

In the given problem, ∵ we do not know β_{x_i} so we replace it with its estimate (i.e; $b_1, b_2, b_3 \cdots b_6$)

contd to ... Page 4

For :

Male (x_1)

$t_{0.05,400} = \pm 1.9659$	$t_{0.01,400} = \pm 2.5882$	$t_{0.001,400} = \pm 3.3150$
$t_{observed} = 1.1268$	$t_{obs} = 1.1268$	$t_{obs} = 1.1268$

$\therefore t_{obs} < t_{crit}$, so we failed to reject H_0 at 5% level of significanc
$\therefore$ The test is not statistically significant at 5% level

Similarly, for both 1% and 0.1% level, the test is also statistically no Significant since we could see that $t_{obs} < t_{crit}$. $\therefore$ we failed to reject H_0.

Drugs (x_2)

In a similar vein as calculated and explained above for variable X_1, we can see from Table 1, that for Drugs (x_2), the test is statistically significant at 5% [$t_{obs} = -2.1645 > t_{crit} = \pm 1.9659$] but not at 1% and 0.1% level. However, the effect of drugs seems to have a negative influence by decreasing math performance of students as shown by the minus(-) sign of slope coefficient.

Enrollment (x_3)

Here, $t_{obs} = 0.9392$ which is far less than t_{crit} at all levels (5%, 1% and 0.1%) of Significance. So, we failed to reject H_0. The test result is therefore statistically not significant.

Math Score 4 years ago (x_4)

Here, $t_{obs} = 17.1344$ which is far above the critical values at 5%, 1% and 0.1% level of significance. As such, we reject H_0 and conclude that the test is statistically significant at all the given levels. This particular variable (x_4) has the most

Positive effect on the regression model.

Socio-economic status (x_5)

Here, $t_{obs} = 3.4300$ which surpasses all the t_{crit} values at 5%, 1% and 0.1% level of Significance. Therefore, we reject Ho and conclude that the test is statistically significant for this variable at the given levels.

Urban (x_6)

In this case, $t_{obs} = 0.5556$ which is far below the threshold critical values of t at 5%, 1% and 0.1% level of Significance. Therefore, we failed to reject Ho. The test is statistically not significant for this variable (x_6).

Solution to Question No.2

Let us first of all calculate the F-statistic of the given regression model based on the information provided, where $R^2 = 0.5501$ & $N = 407$.

$$F_{stat} = \frac{R^2/k}{(1-R^2)(n-k-1)} = \frac{0.5501/6}{(1-0.5501)/407-6-1} = \frac{0.0916833}{0.4499/400}$$

$$= \frac{0.0916833}{0.00112475} = 81.5144$$

Now, from the F-distribution table we find that

$F_{crit(6, 400)} = 2.1212$ at 5% Significance level

$F_{crit(6, 400)} = 2.8472$ at 1% Significance level

$F_{crit(6, 400)} = 3.8305$ at 0.1% Significance level

The F-test for overall significance has the following two hypotheses:

H_0 : the model with no independent variables fits the data as well as the model

H_1 : the given model fits the data better than the intercept-only model.

Based on the calculations done earlier in page 5, we can see that

$F_{Statistic}$ or $F_{observed} = 81.5144$ is considerably $> F_{critical}$ values at all the given level of significance (i.e; 5%, 1% & 0.1%)

Thus, at this point, we can reject H_0 and conclude that the overall F-test is statistically significant, which means that the independent variables in our given model improve the fit.

Explanation and Interpretation :

At this stage, based on our calculation and observation, it is important to note that the results of our F-test and t-tests do not concur or agree with each other entirely. This can happen in statistical analysis because F-test is considered an overall significance test and therefore assesses all of the coefficients jointly, and in our case find that the coefficients are statistically significant. On the other hand, the t-test examines for each coefficient individually and therefore can fail to find significance. For example, if we refer back to table 1 in page 3, the effect of the independent variables : Male (x_1), Enrollment (x_3) and Urban (x_6) in predicting math performance by students (the dependent variable) is found to be not statistically significant while conducting the t-test at all the given α levels (5%, 1% and 0.1%). Therefore, in such given circumstances

if we depend only on the results of the F-test, without additionally carrying out t-tests of our variables of interest, we may run the risk of false-rejection probability, i.e., rejecting the null hypothesis (H_0) when in reality it is true.

