Indian Institute of Management Bodh Gaya

Econometric Analysis for Business Management

Final Project:

Predicting Salary, Analysing various cause and effect between Age, Gender, Experience of Professionals.

Feb 25, 2023

Submitted To

Prof. Hari Venkatesh

Submitted By

Vivek Saurav
MBA/1291/08
Section – B

Aiming to predict salary of individuals from a dataset, considering factors such as Age, Hours worked per week, Education Years, analyzing gender factors & explaining various parameters and tests performed.

Dataset – Data of 26,000 individuals was taken (Dataset made available through a case competition on UnStop) where various parameters are provided regarding individuals of multiple countries.

Sheet 2 of Excel Data – Predicted model will be used in second sheet.

- Finding correlation between Age & Salary

```
> # finding correalation between factors
> cor1<-cor(Dataset$Salary, Dataset$Age)
> cor1
[1] -0.07470506
>
```

Age and Salary have a weak negative link with a correlation coefficient of -0.07. This suggests that as an individual ages, their wage tends to fall slightly, but the association is weak.

Correlation does not indicate causation, as other factors may affect the relationship between age and salary. For instance, elderly people may have different employment, education, or experience than younger people.

- Finding correlation between Salary & No. of years of education

```
> cor2<-cor(Dataset$Salary, Dataset$EducationYrs)
> cor2
[1] -0.04326789
>
```

A correlation coefficient of -0.04 indicates a very weak negative correlation between Salary and Education Years. This means that there is a slight tendency for individuals with more years of education to have slightly lower salaries, but the relationship is not strong.

- Finding correlation between Salary & Hours worked per week
-

```
> cor3<-cor(Dataset$Salary, Dataset$HoursPerWeek)
> cor3
[1] -0.01632606
>
```

A correlation coefficient of -0.01632606 indicates a very weak negative correlation between Hours Worked and Salary. This means that there is a slight tendency for individuals who work more hours to have slightly lower salaries, but the relationship is not strong.

Multiple Linear Regression Model Equation:

Y = βo + β1*Age + β2*HoursPerWeek + β3*EducationYrs + μi (error estimate)

Code provided in R file

Summary:

```
> Multiple_Regression <- lm(Salary~Age+EducationYrs+HoursPerWeek, data=Dataset)
> summary(Multiple_Regression)

Call:
lm(formula = Salary ~ Age + EducationYrs + HoursPerWeek, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-187899  -70851  -10653   46934 1286472

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  229873.64    3581.47  64.184  < 2e-16 ***
Age            -561.45      47.66 -11.780  < 2e-16 ***
EducationYrs  -1642.09     256.67  -6.398  1.6e-10 ***
HoursPerWeek    -47.33      53.53  -0.884    0.377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104900 on 25995 degrees of freedom
Multiple R-squared:  0.007272,  Adjusted R-squared:  0.007158
F-statistic: 63.48 on 3 and 25995 DF,  p-value: < 2.2e-16
```

Coefficents analysis:

When all independent variables are zero, Salary is 229873.64, the calculated intercept coefficient (i.e., a person with zero age, zero education years, and zero hours worked per week). The p-value is less than 0.05.

Aging is anticipated to decrease salary by 561.45 units per year. This coefficient has a p-value less than 0.05.

EducationYrs is anticipated to lower Pay by 1642.09 units per year. This coefficient has less than 0.05 p-value.

The predicted coefficient for HoursPerWeek is -47.33, meaning that for every hour worked per week, Pay falls by 47.33 units. This coefficient is not statistically significant because the p-value is 0.377, which exceeds the alpha level of 0.05.

The residual standard error (RSE) measures unmodeled variability in the response variable (Salary). The RSE for this model is 104900, indicating that the usual difference between observed Salary values and model predictions is 104900 units.

