CA2- STATISTIC

Student Number: x16149645

1 Multiple Regression

Determine the overall score of student using the multiple regression

The given dataset depicts the school propel report of understudies.

The stages of predictor variable and response elements:

In this study, three stages of measurement are used to analyze the complete performance of student. The response variable is Overall score and the two predictor variables are Environment category score and Performance category score.

Expectations for multiple regression model:

By Using values of two or more variables, we can calculate the value of the response variable.

1 ORDINAL Variable: Response element should have scaled on ORDINANL SCALE for that we need to carry out the ordinal regression.

2 More than two predictor factors should be open while figuring the various backslide which can either be steady or ordinal

3 By utilizing DURBIN-WATSON measurement we can figure the autonomy of the residual.

4 At the time of Calculating multiple regression, there must be a linear relationship among the predictor variable and response variables. To represent the liner relationship among the predictor variable and response variable, scatter plots are used.

5 Homoscedasticity: It tells that how information is best fit in to the model.

6 The predictor variable and response variable should be related to each other and furthermore predictor variable should also be related to each other this shows the multicollinearity of the elements.

Determine the collinearity among predictor variables.

Correlations

		OVERALL SCORE	ENVIRONMENT CATEGORY SCORE	PERFORMANCE CATEGORY SCORE
Pearson Correlation	OVERALL SCORE	1.000	.449	.698
	ENVIRONMENT CATEGORY SCORE	.449	1.000	.277
	PERFORMANCE CATEGORY SCORE	.698	.277	1.000
Sig. (1-tailed)	OVERALL SCORE	.	.000	.000
	ENVIRONMENT CATEGORY SCORE	.000	.	.004
	PERFORMANCE CATEGORY SCORE	.000	.004	.
N	OVERALL SCORE	92	92	92
	ENVIRONMENT CATEGORY SCORE	92	92	92
	PERFORMANCE CATEGORY SCORE	92	92	92

Fig 1.1 Correlation

The Above table shows the correlation among the two predictors variables. By using Pearson's correlation, correlation among variables can be defined. Environment category score correlation with overall score is 0.449 as well as the correlation among performance category score and overall score is 0.698.

The two-predictor variable are also corelated to each other i.e. 0.277. The correlation among all the variables must be less than 0.80, then it satisfies the data in the table is linear and not multicollinear.

Deciding in what way the model fit:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.747[a]	.558	.548	9.877672148	2.214

a. Predictors: (Constant), PERFORMANCE CATEGORY SCORE, ENVIRONMENT CATEGORY SCORE

b. Dependent Variable: OVERALL SCORE

Fig:1.2 Model Summary

The above table show the model summary for multiple regression model in which in what way data can be fit is described using 3 values R, R square and adjusted R square.

To forecast the quality of the response value R is considered. R represents the multiple correlation coefficient (R= 0.747). The second column represent the value of R square which represent the coefficient of correlation. (R square =0.558) that means 55.8%

response variable can be explained by predictor variable. To report data appropriately we use adjusted r square (Adjusted R square = 0.548)

ANOVA for statistical significance:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10976.523	2	5488.262	56.250	.000[b]
	Residual	8683.588	89	97.568		
	Total	19660.112	91			

a. Dependent Variable: OVERALL SCORE

b. Predictors: (Constant), PERFORMANCE CATEGORY SCORE, ENVIRONMENT CATEGORY SCORE

Fig 1.3 ANOVA for statistical significance

To determine the overall regression prototype F- ratio F [(2,89) =56.250] is used from the ANOVA table. Significance values i.e. $p<0.05$ ($0.000016<0.05$) tells that data is significantly fits into the multiple regression model.

Evaluated exhibit coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	14.568	4.759		3.061	.003					
	ENVIRONMENT CATEGORY SCORE	1.582	.418	.277	3.780	.000	.449	.372	.266	.924	1.083
	PERFORMANCE CATEGORY SCORE	2.052	.242	.621	8.478	.000	.698	.668	.597	.924	1.083

a. Dependent Variable: OVERALL SCORE

Fig 1.4 Coefficients

The above figure shows the coefficients for best fit model. We can predict the multiple regression equation from the table. First column indicates the unstandardized coefficient for two predictor variables i.e. Environment category score and second is Performance category score. by using that, calculate the best fit model equation. Equation for model is overall score is equal to 1.582 times the first predictor variable (Environment category score) plus 2.052 times the second predictor variable (performance category score) minus the error (14.568). using this model, we can fit the data into multiple regression model.