Assessing overall model fit:

∵ Regression is essentially drawing a line of best-fit across all observations, so as R^2 decreases, the actual relationship between the observed values (y_i) and the line of best-fit ($\hat{y}$) becomes weaker. In our regression model, the given value of $R^2 = 0.5501$ which means that 55.01% of the variation in the dependent variable (Y) is accounted for or explained by the independent variables. This is not a very strong model fit, but reasonably good. Since, the value of R^2 (coefficient of determination) varies from 0 to 1, so when $R^2 = 1$, we have a perfect linear relationship, and as R^2 becomes smaller, the linear relationship gets weaker.

According to the results of the t-tests as given in the table-1 in page 3, we figured out that some of the independent variables currently included in the model, such as Male (x_1), Enrollment (x_3) and urban (x_6) are found to be statistically non-significant, and therefore, do not contribute in predicting the dependent variable. So, such variables which are not significant could be excluded from the model to result in a better fit. Moreover, the unnecessary inclusion of weak independent variables may account for more variance, while adding nothing to the model.

Solution to Question No. 3

In regression analysis, we use dummy variables to represent categorical information, especially when one or more of the variables is categorial, often qualitative data, such as [Yes or No answer] or geographical directions [East, West, North, South] and so forth.

So, in the given problem, let us go ahead and calculate the predicted values of y for the two groups as per the provided regression equation : $\hat{y} = 8 + 2A - \frac{1}{4}B - \frac{1}{4}AB$.

When $\boxed{A = 0}$, $B = 0$

$$\hat{y} = 8 + 2(0) - \frac{1}{4}(0) - \frac{1}{4}(0)(0) = 8 + 0 - 0 - 0 = 8$$

$B = 4$

$$\hat{y} = 8 + 2(0) - \frac{1}{4}(4) - \frac{1}{4}(0)(4) = 8 + 0 - 1 - 0 = 8 - 1 = 7$$

$B = 8$

$$\hat{y} = 8 + 2(0) - \frac{1}{4}(8) - \frac{1}{4}(0)(8) = 8 + 0 - 2 - 0 = 8 - 2 = 6$$

$B = 12$

$$\hat{y} = 8 + 2(0) - \frac{1}{4}(12) - \frac{1}{4}(0)(12) = 8 + 0 - 3 - 0 = 8 - 3 = 5$$

$B = 16$

$$\hat{y} = 8 + 2(0) - \frac{1}{4}(16) - \frac{1}{4}(0)(16) = 8 + 0 - 4 - 0 = 8 - 4 = 4$$

$B = 20$

$$\hat{y} = 8 + 2(0) - \frac{1}{4}(20) - \frac{1}{4}(0)(20) = 8 + 0 - 5 - 0 = 8 - 5 = 3$$

When $\boxed{A = 1}$, $B = 0$

$$\hat{y} = 8 + 2(1) - \frac{1}{4}(0) - \frac{1}{4}(1)(0) = 8+2-0-0 = 10$$

$B = 4$

$$\hat{y} = 8+2(1) - \frac{1}{\cancel{4}_1}(\overset{1}{\cancel{4}}) - \frac{1}{\cancel{4}_1}(1)(\overset{1}{\cancel{4}}) = 8+2-1-1 = 10-2 = 8$$

$B = 8$

$$\hat{y} = 8+2(1) - \frac{1}{\cancel{4}_1}(\overset{2}{\cancel{8}}) - \frac{1}{\cancel{4}_1}(1)(\overset{2}{\cancel{8}}) = 8+2-2-2 = 10-4 = 6$$

$B = 12$

$$\hat{y} = 8+2(1) - \frac{1}{\cancel{4}_1}(\overset{3}{\cancel{12}}) - \frac{1}{\cancel{4}_1}(1)(\overset{3}{\cancel{12}}) = 8+2-3-3 = 10-6 = 4$$

$B = 16$

$$\hat{y} = 8+2(1) - \frac{1}{\cancel{4}_1}(\overset{4}{\cancel{16}}) - \frac{1}{\cancel{4}_1}(1)(\overset{4}{\cancel{16}}) = 8+2-4-4 = 10-8 = 2$$

$B = 20$

$$\hat{y} = 8+2(1) - \frac{1}{\cancel{4}_1}(\overset{5}{\cancel{20}}) - \frac{1}{\cancel{4}_1}(1)(\overset{5}{\cancel{20}}) = 8+2-5-5 = 10-10 = 0$$

The results are summarized and presented systematically in Table: 2 and chart 1 and chart 2 in page 10.

