Assignment 1

JÖNKÖPING UNIVERSITY

Jönköping International Business School

JIBS

Predictive Analysis with Machine Learning

Prince Djanku -

Date of submission - 2024-09-13

Part I: Get familiar and inspect the data

As part of the assignment, the first step was to open the Salary.csv file in Python and get familiar with the variables. The dataset contains information on various factors that may influence an individual's salary, including demographic characteristics, educational background, and work experience. To begin the exploratory data analysis (EDA) process, we started by reading the Salary.csv file into a Pandas DataFrame using the Salary.read_csv() function. This allowed us to work with the data in a structured and organized manner, leveraging the powerful data manipulation and analysis capabilities of the Pandas library.

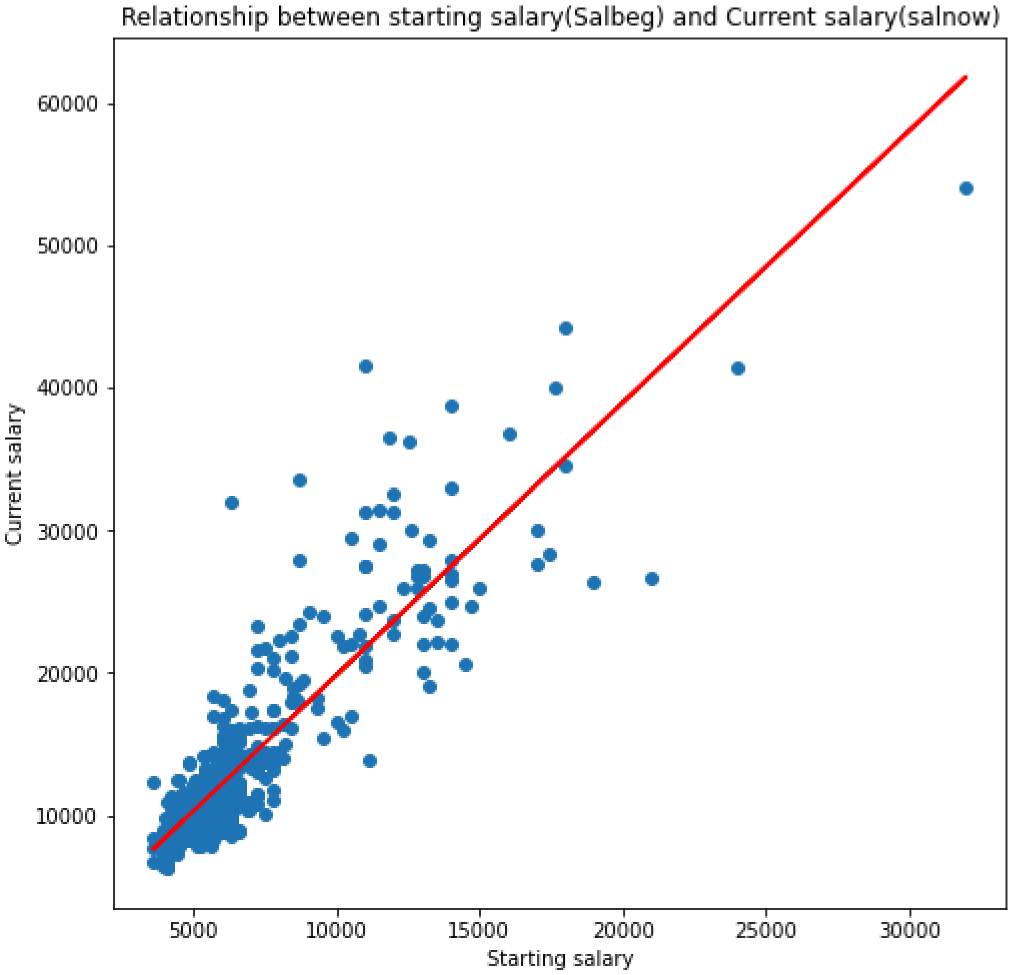
Next, we utilized several Pandas functions to gain a deeper understanding of the dataset. The Salary.head() function provided a quick overview of the first five rows of the DataFrame, giving us a glimpse into the data structure and the variables. To obtain more comprehensive information about the dataset, we used the Salary.info() function, which revealed that the dataset had no null values as well as the associated datatypes of each varaible.

To further familiarize ourselves with the dataset, we employed the Salary.describe() function, which generated summary statistics for the numeric variables, such as the mean, standard deviation, minimum, and maximum values. This information was crucial in understanding the range and distribution of the variables, which later informed the development of the prediction models. By utilizing these Pandas functions and methods, we were able to thoroughly familiarize myself with the Salary.csv dataset, laying the groundwork for the next stages of the analysis.

