Multiple Regression Analysis Using ANCOVA in University Model

Maneesha and Priti Bajpai

***Abstract*—The government of UAE is promoting Dubai as an academic hub. Dubai International Academic City (DIAC) is a free zone area with many national and international universities promoting higher education in almost all disciplines. The aspiration of every graduating student from the university is to get a good placement. In Dubai diverse job opportunities in national and multinational organizations are available. The objective of the paper is to review the placement opportunities in Dubai for the universities offering programs in Engineering. This paper attempts to study the effect of three independent variables namely Cumulative grade point average (CGPA), Engineering disciplines and types of jobs that graduating students are offered on the dependent variable salary. Engineering discipline understudy are Mechanical, Electronics and Communication, Computer Science and Electrical and Electronics Engineering. The type of jobs taken into consideration are marketing, technical marketing, design and logistics. The concepts of Analysis of covariance (ANCOVA) and multiple regression are used for review of placement opportunities vis a vis the salary structure.**

***Index Terms*—ANCOVA, ANOVA, fisher's test statistic, multiple regression.**

I. Introduction

The work by Campbell and Stanley's in the field of experimental design is well regarded [1]. Many methods have been used by statisticians for reducing error variance or to statistically equate comparison groups. Out of the many tools ANOVA and ANCOVA are more frequently used. In agricultural research, however, analysis of variance and covariance (ANOVA/ANCOVA) are used most frequently. J. D. Elashoff [2] studied ANOVA and found it a sensitive tool. B. L. Hamilton [3] has used it for the study of Educational and Psychological Measurement. S. G. Dorsey and K. L. Soeken [4] used Johnson-Neyman technique as an alternative to ANOVA. Steven V. Owen, Robin D. Froman [5] have illustrated legitimate uses of ANCOVA and summarized statistical packages approach to the method and reviewed assumptions and how ANCOVA is used in contemporary nursing research. There are many statistical packages like BMDP, SPSS, SAS, and SYSTAT but their approach is different to ANCOVA. Gregory. A. Miller and Jean P. Chapman [6] find ANOVA is the most misused approach when it comes to substantive group differences on potential covariates, particularly in psychopathological research. H. J. Keselman, C. J. Huberty, L. M. Lix, S. Olejnik, and R. A. Cribbie, analyzed ANOVA, MANOVA, and ANCOVA [7].

Manuscript received June 1, 2013; revised September 28, 2013.

The authors are with Department of Mathematics, BITS, Pilani-Dubai Campus, Dubai, UAE (e-mail: maneesha@bits-dubai.ac.ae, priti@bits-dubai.ac.ae).

A. Background

Dubai International Academic City (DIAC) is the educational hub of Dubai. More than 20,000 students seek higher education in various disciplines every year. Many national and international universities cater to the needs of local students and expatriates. There is a small proportion of student populace that leaves the country to pursue higher studies after completing graduation. However, majority of them seek employment in UAE as the multicultural atmosphere and ease of living is the major attraction. The paper attempts to study the impact of Cumulative Grade Point Average (CGPA), engineering discipline and types of jobs offered on the salary structure. The study is confined to the universities in DIAC that are offering Engineering programs. The four major disciplines that are opted by majority of students are Mechanical Engineering, Electronics and Communication Engineering, Computer Science Engineering and Electrical and Electronics Engineering. The four major trends as observed in the job market of UAE are in the fields of marketing, technical marketing, design and logistics. The students with engineering background have been offered job in these fields.

II. Multiple Regression and Analysis of Covariance (ANCOVA)

A. Multiple Regression

Multiple regression is an extension of simple regression from one to several quantitative explanatory variables. It involves more than one independent variable and the curves obtained are not only used to make predictions rather for the purposes of optimization. The structural model is of the form

$$E(Y/x_1; x_2; x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (1)$$

i.e. the expected value of Y (the outcome) given the values of the explanatory variables x_1 through x_3. β's are fixed (but unknown) constants [8].

B. Interaction

Interaction is a concept in statistics that applies whenever there are two or more explanatory variables. The term interaction applies to both quantitative and categorical explanatory variables. The definition of interaction is that the effect of a change in the level or value of one explanatory variable on the mean outcome depends on the level or value of another explanatory variable. Therefore interaction relates to the structural part of a statistical model. Interpretation for interaction is done using p-value for the interaction line of the

DOI: 10.7763/IJAPM.2013.V3.232

regression results. If the interaction has a significant p-value, both explanatory variables affect the outcome (except in certain special circumstances). On the other hand, if the interaction is not significant, generally the appropriate next step is to perform a new multiple regression analysis excluding the interaction term, i.e., run an additive model.

C. Covariance

Covariance is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation [9].

