**Modeling & assessing pay** **disparity**

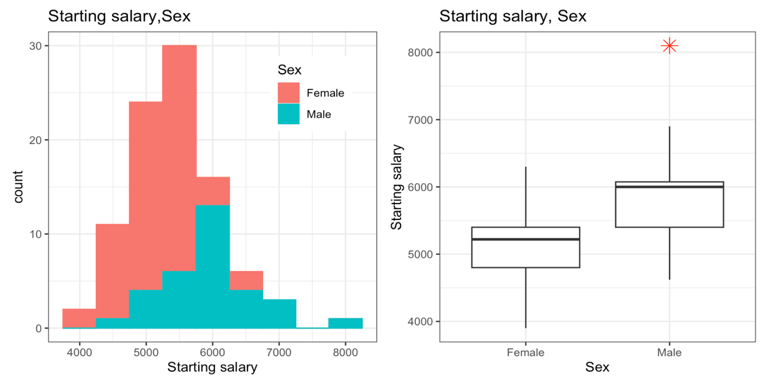
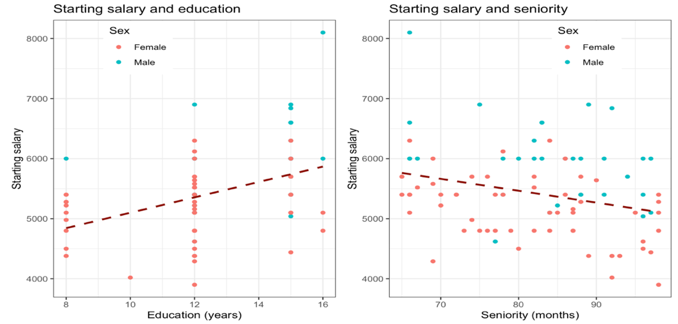
COMP 4441 Introduction to Probability and Statistics for Data science

Dawnena Key, Katrin Jõgi

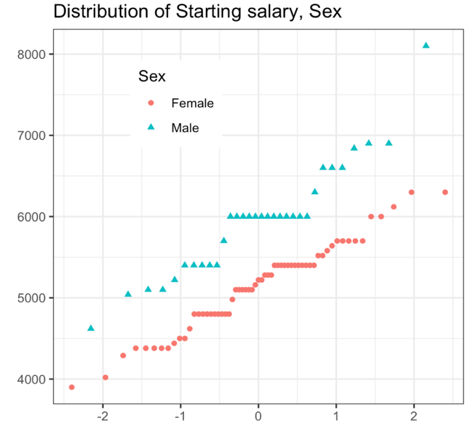
The goal of this analysis is to determine if there is evidence of pay disparity between men and women at a bank using regression modeling. In addition to testing for statistical differences, an important objective is to perform a power analysis to assess if the sample size is adequate to reliably detect gender discrimination. Proving pay discrimination requires sufficient statistical power, so estimating the required sample size and evaluating the study data against this benchmark is a key component. Using regression modeling to control for appropriate factors, the power analysis aims to provide greater confidence in conclusions regarding gender pay equity at this bank. This report details the regression analysis, power calculations and evaluation of the results in the context of the available data. By including power considerations along with robust statistical tests, this work provides a thorough investigation into potential pay discrimination allegations at the bank under this study.

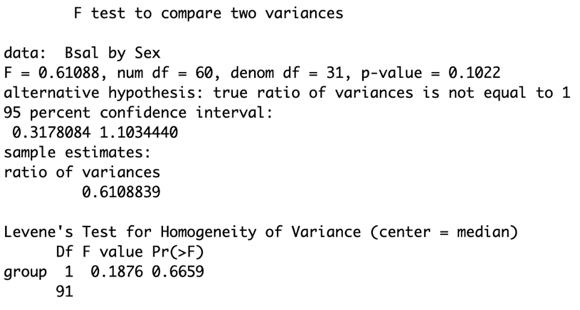
The method of multiple regression is used in the project to find statistical evidence for possible discrimination. Regression analysis explores the relationships between the input variables and the response variable, in this case, the starting salary. Using regression provides controls for other factors while quantifying the magnitude of impact. It is also flexible to model nonlinearity. A Student’s *t* test is intrinsically linear.