The 4th column and 5th column indicated the significance among each predictor variable [the significance should be less than 0.05 (P<0.05)] i.e. Environment category score is 0.000283 and another variable i.e. Performance category score is 0.000004 that means the data is statistically significant

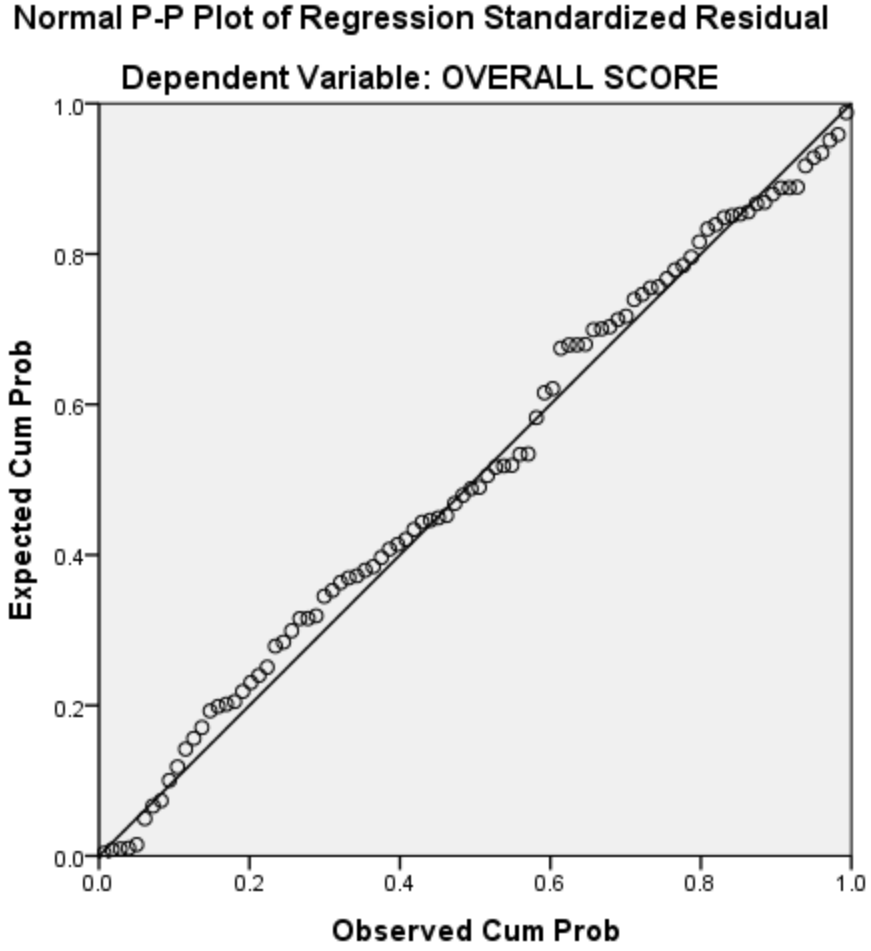

Fig 1.5 Linear scatter plot

The above figure shows the scatter linear plot or P-P plot of regression residual.

Result for multiple regression model:

The overall score is calculated from environment category score and performance category score. By Analysis of variance, we can predict whether data is statically significant or not with the help of F statistic F(2,89) =56.250,P=0.000 (P<0.05).This shows that the data which is taken is statically significant

2 Logistic regression

Binary logistic regression for probability of ethnicity depend on count and rank

dependent and independent elements:

In this analysis, one dependent variable and 2 independent variables are used to calculate the binary logistic regression.

Assumptions taken for binary logistic regression:

Subordinate variable must be measured on dichotomous scale. For instance, gender (“male” and “female”) or numbers (“0” and “1”) etc. In the event that a response scale isn't measured on the dichotomous scale and on the off chance that it is measured on a constant scale then multiple regression should be performed. On the off chance that the dependent variable is measured on an ordinal scale then it should be measured by ordinal regression technique.