D. Analysis of Covariance (ANCOVA)

Analysis of Covariance (ANCOVA) is used to infer the effect of independent variables on the dependent variable and to verify if a linear model makes sense. The ANCOVA method belongs to a larger family of models called GLM (Generalized Linear Models) as do the linear regression and Analysis of Variance (ANOVA) [10]. The specificity of ANCOVA is that it mixes qualitative and quantitative explanatory variables. ANCOVA in terms of General Linear Model is given by

$$Y = GMy + \tau + [Bi\ (Ci – Mij) + \ldots] + E \quad (2)$$

where
Y: is a dependent variable
GMy : is grand mean of dependent variable
τ: is treatment effect
Bi: is regression coefficient for *ith* covariate, *Ci*
M: is the mean of ith covariate
E: is error

1) Levene test for equality of variances

It is used to test if *k* samples have equal variances. Equal variances across samples is called homogeneity of variance. The Levene test rejects the hypothesis that the variances are equal if Levene test statistic (W)

$$W > F_{\alpha,\, k\text{-}1,\, N\text{-}k}$$

where $F_{\alpha,\, k\text{-}1,\, N\text{-}k}$ is the upper critical value of the *F* distribution with *k*-1 and *N*-*k* degrees of freedom at a significance level of α.

2) Goodness of fit

Apart from the other goodness of fit statistics, R^2 i.e coefficient of determination is of most interest to us. R^2 gives the percentage of variability of the dependent variable which is explained by the explanatory variables. R^2 has a range from 0 to 1. R^2 approaching 1 indicates a better goodness of fit between the dependent and explanatory variables [11].

3) Analysis of variance (ANOVA)

ANOVA enables us to draw inferences about whether the samples have been drawn from populations having the same mean. It is used to test for differences among the means of the populations by examining the amount of variation within each of the samples, relative to the amount of variation between the samples. In terms of variations within the given population, it is assumed that the values differ from the mean of this population only because of the random effects. The test statistics used for decision making is F = ratio of Estimate of population variance based on between samples variance and Estimate of population variance based on within samples variance [10].

III. Statistical Analysis

A. Assumptions

There are five assumptions that underlie the use of ANCOVA and affect interpretation of the results:

1) Normality of Residuals: The residuals (error terms) should be normally distributed.
2) Homogeneity of Variances: The error variances should be equal for different treatment classes.
3) Homogeneity of Regression Slopes: The slopes of the different regression lines should be equal.
4) Linearity of Regression: The regression relationship between the dependent variable and concomitant variables must be linear.
5) Independence of Error terms: The error terms should be uncorrelated.

B. Objective

The objective is to determine the effect of CGPA, Engineering discipline and types of jobs on the salary structure of graduating students and to verify if a linear model makes sense. The hypothesis is postulated as:

Null Hypothesis (Ho): No effect of CGPA, Engineering discipline and type of job on salary of Graduating students.

Alternative Hypothesis (H_1): CGPA, Engineering discipline and type of job has effect on salary of Graduating students.

C. Data Collection

The data is collected from the placement cell of Universities in DIAC offering Engineering programs over a period of three years i.e. post recession. The type of job is the categorical explanatory variable whereas the CGPA, engineering discipline and salary structure are quantitative explanatory variables. Four samples each of size 20 were formed with data on all the variables under study. The data was tabulated on spreadsheets and further analysed using XLSTAT, statistical software using the concepts of multiple regression and ANCOVA.

IV. Results

A. Summary Statistics

Inference: In all 80 observations were taken into consideration. The minimum & maximum salaries offered

are AED 3000 and AED 10,500 respectively. As shown in Table I on an average AED 6000 is the starting salary for the graduating students in UAE. The students getting placed had CGPA from 5 to 9.27.

TABLE I: SUMMARY STATISTICS

Variable	Salary	CGPA	Discipline
Observations	80	80	80
Obs. with missing data	0	0	0
Obs. without missing data	80	80	80
Minimum	3000	5	1
Maximum	10500	9.27	4
Mean	6071.25	6.919	2.488
Std. deviation	1460.617	1.007	1.067

B. Job Description

TABLE II: DISTRIBUTION OF TYPES OF JOBS

Variable	Categories	Frequency	%
Type of job	Design	20	25.000
	Logistics	19	23.750
	Marketing	21	26.250
	Technical Marketing	20	25.000

Inference: The maximum jobs offered are in the field of marketing as Dubai is mainly a trading hub as is evident from Table II.

C. Levene Test for Equality of Variances

TABLE III: LEVENE'S TEST

Levene's Test	
# of Groups (<=6)	4

	SS	df
Between Group	2008625	3
Within Group	26452750	76
Levene's Statistic	1.923625	
Critical Value (a=0.05)	2.724944	
P-value	0.132868	

Inference: As computed in Table III, $W < F_{\alpha,\ k\text{-}1,\ N\text{-}k,,}$ it implies that there is homogeneity of variances.