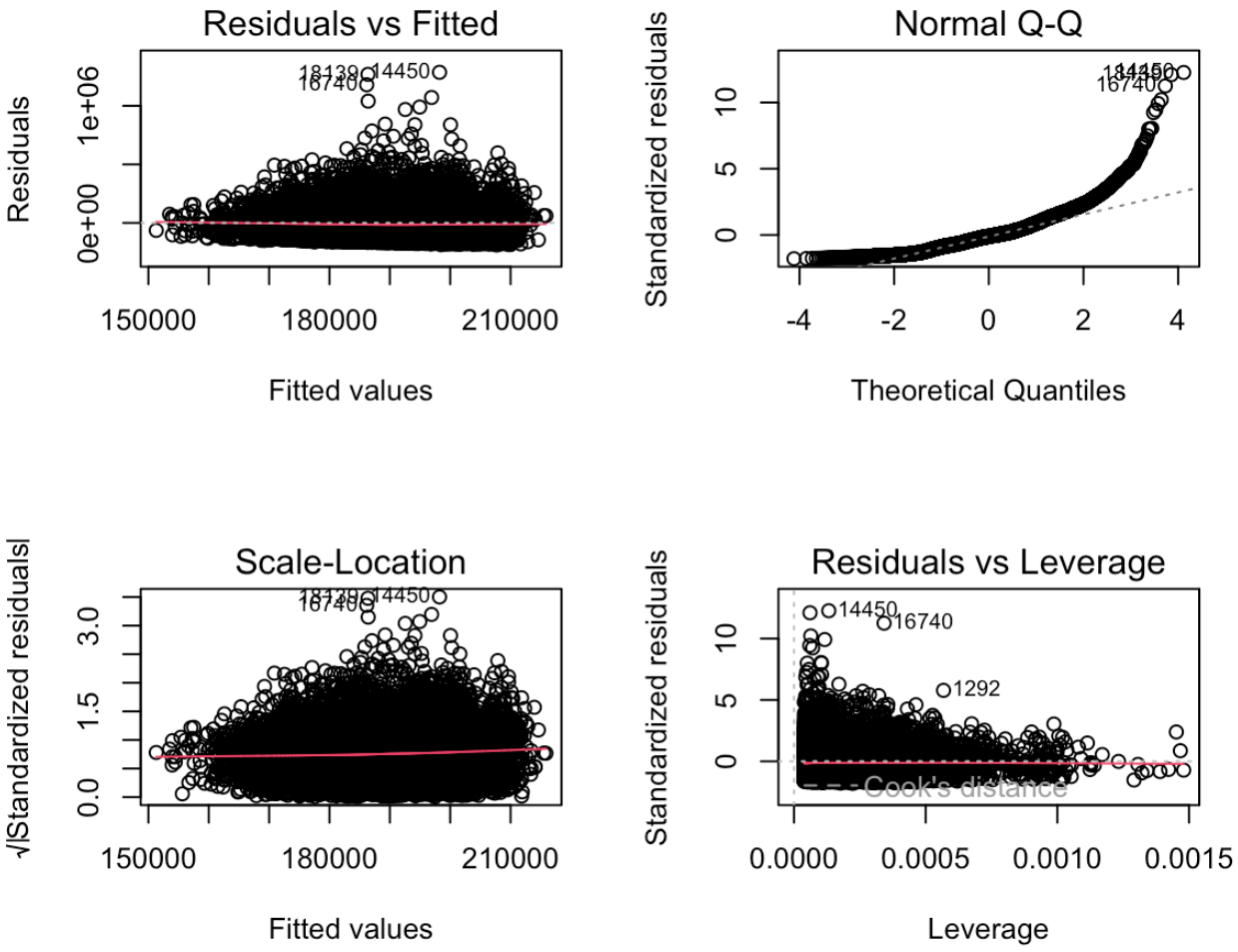
This model's multiple R-squared score is 0.007272, which suggests the independent variables explain just 0.73% of Salary variability. The model does not predict salary well based on the independent factors.

This model's adjusted R-squared value, 0.007158, is comparable to the multiple R-squared value but accounts for the number of independent variables. For comparing models with varying numbers of independent variables, the adjusted R-squared is favoured over the multiple R-squared because it penalises adding extraneous variables that do not improve model fit.

