APPLICATION OF DUMMY VARIABLES IN MULTIPLE REGRESSION ANALYSIS

Abdullahi U. Usman, Hassan S. Abdulkadir and Kabiru Tukur
APPLICATION OF DUMMY VARIABLES IN MULTIPLE REGRESSION ANALYSIS

Abdullahi U. Usman, Hassan S. Abdulkadir and Kabiru Tukur
Jodhpur National University, Jodhpur, Rajasthan, India.

INTRODUCTION

The concept of experimental design was broadly explained (montgomery, 2014). Several researchers use the concept of multiple regression analysis and ANOVA in their statistical research analysis. Multiple regression analysis are frequently used in different aspects of life. (oberkirchner et al, 2010) uses multiple regression to analyses and develop a model for the effect of essential material and process parameters to weight and moisture content of impregnated papers. (Bajpai, 2013) analyse university model using multiple regression and ANCOVA and found very essential. (Syla, 2013) study the significance of active-employment programs on employment levels using multiple regression. (Everarda et al, 2005) uses multiple regression result to study the importance of engaging student to graphical user interface in teaching statistical courses. (Oswald, 2012) shows how viewing multiple regression results through multiple lenses can give a better assessment to the researchers. (Kelley et al, 2003) shows that in multiple regression obtaining accurate parameter contributes more than having statistical significance.

(Pazzani et al, 1981) shows how independent sign regression generate linear model that are almost accurate as multiple regression. (Ludlow, 2014) study suppressor variables and suppression effects in building regression model. (Moya-laraño et al, 2008) encourages ecological researchers to use partial regression in their studies. (Breheny et al, 2013) uses visreg package which is useful tool in visualizing the relationship between an explanatory variables that is estimated. Visreg construct convenient support in regression model.

Background

Jodhpur National University is a private university created under state government private university Act. The university is situated in the western part of Rajasthan, India. Thousands of students seek admissions in various disciplines every year. However, majority of their students seek employment opportunities within India as the unity in diversity atmosphere and abroad. The paper attempts to study the effects of CGPA, types of job and engineering discipline on the salary structure. The study was made in Jodhpur National University for the students offering engineering programs. Marketing, technical marketing, design and logistics are the four major trends as observed in India. Electronics and communication, computer science and engineering, mechanical and electrical engineering are the major engineering discipline chosen by the students. These jobs are offered mostly to the student with engineering background.
Multiple Regression

The statistical techniques of extending linear regression so as to consider two or more independent variables are known as multiple regression analysis. Multiple linear regression takes the following form

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon_i \]

Where \( \beta_0 \) is the intercepts, \( \beta_1, \beta_2, \ldots, \beta_n \) are regression coefficients.

Parameter Estimation

The method of least square is typically used to estimate the regression coefficients in a multiple linear model. The method of least square chooses the \( \beta \)'s in the equation (1) so that the sum of squares of errors \( \epsilon_i \) is minimized. To solve the equations we may use matrix:

\[ Y = X\beta + \epsilon \]

\[ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1n} \\ 1 & x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \]

in general, \( y \) is an \((n \times 1)\) vector of the observations \( X \) is an \((n \times p)\) matrix of the levels of the independent variables, \( \beta \) is a \((p \times 1)\) vector of the regression coefficients, and \( \epsilon \) is an \((n \times 1)\) vector of random errors.

Dummy Variables

These variables usually indicate the presence or absence of the “quality” or an attribute, such as male or female. They are essentially nominal scale variables. One way we could quantify such attributes is by contributing artificial variables that take on values 1 or 0, 1 indicating the presence of that attribute and 0 indicating the absence of that attribute variables that assumed such 0 and 1 values are called dummy variables. Such variables are thus essentially a device to classify data into mutually exclusives categories such as presence and absence. Dummy variables can be incorporated in regression models just as easily as quantitative variables.

Analysis Of Variance

It involves classifying and cross-classifying data and then testing if the means of a specified classification differ significantly. The rationale behind ANOVA is that the amount of variation in a set of data can be attributed to chance and specified causes.