D. Goodness of Fit Statistics

Inference: In the above University model as observed from R^2 computations in Table IV, 43 % of the variability in the salary structure is explained by the CGPA, Engineering discipline and type of job whereas, the remaining 57% variability is due to other variables possibly personality traits, personal preferences and other constraints which were not taken into consideration during the study.

TABLE IV: GOODNESS OF FIT STATISTICS

Observations	80.000
Sum of weights	80.000
DF	67.000
R^2	0.480
Adjusted R^2	0.387
MSE	1308427.120
RMSE	1143.865
MAPE	14.136
DW	1.931
Cp	13.000
AIC	1138.560
SBC	1169.527
PC	0.722

E. Analysis of Variance

TABLE V: ANALYSIS OF VARIANCE TABLE

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	12	80874257.971	6739521.498	5.151	< 0.0001
Error	67	87664617.029	1308427.120		
Corrected Total	79	168538875.000			

Computed against model Y=Mean(Y)

Inference: Fisher's F test is used as the appropriate test statistic. Table V shows that the probability corresponding to F-value is lower than 0.0001. It implies that risk in assuming that the null hypothesis is wrong is quite low i.e. 0.01%. Therefore, we can conclude with 95% confidence that the variables CGPA, Engineering discipline and job type do bring a significant amount of information.

F. Assessment of the Impact of the Independent Variables on Dependent Variable

Assessment of the impact of the CGPA, engineering discipline and type of job on salary structure is further based on Type I and Type III analysis.

TABLE VI: TYPE I SUM OF SQUARE ANALYSIS

Source	DF	SS	MS	F	Pr > F
CGPA	1	12204895.161	12204895.161	9.328	0.003
Discipline	1	4885.313	4885.313	0.004	0.951
Job Type	3	52795620.029	17598540.010	13.450	< 0.0001

TABLE VII: TYPE III SUM OF SQUARE ANALYSIS

Source	DF	Sum of squares	Mean squares	F	Pr > F
CGPA	1	4414153.226	4414153.226	3.374	0.071
Discipline	1	5787936.152	5787936.152	4.424	0.039
Job Type	3	10853965.950	3617988.650	2.765	0.049

Inference: In Type I & Type III SS combined as evident from Table VI and Table VII, the F probability is the least corresponding to the variable type of job. It implies that its impact is the strongest on the salary in the University model in comparison to the other variables CGPA and Engineering discipline.

G. Model Parameters

TABLE VIII: MODEL PARAMETERS

Source	Value	St. Error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	439.4	3651.1	0.120	0.905	-6848.1	7727.0
CGPA	834.6	478.2	1.745	0.086	-119.9	1789.1
Discipline	2074.6	1087.5	1.908	0.061	-96.1	4245.4
Job Type-Design	6226.3	3478.8	1.790	0.078	-717.4	13169.9
Job Type-Logistics	-1200.9	3634.8	-0.330	0.742	-8456.0	6054.2
Job Type-Marketing	-1815.6	2944.8	-0.617	0.540	-7693.4	4062.3
Job Type-Technical Marketing	0.000	0.000				

Inference: From the Table VIII on model parameters we infer that amongst the job types- Logistics has the highest p-value implying its impact is least on the salary of engineering graduates.

H. Graphs

The chart below shows the predicted values versus the observed values. Confidence intervals allow identifying potential outliers (see Fig. 1-Fig. 3).

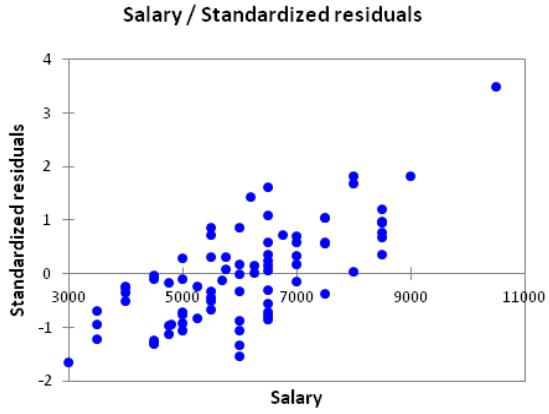

Fig. 1. Observed Salary versus standardized residuals.

