2/25/2024

**Salary Prediction- Regression Project**

**Multivariate Statistics | STAT8031**

**Name: Kina Gajipara**

**Student Number: 8850437**

**Professor: Tao Xu**

**Table of Contents**

Introduction ……………………………………………………………………………….. 2

Dataset Descriptions ………………………………………………………………………. 3

Objective …………………………………………………………………………………... 3

Data Preprocessing ………………………………………………………………………… 3

Exploratory Data Analysis and Visualizations ……………………………………………. 4

Feature Selection, Model Building and Evaluation ……………………………………….. 7

Recommendations …………………………………………………………………………. 11

Conclusion …………………………………………………………………………………. 12

**Introduction**

Salaries of professors can vary significantly based on various factors, and understanding these variations is crucial for individuals and institutions alike. This project focuses on analyzing the salaries of professors and identifying the key factors that influence their pay. Using a dataset of 397 observations, we aim to create a predictive model to estimate salaries based on predictors such as years of experience, rank, discipline, years since earning a Ph.D., and gender.

The primary goal of this analysis is to provide insights into the salary structure of professors and develop a model to accurately predict their earnings. Such a model can help individuals plan their careers and salaries more effectively, while institutions can use it to ensure fair and competitive compensation policies. By considering both quantitative and categorical factors, the analysis seeks to uncover patterns and disparities in salaries, offering a data-driven perspective on wage distribution.

This project also highlights the importance of identifying any salary gaps and ensuring transparency in wage policies. While factors such as rank and years of experience are naturally expected to influence salaries, the dataset allows us to explore the potential effects of other predictors. Ultimately, this analysis will help make informed decisions about salary policies and guide future research into academic wage structures.

**Dataset Description:**

The dataset gives information about the how much professor have been earning. Dataset has some predictors which influence the salary of professor. These predictors include Discipline, Years of experience, Years since PhD, Sex, Rank. There are around 397 observations in this dataset to predict salary.

**Predictors:**

**Discipline**  : A categorical variable indicates Discipline of professor in A and B grade.

**Years of experience:** This is numerical variable. it shows how many years the professor has

been working in academy.

**Years since PhD** : A quantitative variable tells us how many years it's been since the

professor earned their Ph.D. degree.

**Sex** : This is categorical variable specifies the gender of the professor, whether

male or female.

**Rank** : Another categorical variable indicates the rank of the professor, such as

professor, assistant professor, etc.

**Response Variable:**

Response variable of this dataset is "Salary". This numerical variable represents the annual salary of each professor included in the dataset.

* In predicting salary of professor, all predictors mentioned above are important. As Discipline represents soft skill of professors, salary varying due to these reasons. Rank is another predictor, higher rank demand higher salary so, it is also crucial. Years of experience and years since PhD is also important, which correlate with salary. However, sex is not relevant that much, because equal pay for equal work, but it is historical data so, better to consider it as predictor.

**Objective:**

The primary aim of this analysis is to develop a prediction model of salaries of professional based on some predictors. Using some software and statistical technique, our aim is to construct a model to estimate accurate salary. Additionally, this model assist individual to predict their salary and plan their career effectively. Moreover, it also helps institutions to make informed decisions regarding its wages considering its Quality.

**Data Preprocessing:**

* Loading a dataset and saved as a “dt”
* Checking a missing value.
* Checking for the duplication, we found 4 duplicates but it’s not exactly same.
* Remove one unnecessary column, which shows index only.

**Exploratory Data Analytics and Visualizations:**

**Summary:**

Summary helps us to analyse each numerical variable which help to make predictions.

A screenshot of a computer screen

Description automatically generated

**Key Insights:**

* Most individuals fall within the 12 to 32 years since PhD range (based on quartiles), indicating a generally experienced workforce.
* Most people earn between **$91,000** and **$134,185**, showing that many salaries fall in the mid-range.

**Distribution of Salary:**

**Code:** dt %>% ggplot(aes(salary)) + geom\_histogram()

A graph of a graph

Description automatically generated

Here, salary is the response variable. And it is necessary to analyse a graph of it so, that help us to make predictions.

**Key Insights:**

* Most people earn between $91,000 and $134,185, showing that many salaries fall in the mid-range.
* The distribution is **right-skewed**, with a smaller number of individuals earning significantly higher salaries (above $200,000).
* While most salaries cluster around the mid-range, there is notable variability, with a range extending from around **$50,000 to over $200,000**.

Now, we check correlations between response variable and explanatory variables. For that we do factorization to convert categorical variable into numerical variable.

A red and blue squares with black text

Description automatically generated

**Key Insights:**

* Years Since PhD and Years of Service are Closely Related it means the longer someone has had their PhD, the longer they’ve worked in their job.
* People with higher ranks and more years since their PhD or in service get higher salaries.
* Gender (Sex) Doesn’t Strongly Affect Other Factors, there is no strong relation with any factors.

