Predictive Analytics Modeling Exercise

GUILHERME GONÇALVES DE LIMA

Predictive Analytics Modeling Exercise

Nationwide
is on your side

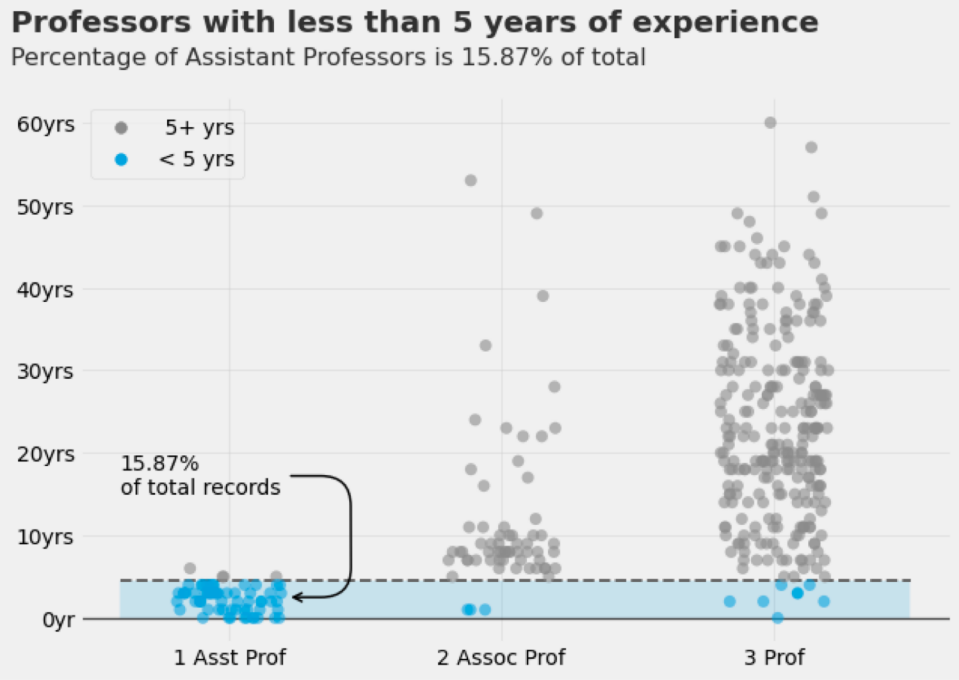
Summary

1. Analysis
2. Model Building
3. Data Set Enhancement

1 ANALYSIS

Analysis

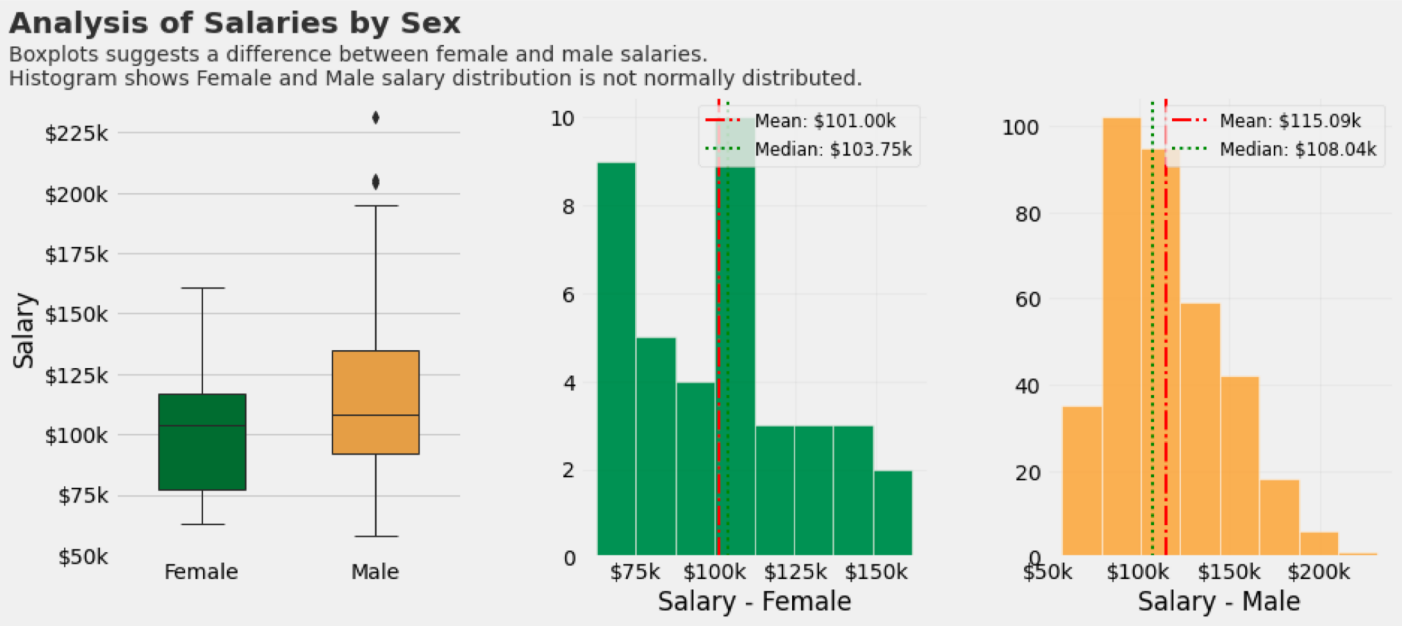
(1) What percentage of records are Assistant Professors with less than 5 years of experience?