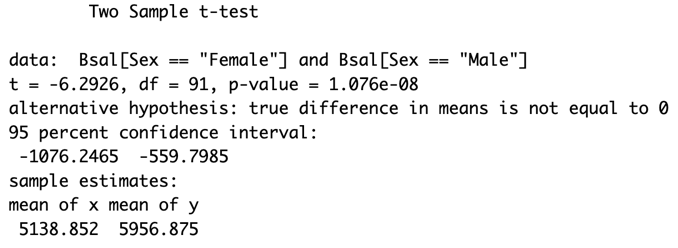
The data {Slueth3 – case1201}[[1]](#footnote-1) is on employees from one job category (skilled, entry–level clerical) of a bank that was sued for sex discrimination. It is a data frame with 93 observations on 7 variables: Bsal – annual salary at time of hire, Sal77 – salary as of March 1975, Sex – sex of employee, Senior – seniority (months since first hired), Age – age of employee (in months), Educ – education (in years), Exper – work experience prior to employment with the bank (months). For the purposes of the current analysis, we omitted Sal77.

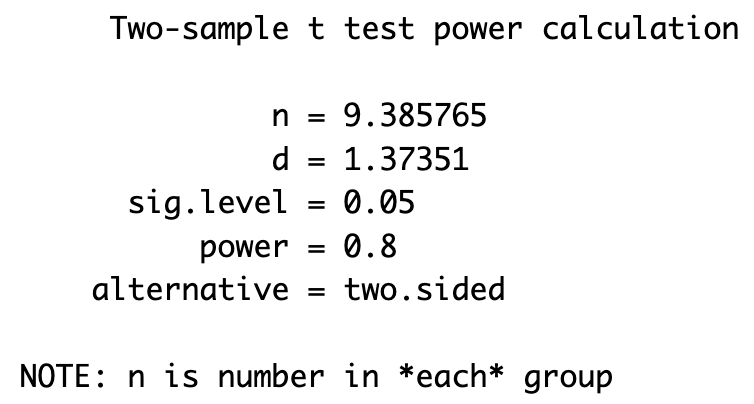
The data are on 32 male and 61 female employees, hired between 1965 and 1975. Women in the study population have on average lower education level than men, 12 and 13.5 years respectively. Out of the 12 workers with only 8 years of education, only one is male. Positive association between education and starting salary is observed. Negative association between seniority and starting salary is observed, which may indicate that the starting salaries increased for everyone over time.

The spread and medians of starting salary for men and women are different. For men, the spread is wider, with a high end outlier and median at 6,000, while the women’s plot exhibits a longer lower tail and median at 5,220.

Non-normality, outliers and serial correlation can all invalidate inferences made by standard parametric test, such as Student’s *t* test. The histogram on starting salary and sex have a general shape of the bell curve, though not perfectly symmetrical. The qqplot have a general shape of a straight line for both groups in the population.

Based on the plots, there is not enough evidence against normality. There is one outlier in the men’s sample and no evidence for serial correlation. Based on F-test and levene’s test, the variances of both groups in the population are equal.

With acceptable sample size, not enough evidence against normality and equal variances, we decided to use the two sample Student’s *t* test on starting salary. Based on the very small p-value, the null hypothesis, that the true mean is equal to 0, was rejected. The confidence interval suggests the difference in means for starting salaries for men and women to range from approximately 550 to 1076.

To determine the sample size needed to detect a difference in salaries between males and females, the power analysis requires specifying the significance level, power, and effect size. The significance level (set to 0.05) represents the probability of rejecting the null hypothesis when it is actually true, also known as a Type I error. The power, or the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true was set to 0.80. This corresponds to an 80% chance of detecting a true difference between groups, if one exists. The effect size represents the magnitude of difference between group. With this power analysis, Cohen’s d was used as the effect size measure. It was calculated using the male and female salary sample means and standard deviation as d = (mean(male salary) – mean(female salary))/pooled standard deviation. This quantifies the difference between male/female salaries in standard deviation units. The estimated Cohen’s d from the sample data was 0.5, indicating medium sized effect based on typical conventions. By specifying 5% significant level, 80% power, and an effect size of 0.5 SDs from the sample means, the power analysis can determine the required sample size to reliably detect salary differences between genders if it exists in the population.