The free factor can be steady or obvious.

Dependent element should have absolutely disconnected and intensive classes.

To test the linearity Box-Tidwell methodology is utilized

Case processing

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	109	100.0
	Missing Cases	0	.0
	Total	109	100.0
Unselected Cases		0	.0
Total		109	100.0

a. If weight is in effect, see classification table for the total number of cases.

Fig 2.1 Processing Summary

The above table states the total number of cases available in the table. In this table 109 cases are available

Encoding variables

Dependent Variable Encoding

Original Value	Internal Value
HISPANIC	0
WHITE NO	1

Fig 2.2 Dependent Variable Encoding

In the above table elements are encoded in to 0's and 1's i.e. dichotomous values.

Summery for Model:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	84.629[a]	.410	.563

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Fig 2.3 Model summary

In the above table of the model summary Cox & Snell R square and Nagelkerke R Square is explained. 56.3% dependent variable is explained by predictor variables. The difference is changes from 41.0% to 56.3%.

Coefficient model:

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	57.538	2	.000
	Block	57.538	2	.000
	Model	57.538	2	.000

Fig 2.4 Model Coefficient

The above table explain model coefficient using chi-square test that the data in the table is statistically significant.

Hosmer and Lemeshow Test.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	18.523	7	.010

Fig 2.5 Hosmer and Lemeshow test

This test is for the evaluation of best fit model so as the p=0.010 it is not statistically significant and so this not the best fit model. The chi- square test value is 18.523 with the degrees of freedom 7.

Omnibus test coefficients:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	57.538	2	.000
	Block	57.538	2	.000
	Model	57.538	2	.000

Fig 2.6 Omnibus model coefficients

In this table Using Chi-Square value, how data can be statistically interpreted and better fit in the model.

Classification Table:

Classification Table					
			Predicted		
			Ethnicity		Percentage
	Observed		HISPANIC	WHITE NO	Correct
Step 1	Ethnicity	HISPANIC	38	1	97.4
		WHITE NO	2	68	97.1
	Overall Percentage				97.2
a. The cut value is .500					

Fig:2.7 Classification table

I this table which shows model is presented precisely by what number of percent is explained. Here in the phase one 97.2 %, Data is appropriately clarified.

Equation Variables:

Variables in the Equation									
								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Count	.115	.022	28.487	1	.000	1.122	1.075	1.170
	Rank	.203	.037	30.538	1	.000	1.225	1.140	1.317
	Constant	-17.582	3.222	29.784	1	.000	.000		
a. Variable(s) entered on step 1: Count, Rank.									

Fig 2.7 Variables in the equation

In the above table, two predictor variables are used, Using Wald test can state that whether the data is statistically significant or not statistically significant. For the first predictor variable, the Wald test value is 28.487 and statistical significance is 0.000. For the second predictable, Wald test value is 30.538 with the statistically significant value 0.000. so, the significant value is less than 0.05 ($P<0.05$) so the data is statistically significant.

3 ANOVA

Newcastle library water consumption Using Analysis of Variance(ANOVA).

The Specified dataset taken from Data.gov.uk, displays water use of Newcastle library in the year 2012,2013,2014.

The amount of levels of estimation of all Independent / Dependent components:

In this analysis, three levels of measurements are used to calculate the water use of Newcastle library. Level one indicates water use of Newcastle library in the year 2012, second level indicates water use of Newcastle library and in third water use of Newcastle library in the year 2014. In this data set water consumption is dependent list where how many cubic meter water is used by Newcastle library. The factors list consists the three years data from how much water is consumed by Newcastle library is calculated.

Hypothesis conditions for ANOVA.

Hypothesis is the significant phase for manipulation of an ANOVA. First have to mention the null hypothesis condition for ANOVA is. Ho= μ1 is equal to μ2 is equal to μ3, that mean there is no difference among the means. Following step is to state the alternative hypothesis for an ANOVA i.e. Ha= μ1 is not equal to μ2 is not equal to μ3, hence there should be at least one difference in the means among the group.