Part 2: The dependent variable is Salnow. Make a graphical analysis of the dependent and the independent variable you intend to include. Use scatterplot to investigate the relationship with the dependent and independent variables. Use boxplots and histograms as well and potentially other suitable graphs.

Fig. 1

Given the variables available in the salary dataset, we thought it appropriate to treat Starting Salary (salbeg) as the independent variable and Current Salary (salnow) as the dependent variable. This is because the starting salary precedes the current salary temporally, establishing a causal relationship where the starting salary influences the dependent outcome of the current salary over time. Additionally, this variable selection aligns with the practical goal of predicting current salary based on starting salary and other factors, which can provide valuable insights for employers, employees, and policymakers regarding compensation and career development decisions.

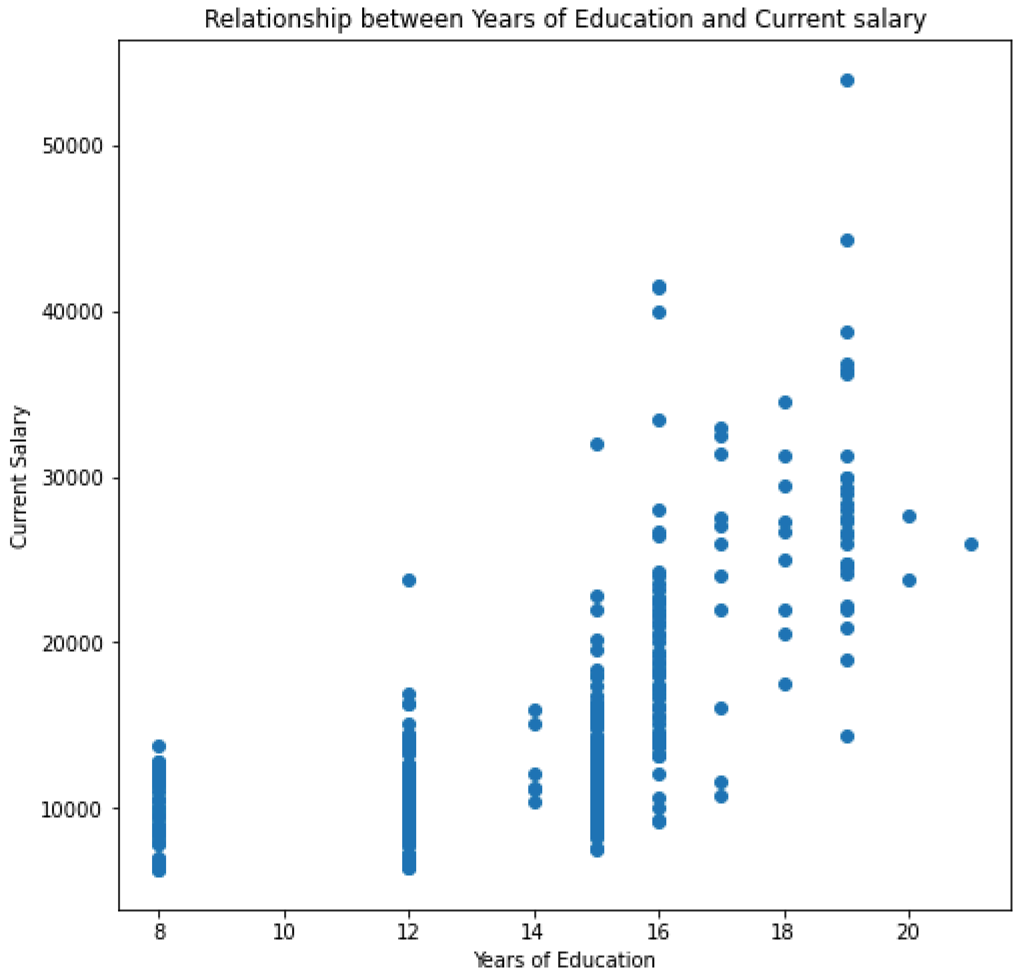


The scatter plot shows the relationship between the starting salary (salbeg) as the independent variable on the x-axis, and the current salary (salnow) as the dependent variable on the y-axis. The direction of the relationship is positive, as the points generally trend upward from left to right. This suggests that the current salary also tends to increase as the starting salary increases. The strength of this relationship appears to be moderately robust, indicated by the clustering of points around a linear model, signifying a solid linear association between the two variables.