The two-sample Student’s *t* test indicated that the difference in starting salaries for men and women is statistically significant, not due to random chance, but it does not explain the strength of the relationships between variables. Regression analysis on the other hand, explores the relationships between the input variables and the response variable. There are many different model building methods for regression. Simple linear regression uses a single predictor variable, while multiple linear regression is more appropriate when the outcome depends on more than one predictor. The assumptions are linearity, homoscedasticity, independence and normality.

When developing a multiple linear regression model, the aim is to identify the smallest number of predictors necessary to provide good predictions. Too many predictors will train the model to follow the data’s random variations (noise) too closely. Too few predictors will produce a model that may not be accurate enough at predicting future values.

Different approaches for model development include forward selection, step-wise regression, backward elimination, setwise regression and automated selection approaches. In the current project, we explored forward selection and compared the outcomes to those of backward elimination and automated subsets selection.

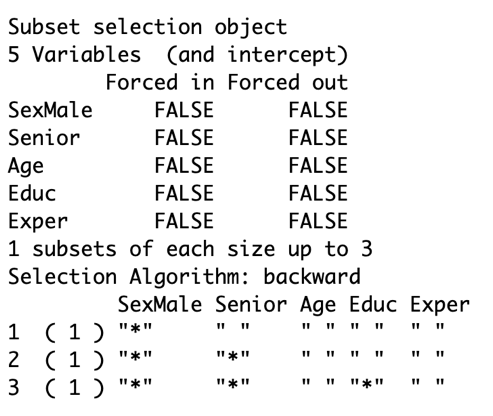
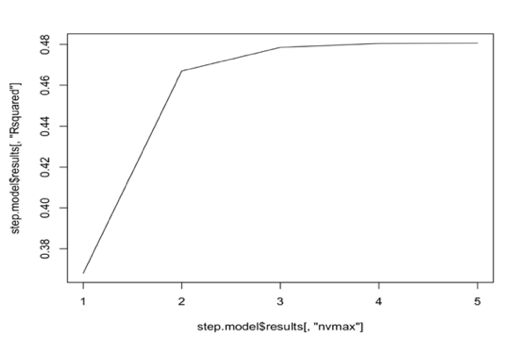
In forward selection, potential predictors are successively tested and added to the model as long as their p-values in the computed model remain below the predefined threshold. A large p-value means that the chance of observing the t-statistic (ratio of estimate to standard error) or larger, assuming the slope or estimate is zero, is fairly likely based on random chance, and the variable is not contributing to the fit of the model. This process continues, one at a time for each predictor, until all predictors have been tested.

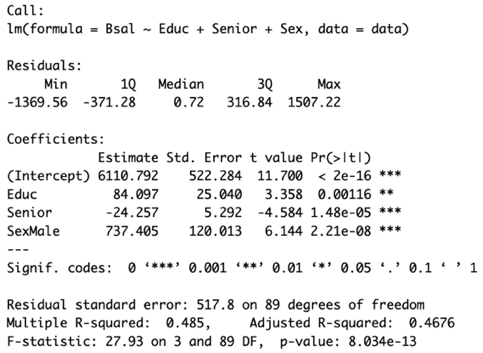
First, all single predictors except Sex (education, seniority, experience, age) were tested, their linear, non-linear and transformed models. The linear model with education explained 16% of the data, while variable Age was omitted from the final model, due to its high p-value. Next, more variables were added to the model. Two-, three and four-variable models and models with their interactions were tested. In case the effect of one variable depends on the level of one or more variables, then it is called an “interaction”. Two-way interaction means variables combine or interact to affect the response.

The non-linear, transformed and interaction models did not add to explaining the data and were omitted from the final model. Similarly, the third and fourth variable contributed very little, and it was decided the model would be better without them. So far, the best model had two-variables (Education and Seniority), explaining 25% of the data.