The Assumptions for doing One-Way Analysis of Variance test

1.the dependent variable should be scattered over each event normally.in this case, taking three group (2012,2013,2014) on their use of water consumption.

2. the leavens test is used for carrying out the test of homogeneity.

3. observations are independent.

Descriptive Statistics:

Oneway

[DataSet3]

Descriptives

waterconsumption

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
2012	5	354.00	12.042	5.385	339.05	368.95	334	365
2013	5	458.80	22.742	10.171	430.56	487.04	436	495
2014	5	420.80	30.285	13.544	383.20	458.40	372	446
Total	15	411.20	49.617	12.811	383.72	438.68	334	495

Fig 3.1: Descriptive Statistic

The above table demonstrates the elucidating measurement for water utilization by Newcastle library. The above table incorporate mean, standard deviation and 95% certainty interim for methods for years (2012,2013,2014). “N” is the number of tests taken for finding out the strategies for each one of the events. For 2012,2013,2014 N is 5 and the total number of elements are 15. The second segment exhibits the mean a motivating force for water use by Newcastle library for consistently. For the year 2012 mean esteem is 354.00. For the year 2013 the mean estimation of water utilization is 458.80 and in 2014 the mean an incentive for water utilization is 420.80. The total mean for water utilization by Newcastle library. The third section show the standard deviation for every one of the years For 2012 is 12.042, for 2013 it is 22.742 and for 2014 it is 30.285 separately the aggregate mean of standard deviation is 49.617.

Test of Homogeneity of Variances:

Test of Homogeneity of Variances

waterconsumption

Levene Statistic	df1	df2	Sig.
1.733	2	12	.218

Fig 3.2 Test of Homogeneity of variences

Above table shows the homogeneity trial of changes utilizing Levene test where it tells that whether the information is homogenous or non-homogenous. the significance level in the table is 0.218 that means the significance value is greater than 0.05 i.e. (P>0.05).The significance in this test of homogeneity among the group is grater than 0.05 so the data is statistically significant and failed to reject the null hypothesis.

ANOVA inspection using SPSS.

ANOVA

waterconsumption

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	28148.800	2	14074.400	26.734	.000
Within Groups	6317.600	12	526.467		
Total	34466.400	14			

Fig 3.3 ANOVA

ANOVA is Analysis of variance test is to check the variation among two or more than two variables. The above table indicates the ANOVA assumptions which includes the sum of square among the group and sum of square within the group. In the above table the sum of square among the group is 21848.800 and sum of square within the group is 6317.600, Total sum of square within and among the group is 34466.400. total sum of square is 34466.400. The mean Square column, it is calculated by dividing sum of square among the group by degrees of freedom i.e. (28148.800/2= 14074.400), Same for within the group sum of square with in the group is divided by degrees of freedom

(6317.600/12=526.467). on this two calculation F statistic is calculate using division of mean square among the group and mean square inside the group (14074.400/526.467=26.734). If the F statistic value is more, then significance value will be less.

Post HOC Tests for Multiple Comparisons:

Post Hoc Tests

Multiple Comparisons

Dependent Variable: waterconsumption

Tukey HSD

Double-click to activate

(I) Years	(J) Years	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2012	2013	-104.800*	14.512	.000	-143.51	-66.09
	2014	-66.800*	14.512	.002	-105.51	-28.09
2013	2012	104.800*	14.512	.000	66.09	143.51
	2014	38.000	14.512	.055	-.71	76.71
2014	2012	66.800*	14.512	.002	28.09	105.51
	2013	-38.000	14.512	.055	-76.71	.71

*. The mean difference is significant at the 0.05 level.

Fig 3.4 Multiple Comparisons

the above table shows the post hoc test for significance. In this table evaluation among the collections and within the collections is shown as follows. 2012 is compare with the year 2013, the significance value among 2012-13 is(P=0.000) and the significance among 2012-14 is (P=0.002). The significance value among 2012-14 is larger than 2012-2013 in the second group similarly significance among 2013-12,2013-2014 is (P=0.000) and (P=0.055) respectively significance difference of 2013-14 is larger than 2012-13. in the third collection evaluation among years is evaluated for significance value. significance value for 2014-12 is(P=0.002) and for 2014-13 is(P=0.055). Hence the significance value of year 2014-13 is bigger that 20014-12.