The nature of the observed relationship is predominantly linear, evident in the data points aligning closely with a straight line rather than forming a curved, nonlinear pattern. It can be seen that as starting salaries get larger, so does the current salaries increase, leading to several potential outliers, which are evident in the scatterplot and deviate significantly from the main cluster of points. These outliers represent unusual data points where individuals have current salaries much higher or lower than expected, given their starting salaries. Overall, the scatter plot provides a helpful visualization of the positive linear relationship between starting and current salaries. The moderate strength of the relationship suggests that starting salary is an essential factor in determining current salary, but other variables may also play a role.

Fig. 2

Incorporating both a Education Level (edlevel) and Starting Salary (salbeg) as independent variables in the analysis of Current Salary (salnow) is a well-justified approach. Theoretically, education level is a key determinant of earning potential, as higher levels of education are associated with the development of more valuable skills and credentials that employers compensate for. Empirically, including education level alongside starting salary can provide incremental predictive power, as it captures additional factors that influence current salary beyond just the initial starting point. Furthermore, there may be important interactions between starting salary and education level in their joint effects on current earnings, which warrant examination. From a practical standpoint, understanding the relative contributions of these variables can inform hiring, compensation, and career development decisions for both individuals and organizations. Therefore, the inclusion of both Starting Salary (salbeg) and Education Level (edlevel) as independent variables is a sound and meaningful approach to analyzing the dependent variable of Current Salary (salnow).

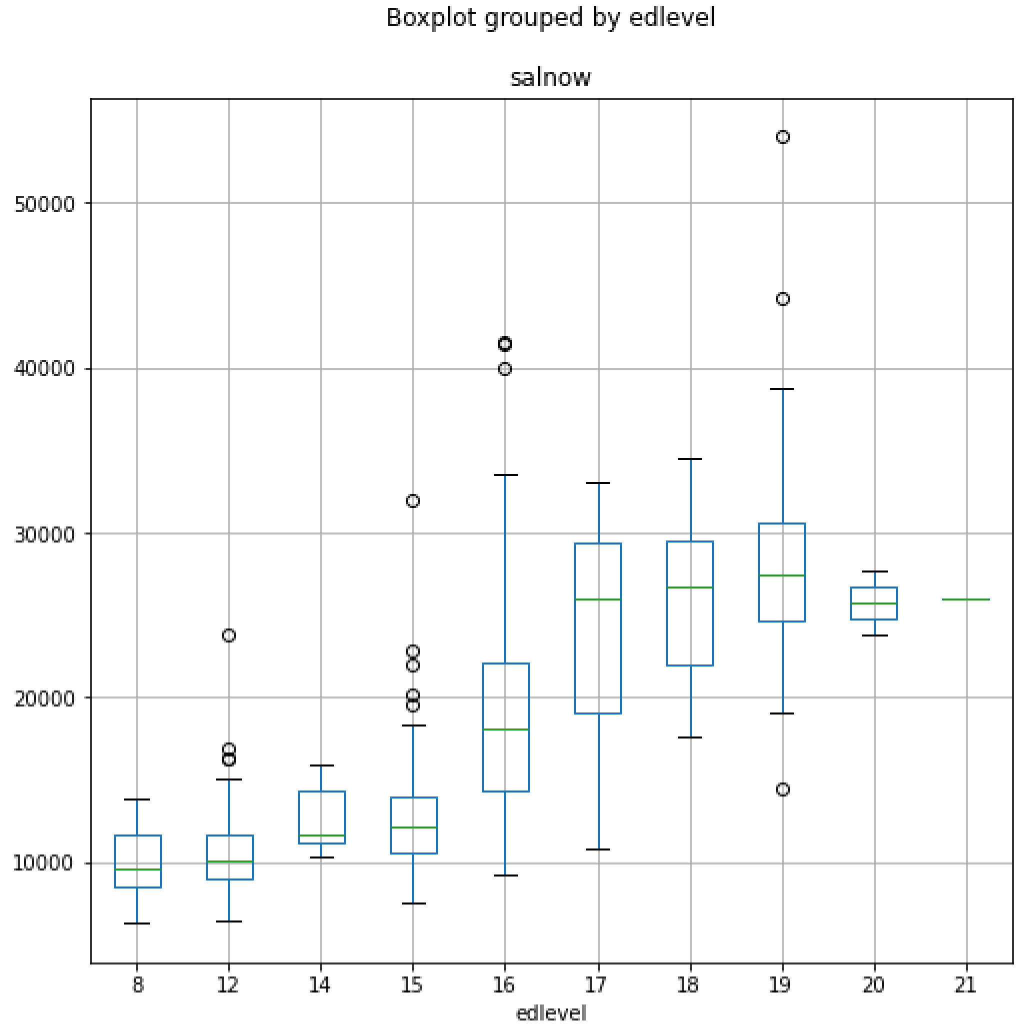


Examining the graph we can see a positive relationship between Years of Education and Current Salary. We can also observe that the relationship does seem to be non-linear as opposed to the previous relationship between the initial salary and the current salary. Here the relationship looks to be more of an exponential relationship. Meaning that at the lower levels years has a weaker positive relationship but as the years of education grows larger so does the impact it has on current salary.

