ADSC1910 Case 8

Neha Malhan(T00757554) & Thiwanka Abeyweera(T00754871)

Introduction

This analysis investigates the key factors influencing salary by examining a range of personal, educational, and geographic attributes. The goal is to identify which characteristics most significantly impact earning potential and to understand how demographic and regional factors may alter their effects on salary. This study provides economic analysts valuable insights into salary patterns, supporting data-driven decision-making for wage structures. The findings aim to better understand the factors behind wage disparities.

Methods

The analysis was focused exclusively on data scientists to simplify the model and minimize potential multicollinearity issues associated with multiple job categories. By filtering the dataset to include only data scientists, complex interactions were reduced, enabling a clearer understanding of the factors influencing salary within this specific role.

Three regression models were constructed to evaluate the effects of various factors on salary. Different combinations of predictors were included in each model based on their relevance and multicollinearity assessments. Predictors such as *Salary.5, GPA, Education, Location*, and *Province* were selectively included, depending on their significance and correlation with other variables. Interaction terms were minimized to enhance interpretability and model stability.

Diagnostic tests were conducted to ensure that the final model met the assumptions of linear regression. Aspects like normality, linearity, constant variance, and independence of residuals were assessed to confirm model reliability and validity.

Findings/Results

In the first model, the influence of various independent/explanatory variables on the dependent variable, *Salary*, is analyzed to understand how each factor contributes to salary variations within the dataset. Geographical influences are accounted for by including *Location* and *Province* variables, with levels redefined to focus on "Metro" for Location and "Ontario" for Province. This releveling allows the salary effect of other location and province categories to be interpreted relative to these baseline groups. This model serves as a foundation for assessing the relative impact of each variable, providing insights into which factors are statistically significant in explaining salary variations.

```
first_model <- lm(Salary ~ Salary.5 + FatherEducation + MotherEducation + GPA
                  + Age + Experience + Education +relevel(Location, "Metro")
                  + relevel(Province,"Ontario"), data = data_ds)
```

The regression analysis indicates that the variables Salary.5, Location (Rural), and Province (Alberta and BC) are statistically significant predictors of Salary. Salary.5 has a highly significant positive effect on Salary ($p < 2e-16$), indicating a strong association between past salary and current salary. Location (Rural) shows a significant negative effect on Salary ($p = 0.00335$), suggesting that individuals in rural areas tend to have lower salaries compared to those in metro areas. Additionally, the provinces of Alberta and BC are marginally significant predictors, with Alberta ($p = 0.05151$) and BC ($p = 0.08585$) showing negative effects relative to the baseline province, Ontario.The model demonstrates an excellent fit to the data, with an adjusted R-squared value of 0.9948, suggesting that approximately 99.48% of the variability in Salary is explained by the predictors. The overall model is also statistically significant (p-value $< 2.2e-16$), indicating that the included variables collectively contribute meaningfully to explaining the salary variations.

To address multicollinearity in the first model, a Variance Inflation Factor (VIF) analysis was conducted. A threshold of 5 was used, with variables exceeding this threshold indicating multicollinearity, which can distort regression results. In this model, FatherEducation, MotherEducation, Age, and Experience each have VIF values above 5, indicating strong correlations with other predictors. FatherEducation (5.07) and MotherEducation (5.05) may reflect similar socioeconomic factors affecting salary. Age (10.52) and Experience (10.95) are also closely related, as experience generally increases with age, making it difficult to separate their individual effects on salary.

```
second_model <- lm(Salary ~ Salary.5 + GPA + Education +
                   relevel(Location, "Metro") +
                   relevel(Province,"Ontario"), data = data_ds)
```