We use Box plots for comparing distributions across multiple categories of Rank, Gender and Discipline, making it easier to see differences and patterns using quartiles.

**Salary Distribution Across Rank:**

A screenshot of a computer

Description automatically generated

**Key Insights:**

* Professors have the highest median salaries, and their salaries vary significantly, with some earning much more (outliers).
* Associate Professors have higher median salaries than Assistant Professors but lower than full Professors, with less variation compared to Professors.
* Assistant Professors have the lowest median salary, with a smaller range of salaries compared to other ranks.

**Salary Distribution Across Sex:**A screenshot of a computer screen

Description automatically generated

**Key Insights:**

* The median salary for males is higher compared to females.
* Male salaries show more variation, with higher maximum values and some outliers.
* Female salaries are more concentrated, with less variation compared to males.

**Salary Distribution across Discipline:**

A screenshot of a computer

Description automatically generated

**Key Insights:**

* The median salary for discipline B is slightly higher than that of discipline A, as shown by the higher central line in the box for discipline B.
* Both disciplines have a wide salary range, but discipline B shows slightly more variability in salaries, with a longer box and whiskers.
* Both disciplines have outliers on the higher end, but discipline A appears to have more prominent high-salary outliers compared to discipline B.

**Feature Selection, Model Building and Evaluation:**

The dataset was split into training (70%) and testing (30%) subsets to ensure proper model training and evaluation.

To find the best factors that predict salary, we used subset selection on the training data. This method tests different combinations of factors to identify the model that best explains the changes in salary.

Using the regsubsets function, we generated models with various subsets of predictors and evaluated their performance using the Adjusted R-squared criterion. Adjusted R-squared accounts for model complexity and helps avoid overfitting.

A screen shot of a computer

Description automatically generated

A plot of Adjusted R-squared values for different models was generated. The Adjusted R-squared indicates the proportion of variance explained by the model, adjusted for the number of predictors included.

* The model with all predictors gives the highest performance based on Adjusted R-squared.
* Including more relevant variables in the model improves its ability to explain salary variations, with the best result coming from using all predictors together.

The best model includes all predictors. This suggests that each variable contributes to explaining variations in salary and should be included in the final predictive model.

**Model Training using cross validation and evaluation:**

A linear regression model is trained using a selected variables and the training dataset. The linear regression model was trained using 10-fold cross-validation to ensure robust performance evaluation.

The model is tested on the test dataset, which was not used for training or cross-validation, to evaluate its performance on unseen data. Predictions are made for the test data, and error metrics (RMSE and MAE) are calculated to assess how well the model generalizes.

* The average Root Mean Square Error (RMSE) was **23,200.5**, indicating the typical error magnitude between predicted and actual salaries during cross-validation.
* The R-squared value was **0.402**, meaning the model explains about 40.2% of the variance in salary.
* The Mean Absolute Error (MAE) was **18,401.67**, representing the average absolute difference between predicted and actual salaries.

**Comparison of Predicted vs actual value:**

**A graph with a red line and black dots

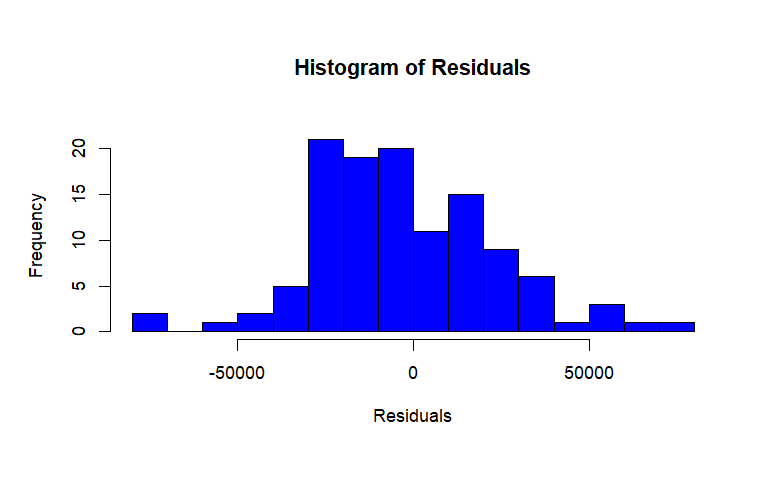
Description automatically generated**

**Key Insights:**

* Many points lie near the diagonal red line (perfect prediction line), indicating that the model predicts salaries reasonably well for a significant portion of the data.
* There is noticeable deviation for higher actual salaries, suggesting the model struggles to accurately predict at the upper range of the salary distribution. This may indicate potential underfitting or missing predictors that better explain higher salaries.
* The scatter points do not show a clear systematic pattern, implying that the model's errors are relatively unbiased across different salary ranges. However, further residual analysis would confirm this.

**Histogram of Residuals:**

The histogram provides another way to assess the distribution of residuals.

