Regression Analysis on R: Salary and Gender

For this analysis, a regression was conducted using multiple variables within the “Placements” dataset obtained from Kaggle. The aim was to determine whether there was a statistically significant correlation between the Y variable i.e., salary, and the X variables, namely gender, work experience, MBA scores, degree scores and SSC board scores. The salary variable represents salaries offered through placements. Salaries with the value 0 indicate that the individual in question was not placed.

The sample regression equation thus obtained was:

Yi = B^0 + B^1(gender)+ B^2(work ex) + B^3(MBA scores) + B^4(degree scores) + B^5(board scores) + ui.

To begin, the null hypothesis states that there is no statistically significant effect of the X variables on the Y variable. The alternative hypothesis states that there is a statistically significant correlation between them.

H0 = B^0+ B^1(gender)+ B^2(work ex)+ B^3(MBA scores) + B^4(degree scores) + B^5(board scores) + ui = 0.

H1 = B^0+ B^1(gender)+ B^2(work ex)+ B^3(MBA scores) + B^4(degree scores) + B^5(board scores) + ui ≠ 0.

Findings

• The Y intercept B^0 = -414058.9 with a standard error 109621.7, t value -3.777and PR>|t| 0.000207

• The slope coefficient B^1(male) = 52827.1 with a standard error 18555.7, t value 2.847 and PR>|t| = 0.004854

• The slope coefficient B^2(work ex YES) = 64383.0 with a standard error 18092.7, t value 3.559 and PR>|t| = 0.000462

• The slope coefficient B^3(MBA score) = -2704.7 with a standard error 1681.3, t value -1.609 and PR>|t| = 0.109185

• The slope coefficient B^4(degree score) = 4592.6 with a standard error 1396.2, t value 3.289 and PR>|t| = 0.001178

• The slope coefficient B^5(board score) = 6241.6 with a standard error 946.4, t value 6.595 and PR>|t| = 3.41e-10

• Residual standard error: 122100 on 209 degrees of freedom

• Multiple R-squared: 0.3923, Adjusted R-squared: 0.3778

Interpretation of Findings

• We find that if all X variables were 0, salaries would have a baseline value of -414059. With the help of the slope coefficients, we infer that being a male, having work experience, higher SSC scores and higher degree scores have a positive effect on salaries, while, interestingly, MBA scores have a negative effect on salaries.

❖ If the candidate is a male, his salary will increase by Rs. 53,000.

❖ If the candidate has had work experience, his/her salary will increase by Rs. 64,000.

❖ If the candidate’s MBA score increases by 1, his/her salary will decrease by Rs. 2,705.

❖ A unit increase in the candidate’s degree score indicates a Rs. 4,600 increase in salary.

❖ A unit increase in the candidate’s board score indicates a Rs. 6,250 increase in salary.

- The standard errors of the estimators tell us how much the estimates vary from the actual population numbers. Ideally, the error should be lower relative to the estimate. The respective standard errors of all

slope coefficients indicate a relatively very small difference between the estimates and the actual values. Therefore, relying on the estimates for an accurate representation is justified.

- The t values tell us how many standard deviations exist between the estimate and 0. The higher the absolute t value, the more reliable the estimate, and the greater the likelihood of rejecting the null hypothesis, which states 0 correlation. The t values for all slope coefficients except MBA scores indicate a statistically significant correlation.

- The P values, all of which are extremely small for this particular model, indicate that we are very unlikely to wrongly reject the null hypothesis, given a 0.05 level of significance. The relationship between salary and the independent variables is not due to mere chance.

- Another indication of the statistically significant correlation are the R^2 and adjusted R^2 scores. An adjusted R^2 value of 0.38 tells us that 38% of the variance in salaries can be explained by the model. 62% of the variance may be attributed to other causes such as employability test scores, degree college board, etc, all of which get absorbed by the error term.

Assumption of Homoskedasticity

1. Breusch-Pagan Test:

In order to check for heteroskedasticity in the model, the Breusch-Pagan test was conducted on R Studio. The BP value was 6.38, and its corresponding p-value was 0.271. Since the p-value is greater than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to claim that heteroskedasticity exists in the model. Hence the assumption of homoskedasticity is satisfied.

2. Graphical Tests for Heteroskedasticity:

❖ Residual vs Fitted Plot: This graph indicates that as the fitted values increase, the spread of residuals is not enough to warrant remedial measures.

❖ Q-Q Plot: The residuals largely follow a 45-degree line, indicating normally distributed values.

Assumption of No Multicollinearity

To check for multicollinearity among the independent variables, variance inflation factors were taken into account. Each independent variable exhibited a VIF of less than 2, indicating weak multicollinearity which does not warrant corrective measures.

Conclusions Drawn

- With the given information, it is clear that salaries offered upon placement are dependent on gender, work experience, MBA scores, degree scores and board scores. In this particular model, gender and work experience have the greatest influence on salaries. These qualitative variables explain that if one is male, one may expect greater compensation than if one is female; and prior work experience predisposes one to be offered greater salaries.

- The effect of the gender variable could be attributed to certain biases on the part of the employer. For example, hiring managers may offer lower salaries to women based on the assumption of possible maternity leave in the near future. The effect of work experience suggests that those who have been previously employed have a proven track record and good recommendations, and consequently greater bargaining power and greater opportunity to be hired in top positions.

- The MBA variable suggests that MBA scores have a negative effect upon salaries. However, upon observing its respective P value, we conclude that at the 5% level of significance, the MBA variable is not statistically significant.

- There are a couple of precautions to be taken. It is important that we do not overestimate the impact of any particular variable, as the model explains only 38% of the variation in salaries. Many other factors such as degree college board, SSC board and employability test scores might be capable of influencing the model, but are being subsumed by the error term.

- However, given that the assumptions of homoskedasticity and no multicollinearity were satisfied, the slope coefficients may be trusted. We may thus conclude that the independent variables used in the model have a statistically significant effect on salary.