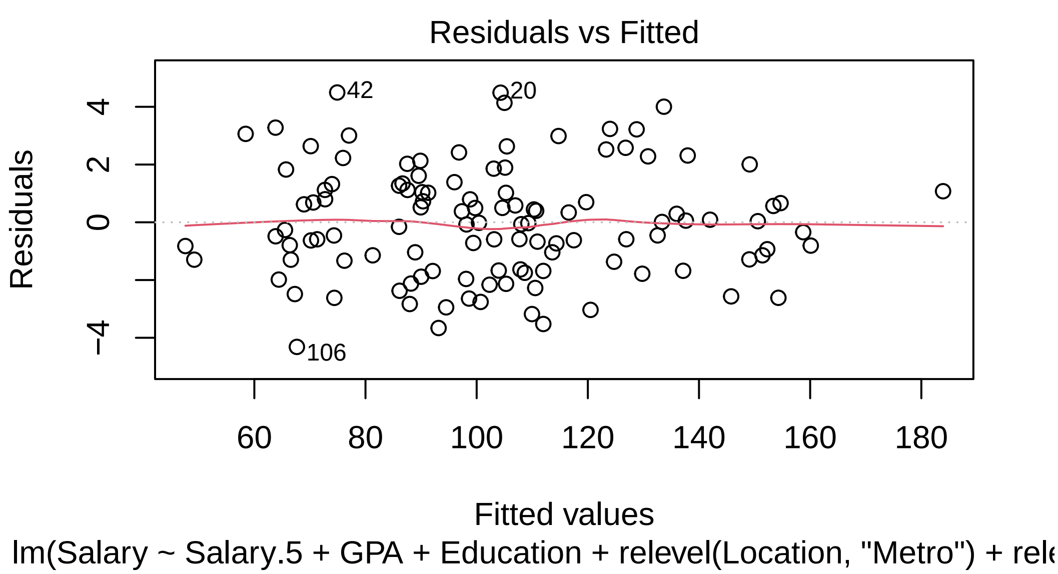
Based on the VIF findings from the first model, a second model was created by removing all highly correlated variables. This adjustment slightly improved the adjusted R-squared value to 99.49%, indicating a minor enhancement in the model's explanatory power while reducing multicollinearity.

```
third_model <- lm(Salary ~ GPA + Experience*relevel(Location, "Metro") +
                  Education*relevel(Province,"Ontario") +
                  relevel(Location, "Metro") +
                  relevel(Province,"Ontario"), data = data_ds)
```

To address the issue of time dependency, a third model was created by removing the *Salary.5* variable, as well as *FatherEducation, MotherEducation*, and *Age*, which exceeded the VIF threshold in the initial analysis. Although *Experience* also showed a high VIF value, it was retained due to its established influence on salary. In this model, interaction terms between Experience and *Location*, and between *Education* and *Province*, were included to explore potential regional and educational effects on salary. The adjusted R-squared value was reduced to 74.15%, indicating that this model has comparatively weaker explanatory power than previous models.

By comparing the adjusted R-squared values across all three models, the second model emerged as the best fit. To further validate its reliability, diagnostic tests were performed on the second model.

A residuals vs. fitted plot was used to assess the linearity assumption.



In this plot, residuals are scattered around the horizontal line at zero, with no clear pattern, suggesting that the model's linearity assumption is reasonably met. A few points, such as those labeled “42” and “20,” stand out as potential outliers, but they do not appear to significantly disrupt the overall trend in the plot.

The Shapiro-Wilk test was conducted to check for normality, yielding a p-value of 0.4798, which suggests that the residuals are likely normally distributed, as the p-value is above the standard significance threshold of 0.05. The ncvTest was used to assess homoscedasticity, and it produced a p-value of 0.41229, indicating that the residuals have constant variance, thus satisfying the assumption of homoscedasticity. The Durbin-Watson test was applied to check for independence, resulting in a p-value below 0.05, indicating potential autocorrelation in the residuals and a possible violation of the independence assumption. This violation is justified, as the time-dependent variable Salary.5 was included in the second model, which may introduce autocorrelation. While normality and homoscedasticity assumptions are reasonably met, the time dependency in the model should be considered when interpreting the results.

Conclusion

In conclusion, the analysis identified the second model as the best fit due to its balance of explanatory power and reduced multicollinearity, as evidenced by its high adjusted R-squared value. By refining the predictor set, this model provides a clearer and more reliable understanding of the factors influencing salary, while diagnostic tests confirmed that most regression assumptions were reasonably met. Although the inclusion of a time-dependent variable introduced some autocorrelation, the overall strength and stability of the second model make it a suitable choice for interpreting salary determinants.

Recommendations

To improve salary modeling, economic analysts should focus on key factors like past salary, location, and province, while carefully considering the impact of time-related variables. Reducing overlaps between similar factors will make the model more stable and reliable. Incorporating combinations of demographic and geographic variables can add depth to the analysis, and models that account for variations across groups may better capture salary patterns and trends.