Considering Years of Service as Years of experience, **94.03%** of Assistant Professors has less than 5 years of experience, which represents **15.87%** of total records.

Analysis

(2) Is there a statistically significant difference between female and male salaries?

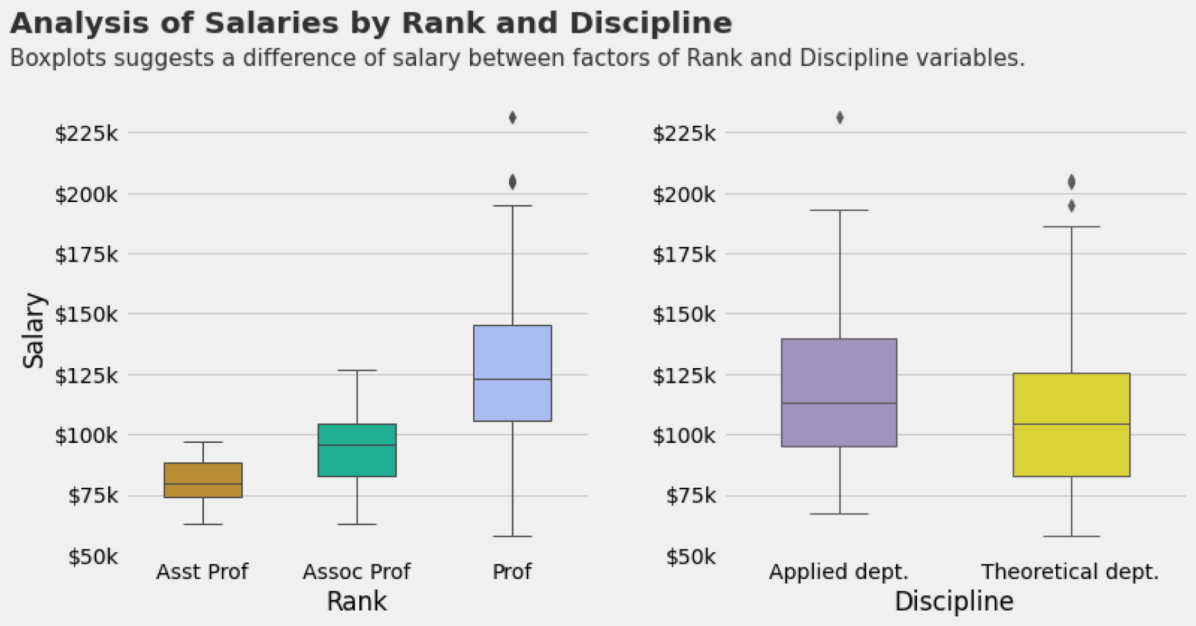


The boxplots shows that male professors has higher salaries (approx. 50% with more than $100k) than Female professors (approx. 50% with less than $100k).

There is significant difference between female and male salaries (Mann-Whitney U test: p-value=0.004).

Analysis

(3) What is the distribution of salary by rank and discipline?