The graph shows a favorable association between years of education and current salary. We can also see that the link appears to be nonlinear, as opposed to the preceding relationship between Initial Salary and Current Salary. Here, the link appears to be more of an exponential one. Meaning that at lower levels, there is a weaker positive association, but as the years of schooling increase, so does the impact on current wage.

The overall trend in the data shows a positive correlation between years of education and salary. As the number of years of education increases, the individual's current salary tends to rise as well. This suggests that investing in higher levels of education is generally associated with higher earning potential. However, the scatter plot exhibits a wide range of salary values for any given number of years of education, indicating that other factors beyond just education level also influence an individual's compensation.

Fig. 3



The boxplot shows the relationship between education level (edlevel) and Current salary(salnow). The visual representation shows a discernible positive correlation between years of education and current salary. As the number of years of education increases, the median wage is a concomitant rise. This trend aligns with human capital theory, which posits that investments in education yield returns in the form of higher earnings (Harmon et al., 2003). The median salary, represented by the horizontal line within each box, shows an uneven increase across various years.

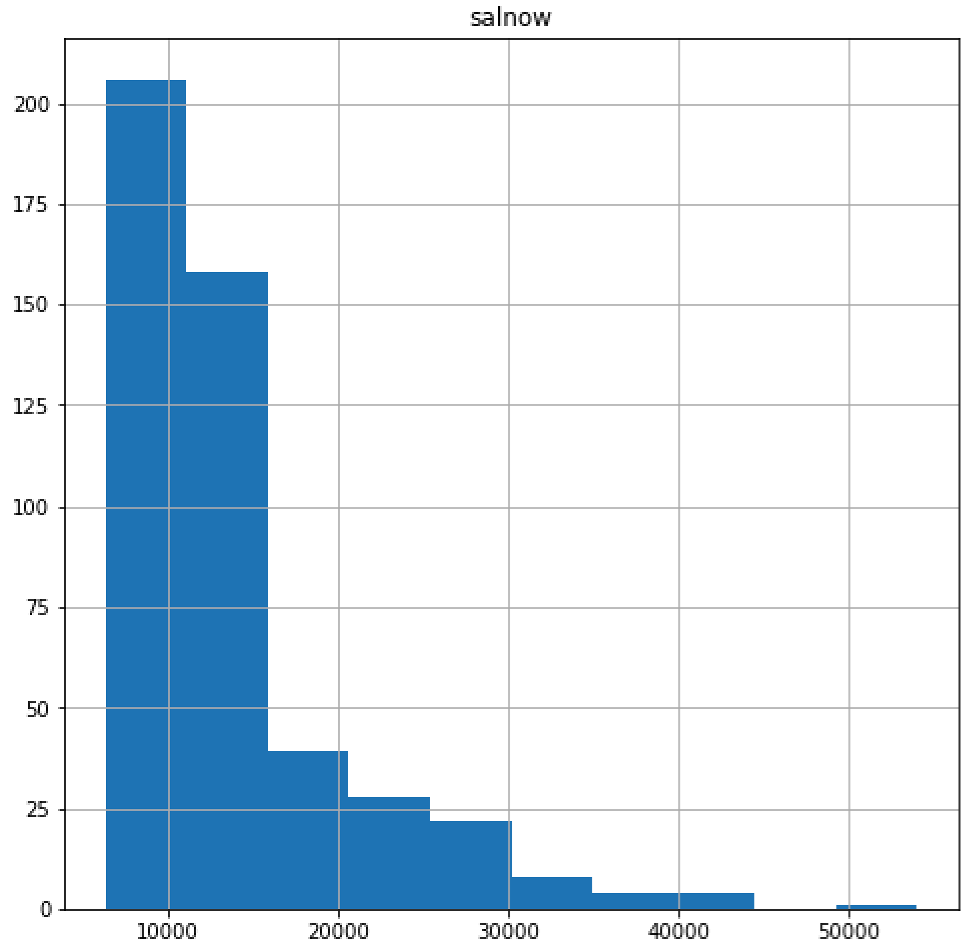
People with 8-15 years of education have the lowest median and the smallest interquartile ranges, suggesting less variability in earnings. This group also shows the presence of several high outliers, indicating that some individuals at this education level achieve salaries significantly above the average. Those with 16-19 years of education display a comparatively higher median salary and a larger interquartile range. This suggests greater salary variability among individuals with this level of education. Multiple high outliers indicate that some individuals at this level attain notably high salaries.