References

[1] Hellingman, S. (2024). Linear Regression II and Linear Regression III. ADSC 1000, Fall 2024.

Appendices

This case study was collaboratively developed by Neha Malhan and Thiwanka Abeyweera.

```
Salary_data <- read.csv("Wages.csv")

#Load necessary libraries
#install.packages("readr")
#install.packages("car")
library(readr)
library(dplyr)
library(car)


data_ds <- Salary_data %>% filter(Job == "Data Scientist")

#table(data_ds$Location)
#table(data_ds$Province)

data_ds$Province <- as.factor(data_ds$Province)
data_ds$Location <- as.factor(data_ds$Location)

#First model with all variables
first_model <- lm(Salary ~ Salary.5 + FatherEducation +
                    MotherEducation + GPA + Age + Experience +
                    Education + relevel(Location, "Metro") +
                    relevel(Province,"Ontario"), data = data_ds)


#first_model1 <- lm(Salary ~ Salary.5 + FatherEducation +
#MotherEducation + GPA + Age + Experience + Education
#                    , data = data_ds)

#vif(first_model1)


#FatherEducation & MotherEducation, Age & Experience are
#highly correlated-> based on vif
second_model <- lm(Salary ~ Salary.5 + GPA + Education +
                  relevel(Location, "Metro") +
                  relevel(Province,"Ontario"), data = data_ds)
```

```r
#Salary.5 -> time-dependent
#FatherEducation, MotherEducation, Age -> based on vif
third_model <- lm(Salary ~ GPA + Experience*relevel(Location, "Metro") +
                  Education*relevel(Province,"Ontario") +
                  relevel(Location, "Metro") +
                  relevel(Province,"Ontario"), data = data_ds)
```

```r
# Summaries of models

summary(first_model)
```

```
Call:
lm(formula = Salary ~ Salary.5 + FatherEducation + MotherEducation +
    GPA + Age + Experience + Education + relevel(Location, "Metro") +
    relevel(Province, "Ontario"), data = data_ds)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5317 -1.3316 -0.0854  1.2989  4.3818

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         13.85534    3.41025   4.063 9.37e-05 ***
Salary.5                             0.96784    0.01529  63.302  < 2e-16 ***
FatherEducation                      0.04036    0.19520   0.207  0.83658
MotherEducation                      0.08903    0.17988   0.495  0.62166
GPA                                  0.08480    0.50035   0.169  0.86574
Age                                 -0.02143    0.09646  -0.222  0.82460
Experience                           0.06766    0.09638   0.702  0.48422
Education                           -0.13972    0.10523  -1.328  0.18713
relevel(Location, "Metro")Rural     -1.22888    0.40932  -3.002  0.00335 **
relevel(Province, "Ontario")Alberta -1.62590    0.82546  -1.970  0.05151 .
relevel(Province, "Ontario")BC      -1.06763    0.61569  -1.734  0.08585 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.988 on 105 degrees of freedom
Multiple R-squared:  0.9952,    Adjusted R-squared:  0.9948
F-statistic:  2199 on 10 and 105 DF,  p-value: < 2.2e-16
```

```r
summary(second_model)
```