ANOVA allows for investigating any number of factors which are hypothesized to influence the dependent variable. The assumption of equal variances also implies that for all practical purposes, the means also comes from same population since any normal population is defined two parameters mean and variance. Thus the variance within samples is only because of random effect whereas the difference between samples is due to specified factor whose effect we are trying to study.

\[ F = \frac{\text{Estimate of population based on between sample variance}}{\text{Estimate of population based on within sample variance}} \]

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>SS_treatments</td>
<td>( a - 1 )</td>
<td>( \frac{SS_{treatments}}{a-1} )</td>
<td>MS_treatments</td>
</tr>
<tr>
<td>Blocks</td>
<td>SS_blocks</td>
<td>( b - 1 )</td>
<td>( \frac{SS_{blocks}}{b-1} )</td>
<td>MS_E</td>
</tr>
<tr>
<td>Error</td>
<td>SS_E</td>
<td>( (a-1)(b-1) )</td>
<td>( \frac{SS_E}{(a-1)(b-1)} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SS_E</td>
<td>( N - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Statistical Analysis

Assumptions

i. All samples are drawn from normally distributed population

ii. All samples are drawn independently of each other

iii. Within each sample, the observations are sampled randomly and independently of each other

iv. Factor effects are additive

v. All populations have a common variance

Objective

The objective here is to determine the impact of CGPA, types of jobs and engineering discipline in the salary structure of graduating student. And to determine whether can be fitted linearly. The hypothesis declared below;

\( H_0 \): There is no significant effect of CGPA, types of job and engineering discipline on the salary of graduating students.

\( H_1 \): CGPA, types of job and engineering discipline has significant effect on salary of graduating student.

Data Collection

The data is collected from office of placement Jodhpur National University offering engineering programs over a period of four years. The types of jobs and engineering discipline are categorical explanatory variables. CGPA and salary structures are quantitative explanatory variables. Four samples are collected each of size 20. The data was presented.
on spreadsheets. It has been analyzed using MINITAB and SPSS by the concept of multiple regression analysis.

**Analysis**

**Table 2** Summary observation Taken Into Consideration Was 80. Below Table Summarizes All The Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>No. of obs</th>
<th>mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALARY(1000)</td>
<td>80</td>
<td>4.234</td>
<td>4.234</td>
<td>11.000</td>
<td>27.000</td>
</tr>
<tr>
<td>CGPA(%)</td>
<td>80</td>
<td>13.13</td>
<td>13.13</td>
<td>50.00</td>
<td>92.00</td>
</tr>
<tr>
<td>Types of job</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discourse_EE</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>discourse_CSE</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3** Analysis Of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>7</td>
<td>1294.87</td>
<td>184.98</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Residual Error</td>
<td>72</td>
<td>121.02</td>
<td>1.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>79</td>
<td>1415.89</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4** Table Of Coefficient

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>18.412 2.387</td>
<td>7.714 .000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGPA</td>
<td>.029 .033</td>
<td>.091 .885 .379</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TYPES OF JOB</td>
<td>-1.666 .434</td>
<td>-3.96 -3.837 .000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAREER</td>
<td>- .968 .353</td>
<td>-2.82 -2.745 .008</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Remark**

Fisher’s F-test is an appropriate test statistic used. Table 2 above shows that the probability corresponding to p-value is 0.000, it means that risk in assuming the null hypothesis is wrong is quite low. Therefore, we can conclude that with 95% confidence that the variables CGPA, types of job and discipline do bring a significant amount of information.

**Table 4** Table Of Coefficient

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>16.665</td>
<td>1.274</td>
<td>13.08</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>CGPA(%)</td>
<td>-0.00216</td>
<td>0.01180</td>
<td>-0.18</td>
<td>0.855</td>
<td>1.127</td>
</tr>
<tr>
<td>job_design</td>
<td>-1.8119</td>
<td>0.4771</td>
<td>-3.80</td>
<td>0.000</td>
<td>2.682</td>
</tr>
<tr>
<td>job_logistics</td>
<td>0.1841</td>
<td>0.5308</td>
<td>0.35</td>
<td>0.730</td>
<td>2.338</td>
</tr>
<tr>
<td>job_marketing</td>
<td>9.7530</td>
<td>0.5835</td>
<td>16.71</td>
<td>0.000</td>
<td>2.066</td>
</tr>
<tr>
<td>discipline_CSE</td>
<td>-0.9473</td>
<td>0.8198</td>
<td>-1.16</td>
<td>0.252</td>
<td>7.277</td>
</tr>
<tr>
<td>discipline_ECE</td>
<td>0.0314</td>
<td>0.8745</td>
<td>0.04</td>
<td>0.971</td>
<td>7.986</td>
</tr>
<tr>
<td>discipline_EE</td>
<td>-1.2510</td>
<td>0.8598</td>
<td>-1.45</td>
<td>0.150</td>
<td>7.016</td>
</tr>
</tbody>
</table>

**Remark**

from the above table CGPA, job_logistics and discipline_ECE has the least impact in the salary of graduating students.

**CONCLUSION**

The analysis above shows that the impact of types of job contributed more on the salary structure in the university model when compared with other variables i.e. CGPA and engineering discipline. Job_marketing contributed more in the types of job when compared with other types of jobs offered in the university.

**References**

10. Quality Control of Impregnated Papers with a Multiple Regression Model. (2010), 5431.

How to cite this article:

*******