Individuals with 17 years of education have the highest median salary. Meanwhile, people with 19 years of school have the most extreme high outliers. These findings indicate that individuals with the highest level of education not only tend to earn more on average but also have access to the highest potential salaries. Suprisingly, we observe that there might be certain thresholds where additional years of education yield diminishing returns in terms of salary enhancement. This phenomena is observable for those who have 20 years of education

Despite the overall upward trend, there is considerable overlap in the salary ranges across different education years. This overlap suggests that while education is a significant factor in determining salary, other variables also likely play essential roles.

Outliers across several educational years indicate that exceptional earners exist across all education years categories. However, the frequency and extremity of these high-value outliers increase with higher educational years.

Fig 4



In the realm of statistical analysis, the histogram has emerged as the fundamental tool for visualizing the distribution of dependent variables (Thaper et al., 2002). This is because they provide a powerful means of gaining insights into the underlying characteristics of data.

For instance the histogram above exhibits a distinct right-skewed distribution, indicating that the majority of the current salary values are concentrated on the lower end of the salary spectrum with a smaller number of higher salary values. This suggests that the salary distribution is not symmetrical or normally distributed around the mean or average salary value. It can be observed that the plot depicts a peak or mode salary value of around 10 000. This means that the salary value is the most occurring value in the data set. The rapid decline in frequency from the peak to higher values shows that higher salary values are much less prevalent in the salary dataset.

Additionally, the plot shows a series of smaller secondary peaks at various salary values which suggest the presence of distinct clusters within the dataset.

Part3: Find a prediction model where Salnow is the dependent variable.
Consider economic/Business administration theory when you develop the model and which variables to include. Include non-linear and interaction terms where applicable. Think of functional form (logarithm transformationetc.) as well.
Evaluate both in-sample statistics and out-of-sample (i.e.out-of sample predictions on a validation set) when developing your model.

As mentioned in Part 2, we chose to use starting salary (salbeg) and educational level (edlevel) as dependent variables to predict current salary (snow) because we found these variables to have a strong R-squared in relation to current salary.

From an economic perspective, we believe these two variables are significant predictors of current salary (snow) as they reflect key factors that commonly influence earning potential over time.

Starting Salary (salbeg): The initial salary at the beginning of a career often sets a foundation for future earnings. It captures the market value of an individual's skill set and experience at the start of their career. Over time, while promotions, experience, and additional skills can influence salary growth, the initial starting point often acts as a baseline from which future increases are negotiated.

Educational Level (edlevel): Education has long been established as a key determinant of earning potential. Individuals with higher educational qualifications have more opportunities for career progression and are more likely to receive salary increases over time. Higher education also signals specialized knowledge, which is often rewarded with higher compensation. Thus, educational level serves as a proxy for an individual's qualifications, skills, and potential value to employers, making it a crucial variable in predicting current salary.