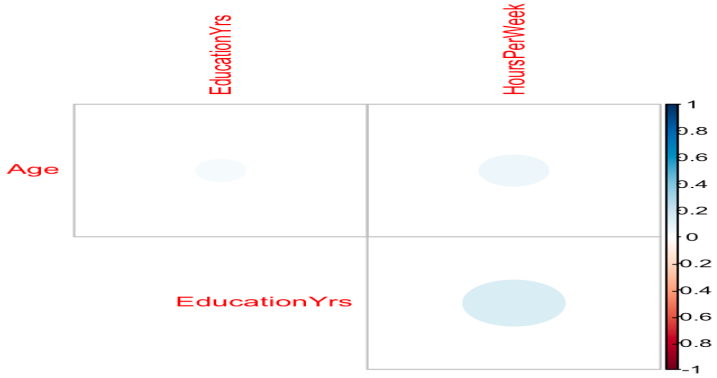
This model's F-statistic is 63.48 with 3 and 25995 degrees of freedom, testing its relevance. The model is statistically significant because its F-statistic p-value is less than 0.05.

The model's multiple and adjusted R-squared are low, showing that the independent variables do not explain much of Salary's variability. The low R-squared values show that additional factors may influence salary more than the model accounts for. The model's statistical significance implies a link between independent variables and salary.

Plot:



Correlation Plot:



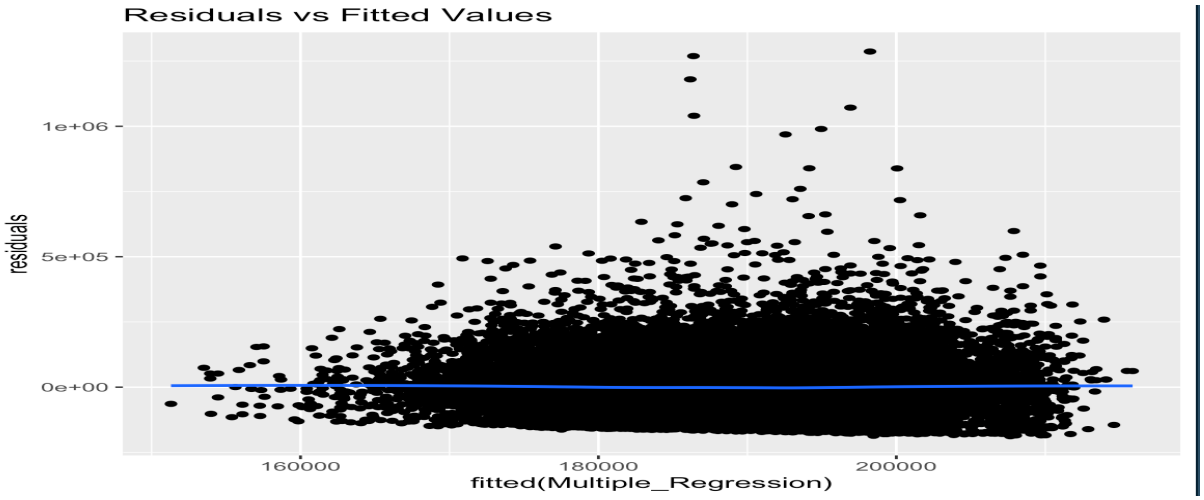
Multi-Collinearity Check:

```
> vif(Multiple_Regression)
       Age EducationYrs HoursPerWeek
  1.005118     1.022788     1.026244
>
```

It is found to be less than 5 which is accepted level according to various research parameters.

The model's independent variables have low VIF values close to 1, indicating less multicollinearity. Age, EducationYrs, and HoursPerWeek have VIFs of 1.005118, 1.022788, and 1.026244, respectively. These values are close to 1, indicating that the model's independent variables are uncorrelated. This shows the model is not multicollinear and the regression coefficient estimations are credible.

Plotting residuals vs fitted values:



Heteroscedasticity Check using Breusch-Pagan Test:

```
> bp_test<- bptest(Multiple_Regression)
> bp_test

        studentized Breusch-Pagan test

data:  Multiple_Regression
BP = 37.106, df = 3, p-value = 4.371e-08

>
```

The output shows the studentized Breusch-Pagan test for Multiple Regression. BP = 37.106 with 3 degrees of freedom yields a p-value of 4.371e-08. This p-value is less than 0.05, indicating model Heteroscedasticity.