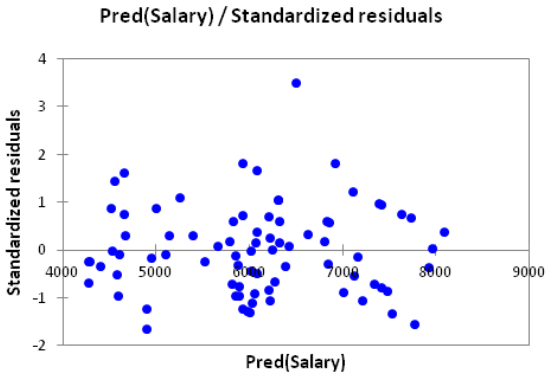

Fig. 2. Predicted Salary versus Standardized residuals.

Inference: The implication of the mean chart is that on an average highest salary structure is for the design jobs followed by technical marketing, logistics and marketing respectively.

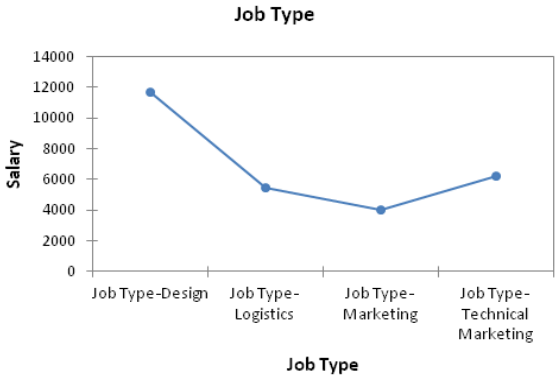

Fig. 3. Mean Chart.

V. CONCLUSION

The above analysis suggests that the impact of type of job is the strongest on the salary structure in the University model in comparison to the other variables which are academic performance as measured by CGPA and Engineering discipline. The jobs in the field of design offer the highest salary on an average in comparison to the jobs offered in other fields in UAE.

REFERENCES

[1] D. T. Campbell and J. C. Stanley, "Experimental and quasi experimental designs for research," *Chicago Rand Mc Nally*, 1963.

[2] J. D. Elashoff, "Analysis of covariance: A delicate instrument," *American Educational Research Journal*, vol. 6, pp. 383–401, 1969.

[3] B. L. Hamilton, "An empirical investigation of the effects of heterogeneous regression slopes in analysis of covariance," *Educational and Psychological Measurement*, vol. 37, pp. 701–702, 1977.

[4] S. G. Dorsey and K. L. Soeken, "Use of the Johnson-Neyman technique as an alternative to analysis of covariance," *Nursing Research*, vol. 45, pp. 363–366, 1996.

[5] S. V. Owen and R. D. Froman, "Focus on Qualitative Methods Uses and Abuses of the Analysis of Covariance," *Research in Nursing & Health*, John Wiley & Sons, pp. 557–562, 1998.

[6] G. A. Miller and J. P Chapman, "Misunderstanding Analysis of Covariance," *Journal of Abnormal Psychology*, vol. 110, no. 1, pp. 40-48, 2001.

[7] H. J. Keselman, C. J. Huberty *et al.*, "Statistical practices of educational researchers," *Review of Educational Research*, vol. 68, pp. 350-386, 1998.

[8] E. J. Pedhazur, *Multiple regression in behavioral research*, 3rd Ed., New York: Harcourt Brace, 1997.
[9] R. A. Johnson, *Probability and Statistics for Engineers*, 6th ed, Pearson Education, 2003, ch. 12, pp. 428-432.
[10] V. K. Rohatgi, *An Introduction to Probability Theory and Mathematical Statistics*, 1988, ch. 12, pp. 513-514.
[11] C. R. Kothari, *Research Methodology*, 2008, ch. 11, pp. 256-259.

Maneesha was born in Mathura, India on 1st February 1973. She received Phd Statistics in the field of sampling theory, from Lucknow University, Lucknow, India in the year 2001. MA Statistics, from Lucknow University, Lucknow, India in the year 1995. BA in statistics, mathematics and economics, from Lucknow University, Lucknow, India in the year 1993.

She is working as an associate professor of mathematics at BITS Pilani, Dubai Campus Dubai since 2002. She has worked in Cabinet Secretariat, New Delhi, India from 2001-2002. Some of her recent publications are Application Of Queuing Model In Dubai's Busiest Megaplex, International Conference on Mathematical Sciences and Statistics 2013 (ICMSS2013), Kuala Lumpur, Malaysia/February 5-7, 2013, publication in AIP conference proceedings. Implementation of Six Sigma Methodology in Academics, International Journal of Sustainable Development and Green Economics, February, 2013, vol. 2, issue 1, 2, pp. 17-21. A Study of Traffic Management at Toll Plaza on Delhi Gurgaon Expessway, Delhi, India, American Journal of Mathematics and Sciences, vol. 2, no. 1, January, pp. 63-69, 2013. Currently involved in research on queuing theory, multiple regression and quality control.

Dr. Maneesha received a commendation award from Government of India in the year 2002.