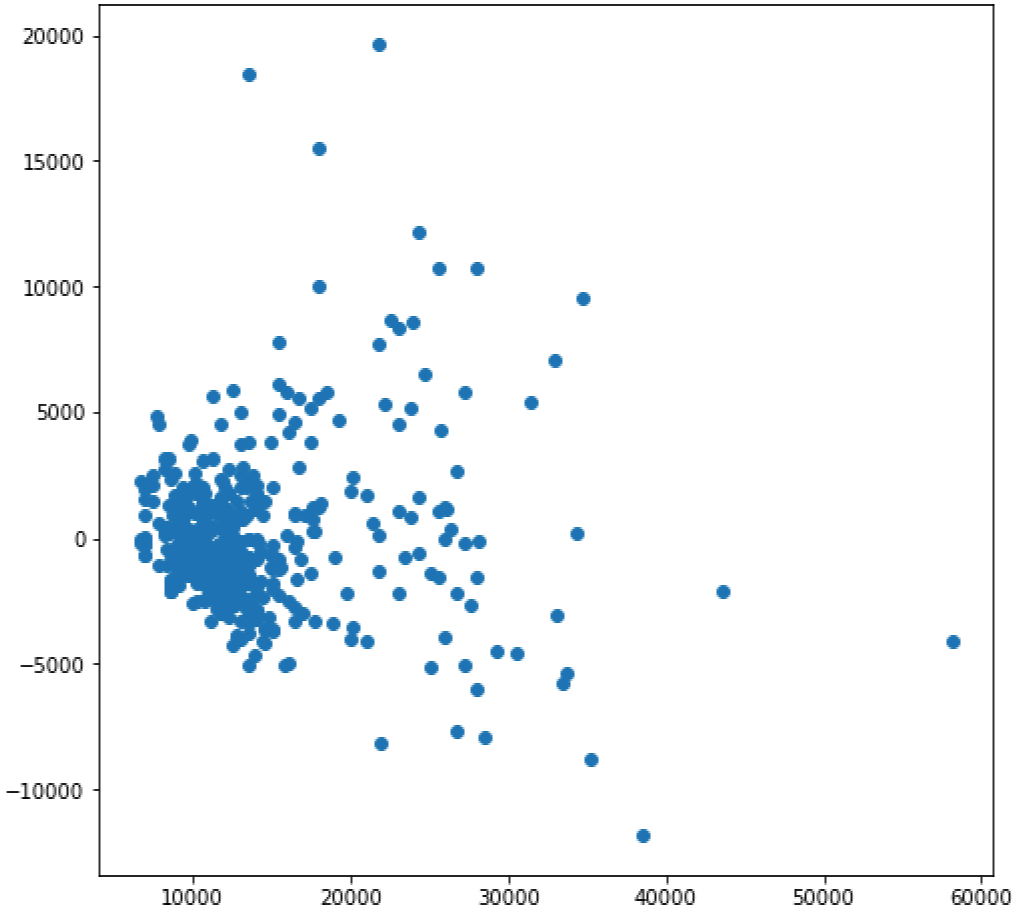
Together, starting salary and educational level provide a robust economic foundation for understanding salary progression. These variables capture both the initial valuation of an individual’s skills at the start of their career and the potential for future earnings growth through education and career advancement.

After conducting various regression analyses—multiple linear regression, multiple linear regression with squared terms, and multiple linear regression incorporating both squared and cross-multiplied variables—we observed the following R-squared values:

- R-squared with squared and cross-multiplied variables: 0.8002815957259627
- R-squared with squared variables: 0.7999683971411102
- R-squared with standard multiple linear regression: 0.791644740799016

These results suggest that including squared and cross-product terms marginally improves the model's explanatory power, as indicated by the higher R-squared value. However, even the standard multiple linear regression yields a reasonable fit, suggesting that starting salary and educational level are relatively effective predictors of current salary.

After this, we ran a diagnostic plot to assess the validity and performance of our regression models. The diagnostic plots show that our prediction model becomes less accurate when predicting higher salary values, as the plots begin to spread further apart from each other. In contrast, the plots are closer and more compact when interpreting smaller salary values.



Following this, we split our total data of 470 individuals into two sets, each consisting of 235 individuals, giving us a training set and a validation set. This split was made at random.

We then conducted multiple linear regression analyses using our training set data on two different models:

Model 1: A standard multiple linear regression with starting salary and educational level as predictors.
Model 2: A squared multiple linear regression, which includes the squared terms for starting salary and educational level, in addition to the original variables.