****

**Key Insights:**

* The residuals are symmetrically distributed around zero, which supports the assumption of normality.
* There are no extreme outliers, and the distribution appears approximately bell-shaped, further confirming that the residuals are normally distributed.

**QQ Plot of Residual:**

The QQ plot assesses the normality of residuals.

A graph of a number of individuals

Description automatically generated with medium confidence

**Key Insights:**

* The points mostly align with the diagonal line, indicating that the residuals are approximately normally distributed.
* However, some deviations at the tails suggest potential mild non-normality, especially at the extremes.

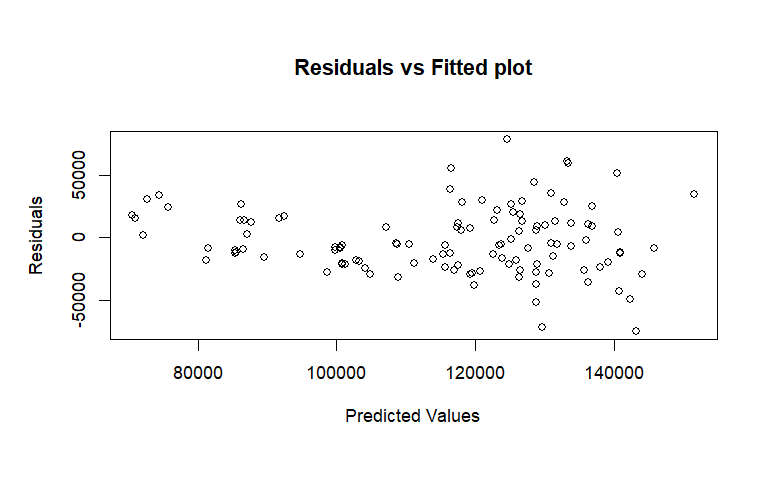
**Statistical Confirmation of normality: Shapiro-Wilk Test for Normality**

The Shapiro-Wilk test assesses whether the residuals follow a normal distribution. The p-value is 0.0696, which is greater than the significance level of 0.05. Since the p-value > 0.05, we fail to reject the null hypothesis, meaning the residuals are approximately normal.

The Shapiro-Wilk test confirms that the residuals are approximately normal (p-value = 0.0696), supporting the normality assumption required for the model.

**Residual vs Fitted plot:**

This plot checks for heteroscedasticity and patterns in residuals.



**Key Insights:**

* The residuals appear randomly scattered around the horizontal line (zero), indicating no clear pattern, which suggests that the model fits the data reasonably well.
* There is no obvious funnel shape, so heteroscedasticity (non-constant variance) is likely not a significant issue.

**Recommendations:**

* All predictors play a role in explaining salary differences. These factors should be carefully considered when designing salary policies or predicting salaries. Even though gender (Sex) does not strongly affect other variables, it is still helpful to include it for historical context.
* While the model generally fits the data well, slight deviations in residuals suggest that other unknown factors may influence salaries. Regularly updating the model with new data and predictors can improve accuracy.
* Gather data on regional salary variations and institutional reputation for a more comprehensive model.
* Analyze salary trends over time to account for inflation, policy changes, and institutional priorities.
* Collect more data for more accurate predictions. Most of the data are in mid range, but few of them earn much higher salaries and may be model struggles to predict higher salaries.
* Conduct qualitative research to understand why salary disparities might exist between genders despite similar qualifications and ranks.

.

**Conclusion:**

This analysis provides important insights into the factors that affect professor salaries. Rank, Years Since PhD, and Years of Experience are the strongest predictors of salary. Professors, who hold the highest rank, earn the most and have the widest salary range, while Assistant Professors earn the least and have less variation in their salaries. Sex and Discipline also show some differences in salaries, but their impact is smaller compared to other factors. Professors in Discipline B earn slightly more than those in Discipline A, and male professors tend to have slightly higher median salaries than females.

Salaries are not evenly distributed; most are in the mid-range, but a few professors earn much higher salaries. The prediction model explained about 40.2% of the variation in salaries and had an average error (RMSE) of around $23,200. While the model works well for predicting mid-range salaries, it struggles to predict very high salaries accurately. This suggests that some important factors, like research output or teaching performance, might be missing from the data. The analysis also showed that the assumptions of the model, like normality and equal variance, were met, which adds confidence to its reliability.

However, the model has limitations. It doesn't explain all the salary variation, and it doesn’t capture high salaries very well. Adding more data, like publications or institutional reputation, could improve its performance. While the model found differences in salaries based on gender, the reasons behind these differences weren’t clear, which means more investigation is needed to ensure fairness.

In summary, the analysis highlights that rank and experience play the biggest roles in determining professor salaries. Institutions can use this information to create fair salary policies. Future studies should include more data and try advanced methods to better understand and predict salaries, especially for those at the higher end.