**ANALYSIS OF MBA Salaries (Placed)**

**What is our data?**

The data set which was given to us had 274 observations and 13 variables named accordingly. We have to fit a model taking independent values as regressor/independent variable and **salary** as a dependent variable.

Out of these 13 variables four of them were categorical, and rest of them were numerical (4 related to GMAT scores and 2 related to his college grades). First we will be performing multiple linear regression on the dataset of students who were placed.

**Early Observations from Boxplots**

1. Majority of the students are in in between their mid and late twenties.
2. On an average work experience is between 2 and 4 years.
3. GMAT columns are have some correlation to each other as the percentile is calculated while taking in mind the total score.
4. Grades improved in the fall semester.

**Some more Observations from Tables**

1. A huge majority of the population has English as their first language. The correlation is satisfactory because English speaking are important from a company’s perspective.
2. 70% of the students are males.
3. Most of the placed students were happy with the MBA course.

**Drawing contingency tables & chi-square tests**

* We fail to reject the Null Hypotheses in every pair of chi square tests done on Categorical Data. This means that the variables are independent.

**Hypothesis testing**

**Assumptions taken throughout:** The data sets is normally distributed. The data points are independent of each other. Some tests involve a dichotomous variable.

**Hypothesis 1:** The salary of a male MBA graduate is more than a female MBA graduate.

**Hypothesis 2:** The salary of a native English speaker is higher.

**Hypothesis 3:** Graduates with higher work experience have higher salaries.

**Hypothesis 4:** Graduates with better GMAT percentile have higher salaries.

**Final Results:**

1. So, the null hypothesis fails in the first two ones. Because the p-value is greater than 0.05. So we fail to reject the Null Hypothesis that there is no relation between salary and the sex or first language of an individual.
2. In the last two ones, the p-value is less than 0.05. So, the hypotheses fails.

**Regression model**

* The salary cannot be zero as all these students are placed in an organisation. . The intercept is just an adjustment term.
* We have observed three linear models.
  1. In the first model, we take into account all the relevant variables.
  2. Then in the second one, we remove the GMAT scores and percentiles.
  3. At last, we removed the grades.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| R squared value | 0.3335 | 0.2933 | 0.2847 |

* These multiple R-squared values indicate how much does the model account for the variance in the Salary. By this statement, Model 1 is the best model for regression.
* The p-value is 6.027e-05. So the model has a 95% confidence interval.

**ANALYSIS OF MBA Salaries (Not Placed)**

**What is our data?**

Now we will look into the data of unplaced students. We will have the same number of columns but the observations would decrease. We would be omitting the people who didn’t fill the salary amount or were not willing to disclose the amount.

**Drawing contingency tables & chi-square tests**

* Out of all the categorical data, only the satisfaction level and first language have no significant relation.

**Logistic Regression**

1. First we convert the salary column into a categorical data set where 0 denotes Not Placed and 1 denotes placed.
2. Only ‘age’ and ‘satis’ are statistically significant having a p-value less than 0.05. This means these two terms have a larger impact of a student placement chances. Negative coefficient of age implies that larger age reduces chances of placements and on the other hand, positive coefficient of satis imply that a satisfied student has larger chances of getting placed.
3. After going through **anova** output, we see that age, satisfaction level of student and his GMAT percentile tend to decrease the overall deviation of the model.