Heteroscedasticity – A situation where the variance of the residuals is not constant across different values of independent variable.

Use a robust standard error estimator or change the data to stabilise the variance is one way to solve this problem.

Heteroscedasticity removal:

```
> # Correct for heteroskedasticity using HC standard errors
> model_HC <- coeftest(Multiple_Regression, vcov = sandwich)
> # Print the corrected coefficients and standard errors
> summary(model_HC)
    Estimate          Std. Error         t value          Pr(>|t|)
 Min.   : -1642.1   Min.   :  45.98   Min.   :-12.211   Min.   :0.00000
 1st Qu.:  -831.6   1st Qu.:  52.10   1st Qu.: -7.846   1st Qu.:0.00000
 Median :  -304.4   Median : 155.53   Median : -3.633   Median :0.00000
 Mean   : 56905.7   Mean   :1003.12   Mean   : 10.852   Mean   :0.09549
 3rd Qu.: 57432.9   3rd Qu.:1106.56   3rd Qu.: 15.066   3rd Qu.:0.09549
 Max.   :229873.6   Max.   :3655.46   Max.   : 62.885   Max.   :0.38194
>
```

In the Estimate and Std. Error columns, respectively, are the corrected coefficients and standard errors displayed. The fact that the corrected standard errors are bigger than the original ones shows that heteroscedasticity caused the original standard errors to be overestimated.

Overall, the corrected coefficients and standard errors indicate that Age and EducationYears continue to have a considerable impact on salary, while HoursPerWeek has no such impact. The total model may be marginally significant at the 0.05 level, according to the mean p-value of 0.09549.

Autocorrelation check using Durbin-Watson Test:

```
> # Check for autocorrelation using the Durbin-Watson test
> dw_test <- durbinWatsonTest(Multiple_Regression)
> # Print the test result
> dw_test
 lag Autocorrelation D-W Statistic p-value
   1    0.0005201893       1.99874   0.942
 Alternative hypothesis: rho != 0
>
```

The Durbin-Watson statistic in this instance is 1.99874, or nearly two for no autocorrelation. The p-value is substantially higher than the significance level of 0.05 at 0.942. As a result, we are unable to rule out the possibility of autocorrelation. This implies that the residuals of the multiple regression model do not show any indication of considerable autocorrelation.

Although there is no significant autocorrelation, still trying to make the model more robust

Corrected :

```
> summary(co_model)
Call:
lm(formula = Salary ~ Age + EducationYrs + HoursPerWeek, data = Dataset)

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)  229867.153   3581.551  64.181 < 2.2e-16 ***
Age            -561.469     47.663 -11.780 < 2.2e-16 ***
EducationYrs  -1640.894    256.672  -6.393 1.655e-10 ***
HoursPerWeek    -47.347     53.527  -0.885    0.3764
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104886.5 on 25994 degrees of freedom
Multiple R-squared:  0.0073 ,  Adjusted R-squared:  0.0072
F-statistic: 63.5 on 3 and 25994 DF,  p-value: < 7.137e-41

Durbin-Watson statistic
(original):    1.99874 , p-value: 4.595e-01
(transformed): 1.99986 , p-value: 4.955e-01
>
```

Model Misspecification Test :

Influential observations, or data points with a considerable impact on the outcomes of the statistical model, are identified using influence measures.

Summary of Corrected Model:

```
Call:
lm(formula = Salary ~ Age + I(Age^2) + HoursPerWeek + EducationYrs,
    data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-187135  -70836  -10688   46920 1286504

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  226943.694   5218.257  43.490  < 2e-16 ***
Age            -363.005    261.426  -1.389    0.165
I(Age^2)         -2.322      3.008  -0.772    0.440
HoursPerWeek    -61.732     56.685  -1.089    0.276
EducationYrs  -1667.286    258.735  -6.444 1.18e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104900 on 25994 degrees of freedom
Multiple R-squared:  0.007295,  Adjusted R-squared:  0.007142
F-statistic: 47.76 on 4 and 25994 DF,  p-value: < 2.2e-16

>
```

When maintaining all other predictors fixed, the Age coefficient estimates pay change per year of age. The correlation is negative (-363.005), suggesting salary decreases with age. We cannot conclude that Age and Salary are linear because this coefficient is not significant (p = 0.165).