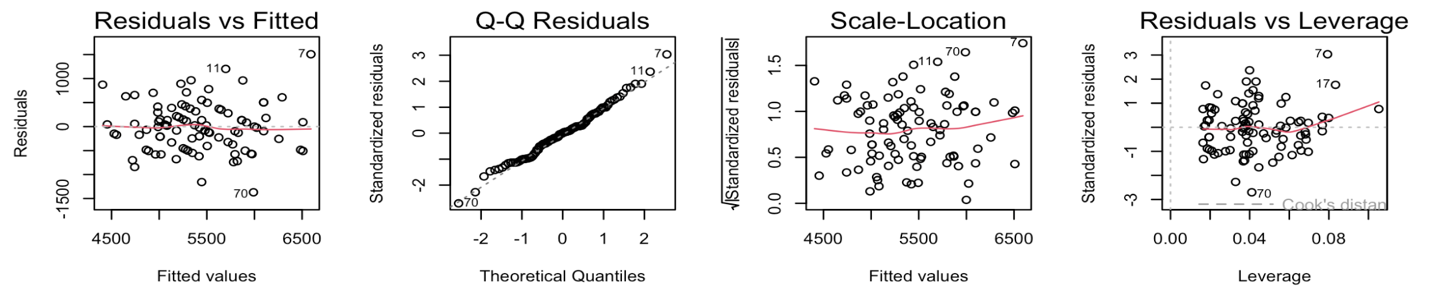
In the next step, variable Sex was introduced to the model. The addition had a significant impact, increasing the Adjusted R-squared by 20% compared to the best previous model. This, in fact, could quantify as statistical evidence of gender-based discrimination at the bank. The Student’s *t* test had established that the differences in starting salaries were not due to random chance and the regression analysis, that these differences depended clearly on gender.

Before picking the final model, two other regression methods – backward elimination and automated subsets – were explored. Backward elimination method starts with a full model. At every step, the variable with the highest p-value is eliminated until all remaining p-values are under the selected threshold. The process led to the three-variable model with Sex, Seniority, Education.

The automated procedures are tempting because it feels that the process will likely test a broader range of possible predictor combinations than could be tested manually. However, the automated procedures often lack intuitive insights into the underlying nature of the system being modeled and should be used with caution. We tested an automated subsets procedure from “caret” package called “leapBackward”, which also suggested the three-variable model with Sex, Seniority and Education as best.

For the final model selection, we also studied the diagnostic information. The plot of r-squared and number of variables made a strong case for a two-variable model with Sex and Seniority, since the third variable adds comparatively little. But, after comparing the Akaike Information Criteria (AIC) values of the models, the three-variable model with Sex, Seniority, Education, with the lowest AIC value among the two runner up models, was chosen.

The coefficients of the final modelsuggest that if a man and a woman with exactly the same education and seniority were to apply for a job at the bank, the man could expect higher starting salary by $737. $737 lies within the 95% confidence interval of the Student’s *t* test.

The residual analysis suggests that assumptions on linearity and normality are reasonable, the variances are equal and that there are no influential observations. 

Among the limitations of the analysis is the observational study design, which means causality cannot be definitively determined. Some variables related to job role may not have been captured in the available data. Also, the current sample size may be underpowered to detect real gender differences in salary. The power analysis indicated that a sample of 128 employees is required to reliably detect a medium effect size difference between male and female salaries with 80% power at 5% significance level. Since the current study includes only 93 observations, it may not have sufficient statistical power to conclusively identify gender discrimination.

In conclusion, the regression analysis found evidence of lower salaries for female employees compared to males with similar education and experience levels. However, due to the limitations of the observational data and potentially insufficient statistical power, more investigation is required to make a determination of unlawful gender discrimination at the bank. The power analysis provides a reference for the sample size needed in future studies of this organization to reliably detect gender pay disparity. Expanding the sample to 128 observations will give 80% power to identify a true gender pay gap if it exists in this population.

**References**

Roberts, H.V. (1979). Harris Trust and Savings Bank: An Analysis of Employee Compensation, *Report 7946*, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.

1. Ramsey, F.L. and Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed)*, Cengage Learning. [↑](#footnote-ref-1)