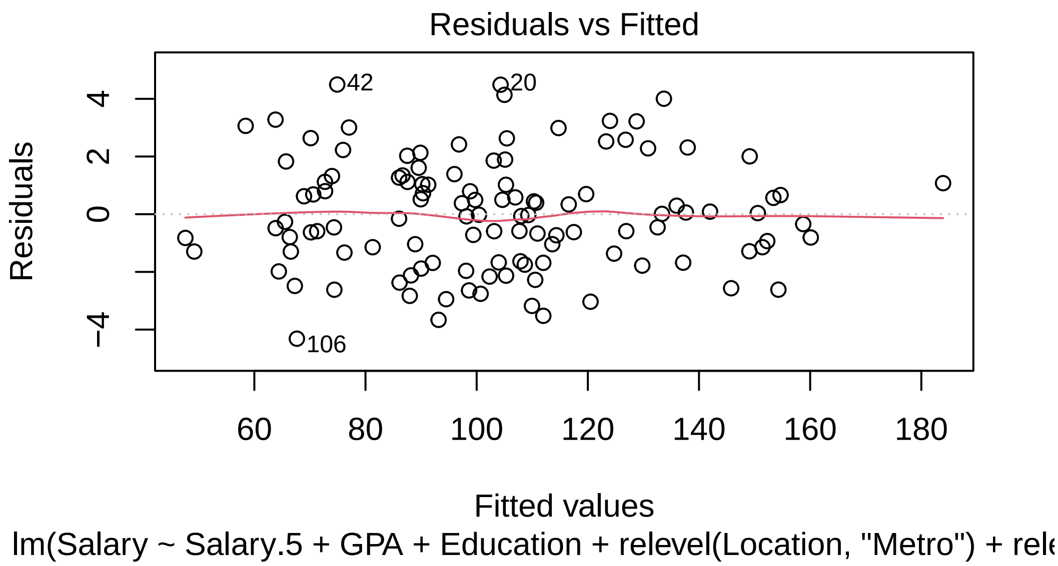
```
Call:
lm(formula = Salary ~ Salary.5 + GPA + Education + relevel(Location,
    "Metro") + relevel(Province, "Ontario"), data = data_ds)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3156 -1.3388 -0.0472  1.1585  4.4956

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         15.082692   2.283469   6.605  1.5e-09 ***
Salary.5                             0.983160   0.009791 100.410  < 2e-16 ***
GPA                                  0.071511   0.492881   0.145  0.88491
Education                           -0.171454   0.102140  -1.679  0.09609 .
relevel(Location, "Metro")Rural     -1.092166   0.380452  -2.871  0.00492 **
relevel(Province, "Ontario")Alberta -0.963354   0.638048  -1.510  0.13398
relevel(Province, "Ontario")BC      -0.683271   0.529028  -1.292  0.19924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.973 on 109 degrees of freedom
Multiple R-squared:  0.9951,    Adjusted R-squared:  0.9949
F-statistic:  3721 on 6 and 109 DF,  p-value: < 2.2e-16
```

```r
summary(third_model)
```

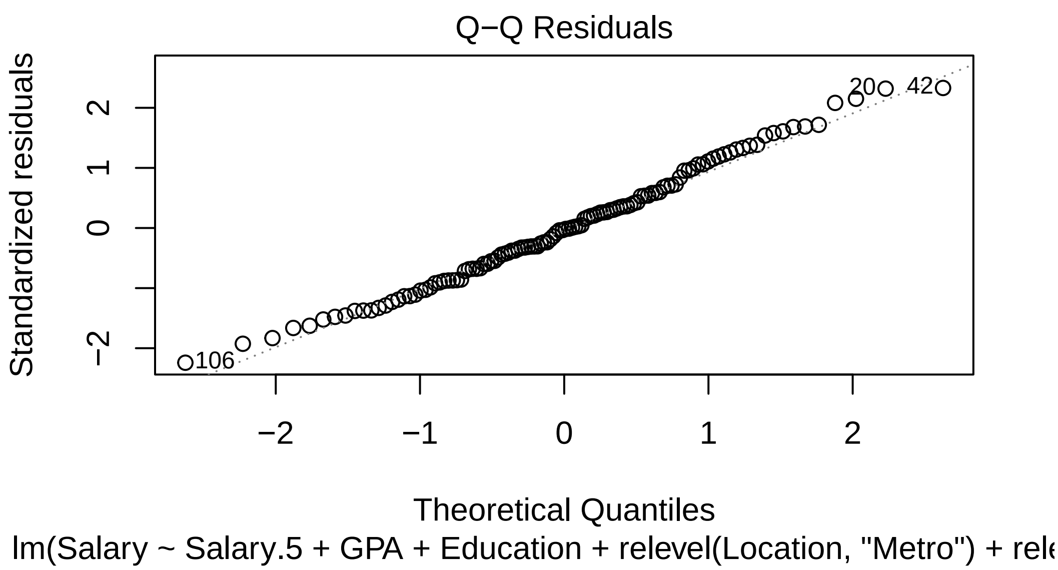
```
Call:
lm(formula = Salary ~ GPA + Experience * relevel(Location, "Metro") +
    Education * relevel(Province, "Ontario") + relevel(Location,
    "Metro") + relevel(Province, "Ontario"), data = data_ds)

Residuals:
     Min       1Q   Median       3Q      Max
-30.2644  -9.0656  -0.5201   9.7372  30.9644

Coefficients:
                                                Estimate Std. Error t value
```

```
(Intercept)                                  67.029247  20.661938   3.244
GPA                                          -0.009707   3.548981  -0.003
Experience                                    1.502990   0.312787   4.805
relevel(Location, "Metro")Rural             -25.357971   9.237146  -2.745
Education                                     2.011155   1.033754   1.945
relevel(Province, "Ontario")Alberta         -41.782141  26.681938  -1.566
relevel(Province, "Ontario")BC              -23.811427  25.521754  -0.933
Experience:relevel(Location, "Metro")Rural    0.739366   0.409997   1.803
Education:relevel(Province, "Ontario")Alberta -0.221167  1.766285  -0.125
Education:relevel(Province, "Ontario")BC     -0.223543   1.648244  -0.136
                                             Pr(>|t|)
(Intercept)                                   0.00158 **
GPA                                           0.99782
Experience                                   5.12e-06 ***
relevel(Location, "Metro")Rural               0.00710 **
Education                                     0.05436 .
relevel(Province, "Ontario")Alberta           0.12034
relevel(Province, "Ontario")BC                0.35295
Experience:relevel(Location, "Metro")Rural    0.07417 .
Education:relevel(Province, "Ontario")Alberta 0.90059
Education:relevel(Province, "Ontario")BC      0.89238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.01 on 106 degrees of freedom
Multiple R-squared:  0.7618,    Adjusted R-squared:  0.7415
F-statistic: 37.66 on 9 and 106 DF,  p-value: < 2.2e-16
```

```
# linearity #

plot(second_model,1)
```