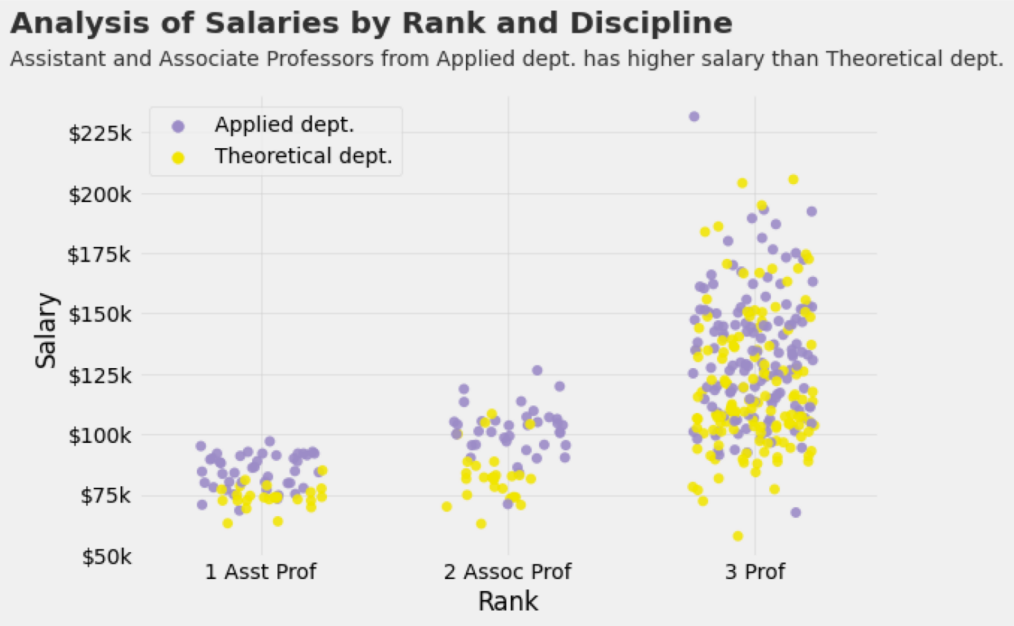


The boxplots shows, in average, Professors has higher salaries (approx. 50% with more than $125k), followed by Associate Professors (approx. 50% with less than $95k) and Assistant Professor (approx. 50% with less than $80k).

For Discipline, It suggests that Applied departments has higher salaries (approx. 75% with more than $100k) than Theoretical departments (approx. 50$ with less than $100k).

Analysis

(3) What is the distribution of salary by rank and discipline?



Applied departments has higher salaries than Theoretical departments for Assistant and Associate Professors.

For Full Professors, It does not suggests a clear difference between salaries of Applied and Theoretical departments.

2 MODEL BUILDING

Model Building – 1st Approach

Using Multiple Linear Regression Model to model Salary level.

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

- $\mathbf{y}$: vector of observed values.
- X: matrix of explanatory variables.
- $\boldsymbol{\beta}$: parameter vector;
- $\boldsymbol{\varepsilon}$: vector of errors (residuals).

Model Building – 1st Approach

Candidates Models

- Salary could not have level zero, i.e., we can model without intercept.
- For modeling Salary as response, we will start with a model with all variables and remove the variables that are not significant. To avoid multicollinearity, if both Years since PhD and Years of Service are significant, we will keep only Years since PhD due high correlation (0.91) between them.
- Other candidates models will be with one explanatory variable to analyze whether only one is sufficient to explain Salary level. See table below.

Model	Explanatory Variables
Model All	Rank, Discipline, Sex, Years since PhD, Years of Service.
Model 1	Rank, Discipline, Years since PhD
Model 2	Rank, Discipline
Model 3	Rank
Model 4	Discipline
Model 5	Sex
Model 6	Years since PhD

Model Building – 1st Approach

First Analysis of Candidates Models:

- Sex is not significant (p-value = 0.3, CI95% [-4375.9, 12579.2]) in Model with all variables.
- Years since PhD is not significant (p-value = 0.3, CI95% [-143.7, 410.2]) in Model 1.
- Other models has all variables are significant (p-value < 0.05).

Comparing results of models (2, 3, 4, 5, and 6) with all variables significance:

- The Models 4 and 5 are significant at 5% but borderline for F Test (hypothesis: all coefficients are zero).
- The Models 2, 3, and 6 are significant for F Test.
- The Models 3, 4 and 5 has Adjusted R-squared lower than .4.
- The Model 2 has lower AIC (7262) and variables explain 44% of the variability in Salary.
- The Model 6 has higher AIC (7804) and variables explain 79% of the variability in Salary.

Based on theses results, we will use Model 2 to predict Salary level.