Model 1: Standard Multiple Linear Regression

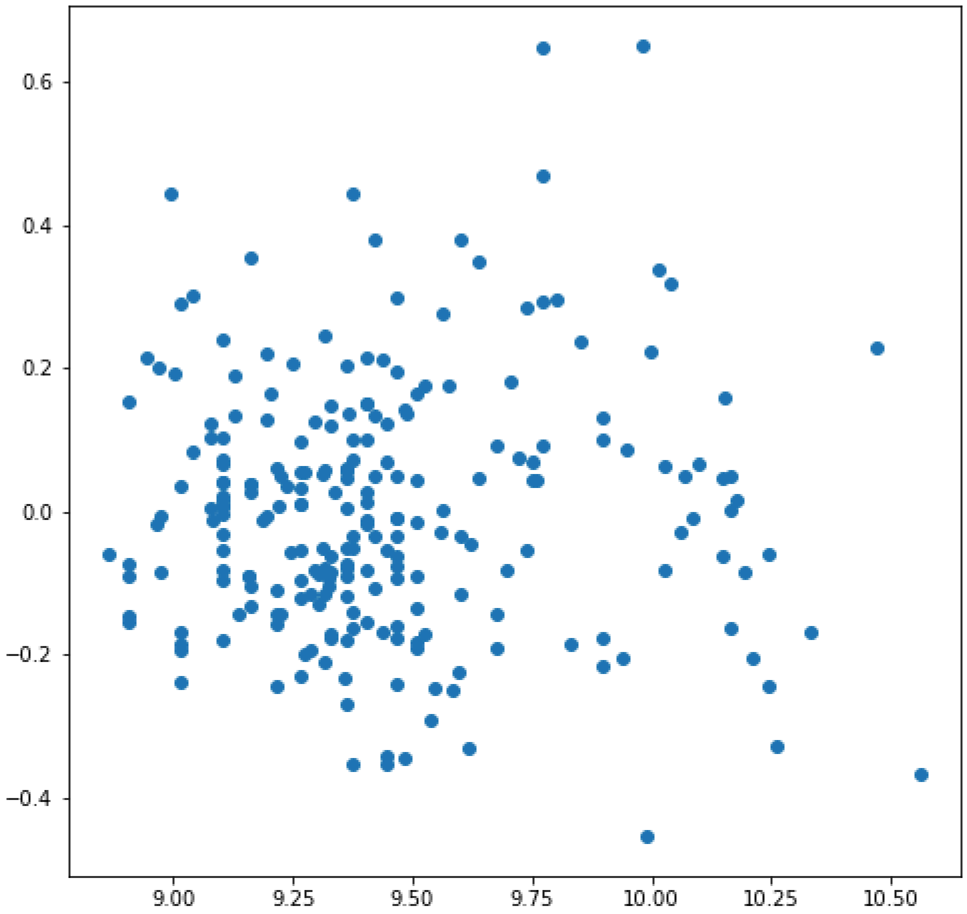
- Training data performance (R^2): 0.742149402383157
- Mean Error (ME): 205.276341935273
- Mean Absolute Error (MAE): 2103.018352307298
- Mean Squared Error (MSE): 9556147.429394409

Model 2: Squared Multiple Linear Regression

- Training data performance (R^2): 0.7563416299381968
- Mean Error (ME): 423.86235226614434
- Mean Absolute Error (MAE): 2206.0012070538564
- Mean Squared Error (MSE): 12494550.009358544

The squared multiple linear regression slightly improves the R^2 value compared to the standard multiple linear regression, indicating a better overall fit to the data. However, this comes at the cost of increased mean error, mean absolute error, and mean squared error, suggesting that the model's predictions might have larger individual errors despite its improved fit.

Given this, we decided to use the logarithmic transformation of our data to see if it would improve our results. This produced the following graph:



The graph representing the logarithmic dataset shows a more balanced distribution compared to previous plots. While there is still a slight concentration of values in the lower range, this concentration is significantly reduced, indicating a more even data spread.

We then recalculated linear regression metrics such as mean error, absolute mean error, and mean squared error for the logarithmic dataset, giving us the following results:

Model 1: Logarithmic Multiple Linear Regression

- Training data performance (R^2): 00.7867936009716338

- Log Mean Error (ME): 0.0151871907596807
- Log Mean Absolute Error (MAE): 0.14269707403308007
- Log Mean Squared Error (MSE): 0.032738813052853026

Model 2: Logarithmic Squared Multiple Linear Regression

- Training data performance (R^2): 0.7867936009716339
- Log Mean Error (ME): 0.01518719075968064
- Log Mean Absolute Error (MAE): 0.14269707403308007
- Log Mean Squared Error (MSE): 0.03273881305285299

The logarithmic multiple linear regression (Model 1) and the logarithmic squared multiple linear regression (Model 2) produce nearly identical R^2 values and error metrics. This indicates that introducing squared terms in the logarithmic model does not significantly improve the model's fit or prediction accuracy, as both models perform similarly across all metrics.

Further calculations were necessary to determine whether the logarithmic dataset provides a better model than the non-logarithmic dataset. Due to the difference in scales between the two models, comparing the mean errors directly was not possible. We recalculated the logarithmic data back into the same scale as the original non-logarithmic data, yielding the following results:

Model 1: Exponential Logarithmic Multiple Linear Regression

- Exponential Log Mean Error (ME): 406.34797367623713
- Exponential Log Mean Absolute Error (MAE): 2066.161660827987
- Exponential Log Mean Squared Error (MSE): 9633176.980719998

Model 2: Exponential Logarithmic Squared Multiple Linear Regression

- Exponential Log Mean Error (ME): 406.34797367623713
- Exponential Log Mean Absolute Error (MAE): 2066.161660827987
- Exponential Log Mean Squared Error (MSE): 9633176.980719998

The recalculated results show no differences between the two logarithmic models. Indicating that adding the squared variable in this scenario has no real impact on the prediction model.