Predicted values of y for the two groups is presented below where A is a dummy grouping variable with two values (0 and 1) and B is a continuous variable.

Table : 2

Group 1 (A=0)				Group 2 (A=1)		
Variable B (X1)	Variable A (X2)	Predicted value of y ($\hat{y}$)		Variable B (X1)	Variable A (X2)	Predicted value of y ($\hat{y}$)
0	0	8		0	1	10
4	0	7		4	1	8
8	0	6		8	1	6
12	0	5		12	1	4
16	0	4		16	1	2
20	0	3		20	1	0

Chart: 1

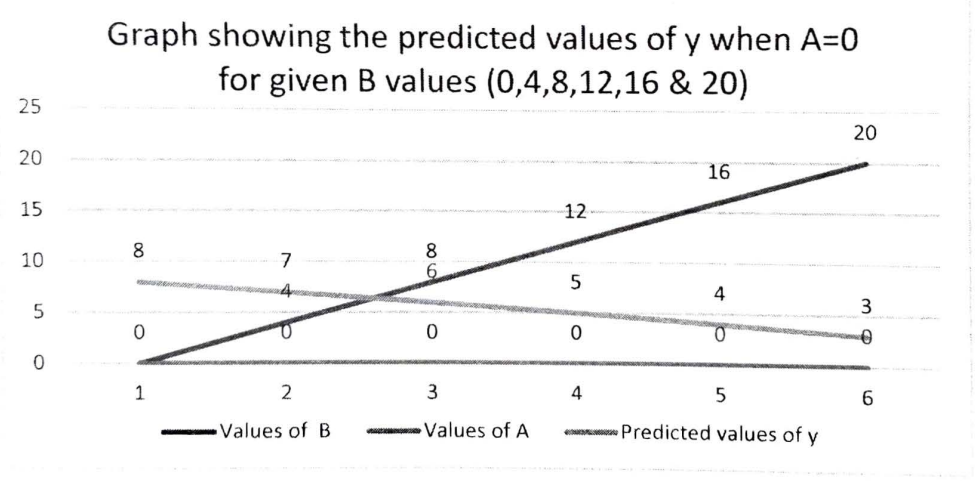

Chart: 2

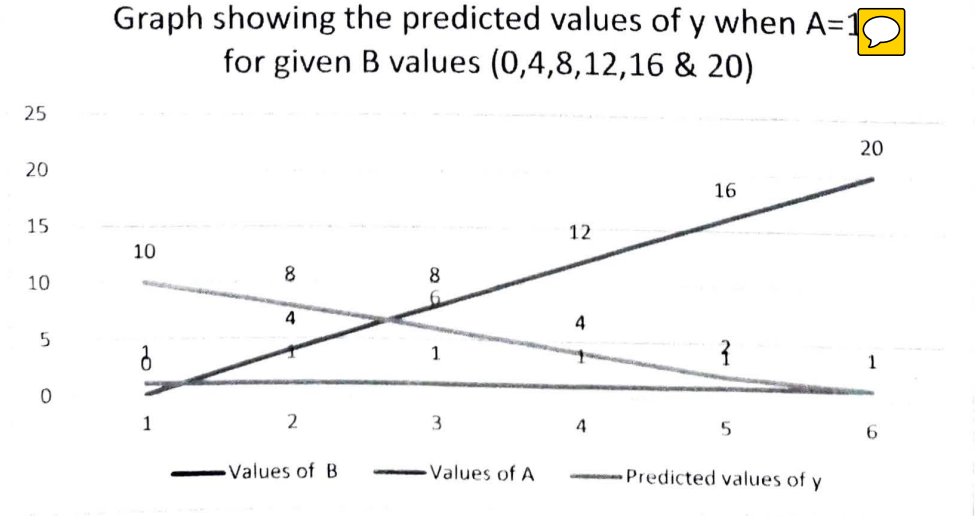

TABLE A.2

t Distribution: Critical Values of *t*

Degrees of freedom	*Two-tailed test:* / *One-tailed test:*	*Significance level* 10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291