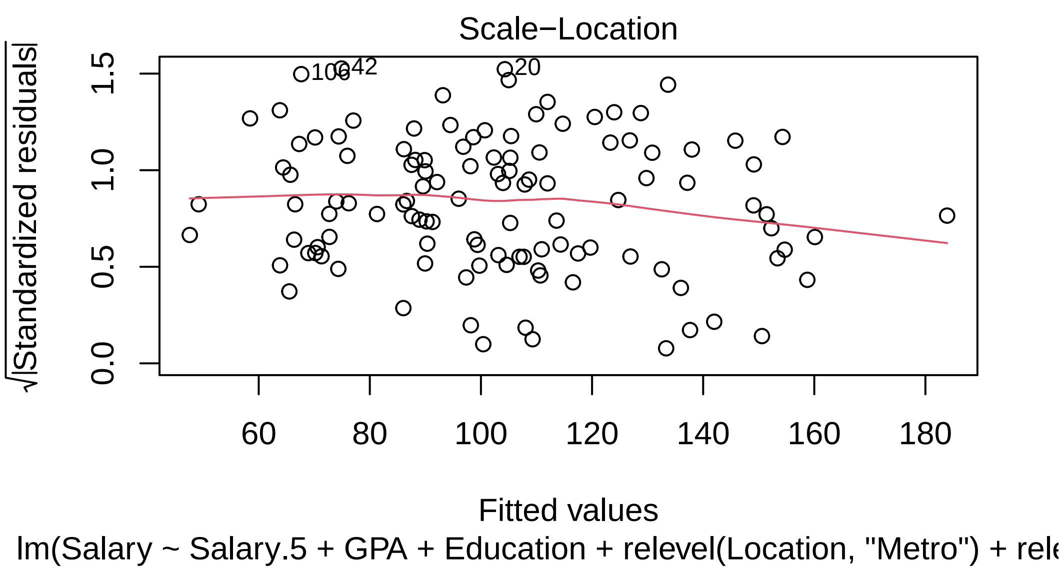

```
# Normality #
plot(second_model,2)
```



```r
shapiro.test(rstandard(second_model))
```

```
    Shapiro-Wilk normality test

data:  rstandard(second_model)
W = 0.98906, p-value = 0.4798
```

```r
# Constant Variance #
plot(second_model,3)
```




```r
ncvTest(second_model)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.6721986, Df = 1, p = 0.41229
```

```r
# Independence #

vif(second_model)
```

```
                                  GVIF Df GVIF^(1/(2*Df))
Salary.5                      2.175697  1        1.475024
GPA                           1.016909  1        1.008419
Education                     1.107491  1        1.052374
relevel(Location, "Metro")    1.078099  1        1.038315
relevel(Province, "Ontario") 2.186346  2        1.215989
```

```r
durbinWatsonTest(second_model)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.1840304      2.350736   0.052
 Alternative hypothesis: rho != 0
```