Analyzing the performance of our four models reveals key insights into their predictive capabilities.

Model 1: Standard Multiple Linear Regression

- Predictors: Starting Salary (salbeg), Educational Level (edlevel)
- R-squared: 0.742149402383157
- Mean Error (ME): 205.276341935273
- Mean Absolute Error (MAE): 2103.018352307298
- Mean Squared Error (MSE): 9 556 147.429394409

This is a straightforward model using starting salary and education level as predictors. While it explains about 74.2% of the variation in current salary, the error metrics show room for improvement.

Model 2: Squared Multiple Linear Regression

- Predictors: Starting Salary (salbeg), Educational Level (edlevel), plus their squared terms
- R-squared: 0.7563416299381968
- Mean Error (ME): 423.86235226614434
- Mean Absolute Error (MAE): 2 206.0012070538564
- Mean Squared Error (MSE): 12 494 550.009358544

Introducing squared terms slightly increases the R-squared value, meaning the model better fits the data. However, the error metrics, especially ME, MAE, and MSE, increase, indicating that individual predictions may be less accurate, potentially due to overfitting.

Model 3: Logarithmic Multiple Linear Regression

- R-squared: 0.7867936009716338
- Log Mean Error (ME): 0.0151871907596807
- Log Mean Absolute Error (MAE): 0.14269707403308007
- Log Mean Squared Error (MSE): 0.032738813052853026
- Exponential Log Mean Error (ME): 406.34797367623713
- Exponential Log Mean Absolute Error (MAE): 2 066.161660827987
- Exponential Log Mean Squared Error (MSE): 9 633 176.980719998

This model transforms the data using logarithmic scaling and performs better in terms of error distribution. The R-squared is higher than the standard model (0.787), and when recalculated into the original scale, its performance (in terms of errors) improves, making it a more effective model for predicting current salary.

Model 4: Logarithmic Squared Multiple Linear Regression

- R-squared: 0.7867936009716339
- Log Mean Error (ME): 0.01518719075968064
- Log Mean Absolute Error (MAE): 0.14269707403308007
- Log Mean Squared Error (MSE): 0.03273881305285299
- Exponential Log Mean Error (ME): 406.34797367623713
- Exponential Log Mean Absolute Error (MAE): 2 066.161660827987
- Exponential Log Mean Squared Error (MSE): 9 633 176.980719998

This model adds squared terms to the logarithmic model. While it has virtually the same R-squared value as the logarithmic model (0.787), introducing squared terms does not improve the model and leads to similar error values as Model 3.

Comparing these models shows that Model 4 has the highest R-squared value at 0.7868, barely surpassing Model 3. However, Models 3 and 4 are practically identical; their differences are negligible. Aslo other models are not that far of at 0.7563, and 0.742.

Model 1 stands out with the lowest Mean Error (ME) at approximately 205, while the other models exceed 400, suggesting that Model 1 offers more accurate individual predictions in terms of error magnitude.

Looking at the Mean Absolute Error (MAE), Models 3 and 4 slightly outperform the others, with a value of 2066, followed by Model 1 at 2103, and Model 2 at 2206. This shows that the difference in MAE is minor, but Models 3 and 4 provide slightly better performance in absolute prediction accuracy.

Finally, analyzing the Mean Squared Error (MSE) reveals that Model 1 again has the lowest value at 9,556,147, closely followed by Models 3 and 4 at 9,633,176. Model 2, however, shows a significant discrepancy with an MSE of 12,494,550, indicating that Model 2 introduces larger individual prediction errors, likely due to the squared terms, which may contribute to overfitting.

In summary, Model 1 minimizes prediction errors, while Models 3 and 4 seams to offer the best overall balance between fit and prediction accuracy. Model 2, despite having a higher R-squared than model 1, shows weaker performance in error metrics, suggesting that adding squared terms does not improve prediction quality, which is further demonstrated as model 4, which adds squared variables to model 3 and barely changes model 3.

References

Harmon, C., Oosterbeek, H., & Walker, I. (2003). The Returns to Education: Microeconomics. *Journal of Economic Surveys*, *17*(2), 115–156. https://doi.org/10.1111/1467-6419.00191

Thaper, N., Guha, S., Indyk, P., & Koudas, N. (2002). Dynamic multidimensional histograms. , 428-439. https://doi.org/10.1145/564691.564741.