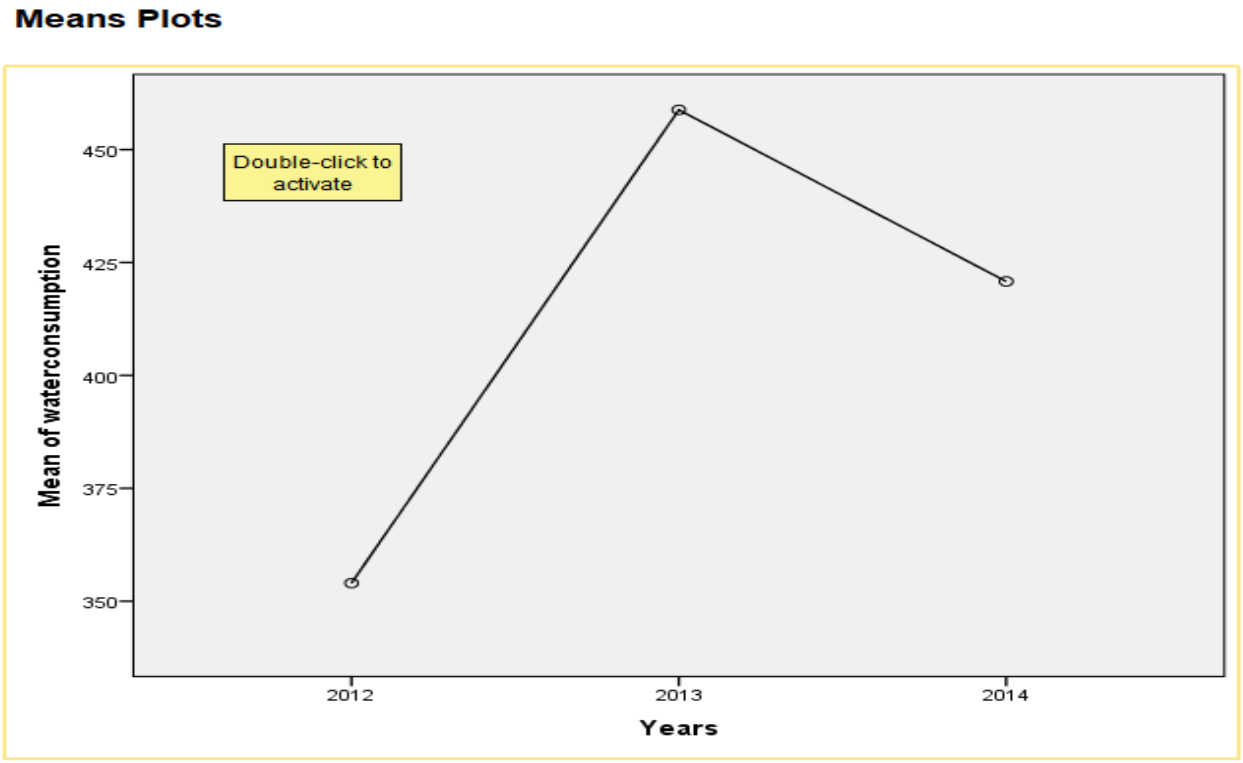

Fig 3.5 Means Plot for ANOVA

Final Result for ANOVA

By utilizing this procedure can compute the measurably huge contrast among the gathering. (F (2,12) =26.734, P=0.000).

In the multiple comparison, Utilizing Tukey post hoc strategy ascertain the centrality level among the gathering. For 2013(459+/- 22.7, P=0.000) and for 2014(421+/- 30, P =0.002) with the comparison of 2012(354 +/- 12.0). By this result can achieve that the data is statistically significant.

DATA SOURCE:

Multiple Regression:

https://catalog.data.gov/dataset/school-progress-report-2007-2008-16883

Logistic Regression:

https://catalog.data.gov/dataset/most-popular-baby-names-by-sex-and-mothers-ethnic-group-new-york-city-8c742

ANOVA techniques:

https://data.gov.uk/dataset/newcastle-city-library-energy-consumption/resource/8e150957-9f35-47ba-bb2e-c7df68047911