**MBA Salaries Analysis**

**By**

Name : Mayank Sagar

[mayanksagar26@gmail.com](mailto:mayanksagar26@gmail.com)

Phone : 9620834479

College: R.V.College Of Engineering, Bengaluru.

File name : MBA Salaries Data.csv

Introduction : This is the analysis of MBA Starting Salaries HBR Case Study. The data set with which we are dealing with here is a classification data set about MBA students and about their salaries , GMAT Scores , percentiles, age, sex etc...So the main purpose of the data is used to tell the significant managerial issue at stake.

The analysis and inferences from the data set are as follows:

* The average of GMAT quantitative percentile is greater than verbal GMAT percentile.
* The spring MBA average score is very well correlated to fall MBA average score.
* The average work experience people have is 3.8 with median of 3, but there are lot of outliers to it also.
* The mean GMAT score is 619.5 and the median is 620.
* The mean spring average is 3.025 and mean fall average is 3.062
* The mean work experience is 3.872yrs
* The mean percentile of the students is 84.52 and the median is 87.
* The number of placed students to those who weren’t is almost in a 1:1 ratio
* Positive correlations are seen between, work\_yrs ,salary and age.
* Positive correlation is seen between quarter, s\_avg and f\_avg, and among gmat\_vpc, gmat\_tpc, gmat\_tot and gmat\_qpc.
* Most important predictor for salary is work\_yrs, gmat\_tpc and age.

1. Observations from Contingency Table and chi square test
   1. We can’t predict the sex of the MBA student from work experience (mytable)
   2. We can predict satisfactory level of the MBA program from work experience(mytable1)
   3. We can’t predict the sex of the MBA student from the his/her first language.(mytable2)
   4. We can predict the work experience of a MBA student from his/her first language.
2. Observations from T\_Test
   1. We can’t say that the average salary of males is greater than females because the p value is more than 0.05.
   2. We can’t say that the average salary of English speaking students is greater than other language speaking students because the p value is more than 0.05.

.

* 1. Also we can’t say the the average GMAT percentile of males is more than that of females as we can’t reject the null hypothesis (p value>0.05)

1. Multiple Regression Model 1
   1. The first model has salary as response variable or dependent variable.
   2. Predictor variables are work\_yrs and gmat\_tot.
   3. The R^2 value is 0.9712 which is a very good model
   4. The model’s, p-value: < 2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero (or in other words, the predictor variable has no explanatory relationship with the response variable).
   5. b0 = -1(assumption), b1 = 3264.2887, b2=146.7158  
      # Model: salary = -1 + 3264.2887\*work\_yrs + 146.7158\*gmat\_tot
2. Multiple regression model 2
   1. The first model has salary as response variable or dependent variable.
   2. Predictor variables are are work\_yrs , gmat\_tot., age, frstlang and sex.
   3. The R^2 value is 0.9785 which is very good model
   4. The model’s, p-value: < 2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero (or in other words, the predictor variable has no explanatory relationship with the response variable).
   5. b0 = -1(assumption), b1 = -958.86681 , b2=2905.80137 ,b3=-15290.15225, b4=40.35853  
      # b5=-2260.92128, b6=794.41173  
      # Model: salary = -1 -958.86681\*work\_yrs + 2905.80137\*age - 15290.15225\*frstlang
3. Multiple regression model 4
   1. The first model has salary as response variable or dependent variable.
   2. Predictor variables are are work\_yrs , and sex.
   3. The R^2 value is 0.2223which is very good model
   4. The model’s, p-value: 3.471e-06 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero (or in other words, the predictor variable has no explanatory relationship with the response variable).
   5. b0 = 99676.944 , b1 = 2629.973, b2= -4860.589  
      # Model: salary = 99676.944 + 2629.973\*work\_yrs + -4860.589\*sex
4. Multiple regression model 3
   1. The first model has salary as response variable or dependent variable.
   2. Predictor variables are are work\_yrs and age.
   3. The R^2 value is 0.2506which is not so good model
   4. The model’s, p-value:5.438e-07is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero (or in other words, the predictor variable has no explanatory relationship with the response variable).
   5. b0 = 36967.5, b1 = 388.8, b2= 2413.8  
      # Model: salary = 36967.5 + 388.8\*work\_yrs + 2413.8\*age
5. CHALLENGE ACCEPTED
   1. In logistics regression the dependent variable can only be categorical , so we selected sex as the dependent variable.
   2. The model is prepared for both who were placed and not placed .
   3. The accuracy of both the models came out to be very less and approximately equal to 4.5% of accuracy.