Age squared coefficient measures wage change per unit increase in Age squared. This coefficient is also not significant (p = 0.440), suggesting no quadratic association between Age and Salary.

When maintaining all other predictors fixed, the HoursPerWeek coefficient calculates the projected pay change per hour worked per week. The coefficient is negative (-61.732), indicating that compensation decreases as hours worked per week increase. We cannot conclude that HoursPerWeek and Salary are linear because this coefficient is not significant (p = 0.276).

The EducationYrs coefficient measures the expected wage change per year of education, holding all other factors constant. The coefficient is negative (-1667.286), indicating that salary decreases with education. EducationYrs and Salary are linearly related (p < 0.001).

The output residuals show the discrepancies between the actual and anticipated response variable (Salary) values for each observation in the dataset. Residual standard error (104900) estimates residual standard deviation.

The model predictors explain 0.007295 of the variance in Salary. Predictors explain only 0.73% of salary variation.

The adjusted R-squared value (0.007142) accounts for model predictors. The modified R-squared value penalises the model for having additional predictors.

The F-statistic (47.76) and p-value (< 2.2e-16) show that at least one predictor has a significant connection with the response variable.

Final Prediction:

```
> Finalpred
       1        2        3        4        5        6        7        8        9       10       11       12
185110.7 180511.1 192321.6 187041.5 190815.1 184522.3 194257.1 184005.1 187030.4 183457.4 188722.2 189819.7
      13       14       15       16       17       18       19       20       21       22       23       24
193839.5 189855.7 187898.8 202470.2 199251.1 195474.9 195038.9 180921.1 178327.7 184330.1 200588.3 192900.7
      25       26       27       28       29       30       31       32       33       34       35       36
179968.5 175189.5 201733.5 180193.5 189310.5 186106.4 194148.7 199365.7 181762.3 194821.6 200234.9 190029.2
      37       38       39       40       41       42       43       44       45       46       47       48
199154.4 202659.5 194440.4 181692.7 202468.1 177037.8 192132.8 186106.4 199251.1 174564.1 183830.4 180664.7
      49       50       51       52       53       54       55       56       57       58       59       60
187347.7 193469.1 197275.2 202799.7 176039.8 177918.3 186043.6 187898.9 197860.9 190584.6 190188.4 196488.8
      61       62       63       64       65       66       67       68       69       70       71       72
189819.7 203811.3 186694.6 178146.9 194210.4 193391.3 195816.9 183706.9 183821.7 197275.2 200560.1 189315.1
      73       74       75       76       77       78       79       80       81       82       83       84
188467.8 198224.1 165866.9 197974.9 186231.5 182071.5 206628.2 204320.6 202799.7 177644.6 187857.2 179474.6
      85       86       87       88       89       90       91       92       93       94       95       96
189001.1 184015.6 186106.4 186007.2 203158.0 177277.8 177898.6 191191.5 196742.9 196797.5 187773.2 194704.2
      97       98       99      100      101      102      103      104      105      106      107      108
173789.0 190697.6 181692.7 195474.9 160131.9 181097.3 178324.4 199612.6 196988.9 195474.9 208122.4 199823.4
     109      110      111      112      113      114      115      116      117      118      119      120
195984.2 190126.5 192750.1 182317.9 181241.3 196742.9 193391.3 189758.8 179574.0 182771.8 179824.4 200297.1
     121      122      123      124      125      126      127      128      129      130      131      132
199154.4 183329.6 189819.7 189702.4 200374.9 180609.1 200353.4 190982.3 193919.2 190112.5 191247.7 199419.2
```

Predicted salaries of 26000 individuals was provided, a second excel sheet is attached in original Dataset where these results will be embedded.

The model used to predict salary of individuals has been attached in another sheet where salaries were not provided. Same model was used in excel sheet.