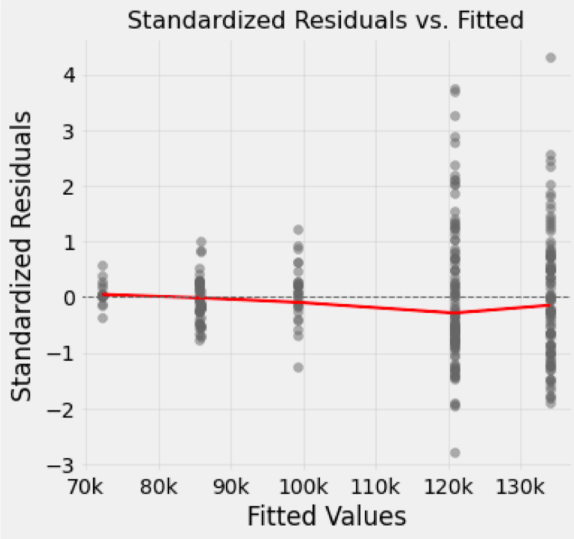
Model Building – 1st Approach

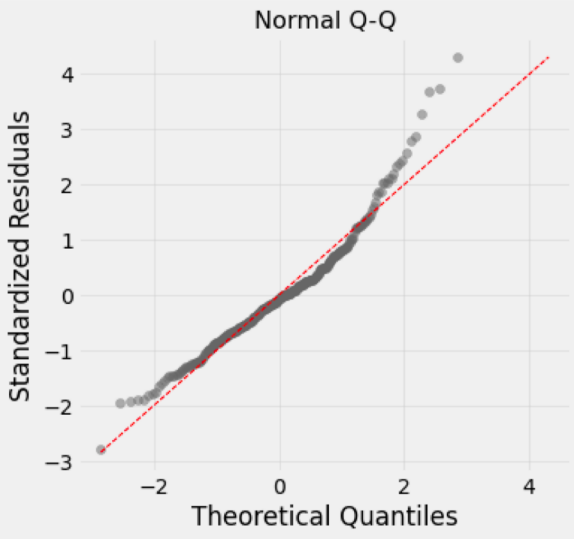
Model 2 results – Ordinary least squares:

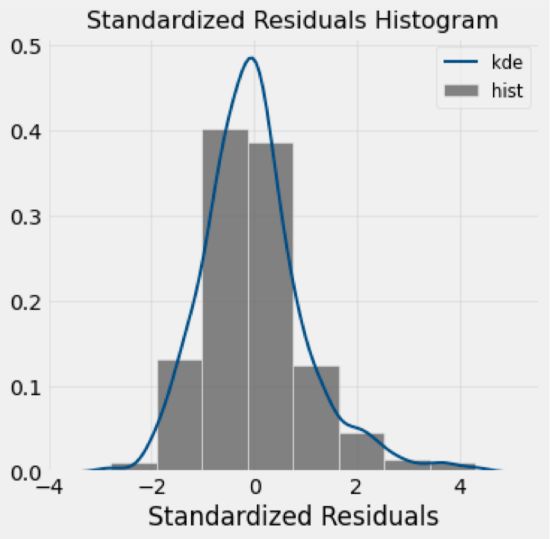
Parameter/Variable	Coef.	Std. Err.	t statistic	p-value	CI 95%
β_1: C(rank)[1 Asst Prof]	85625.7	3213.6	26.6	0.000	[79302.6, 91948.7]
β_2: C(rank)[2 Assoc Prof]	99208.9	3332.1	29.7	0.000	[92652.7, 105765.1]
β_3: C(rank)[3 Prof]	134120.5	2041.6	65.6	0.000	[130103.3, 138137.6]
β_4: C(discipline)[T.Theoretical dept.]	-13287.7	2588.7	-5.1	0.000	[-18381.3, -8194.1]

Model Building – 1st Approach

Model 2 – Residuals Analysis:





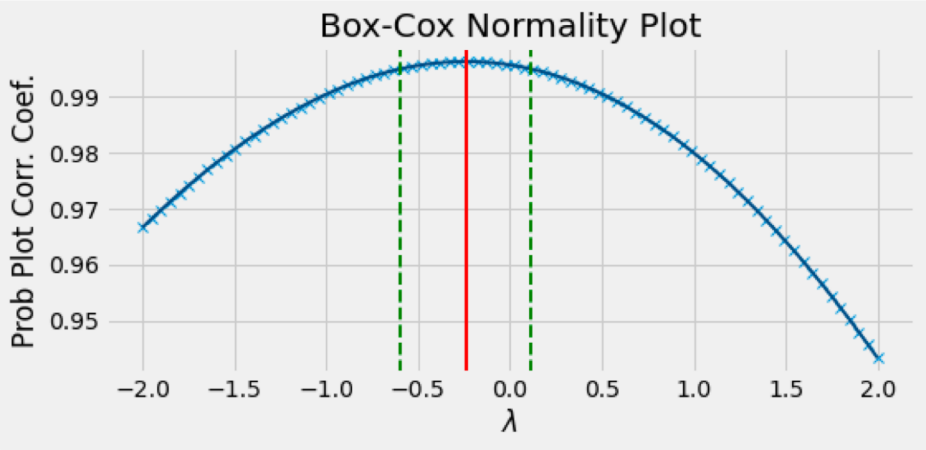


Residuals does not have normal distribution (Jarque-Bera test: p-value < 0.05) and homoscedasticity (Breusch-Pagan test: p-value < 0.05) as well.

Model Building – 1st Approach

Model 2 variable transformation:

To fix heteroskedasticity, we will use Box-Cox Transformation:

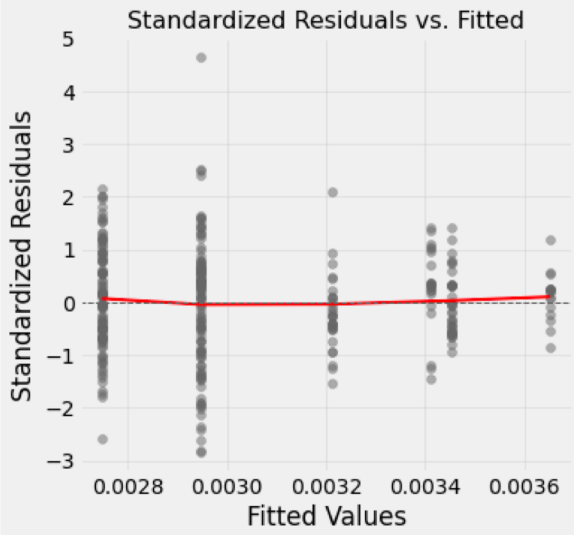


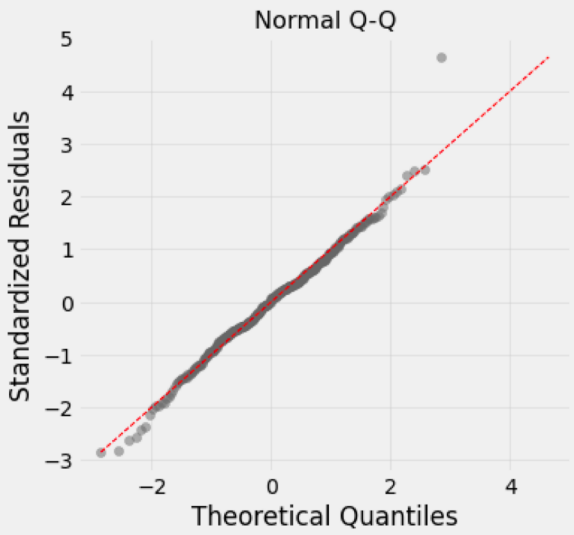
The Box-Cox transformation suggests two:

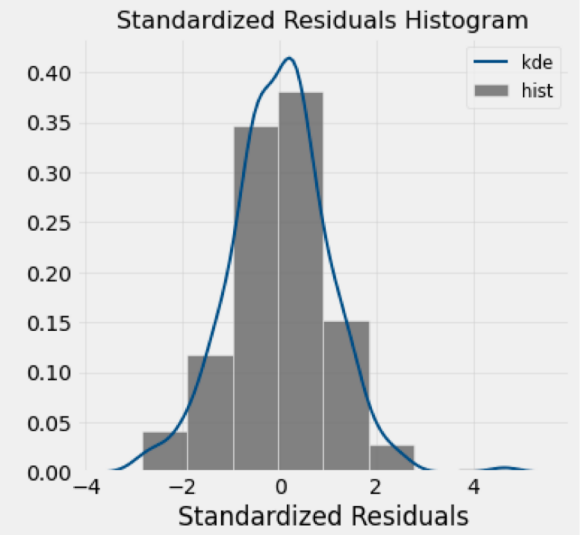
- Inverse Square Root
- Logarithm.

Model Building – 1st Approach

Model 2 Residuals – Transformation: Inverse Square Root of Salary:



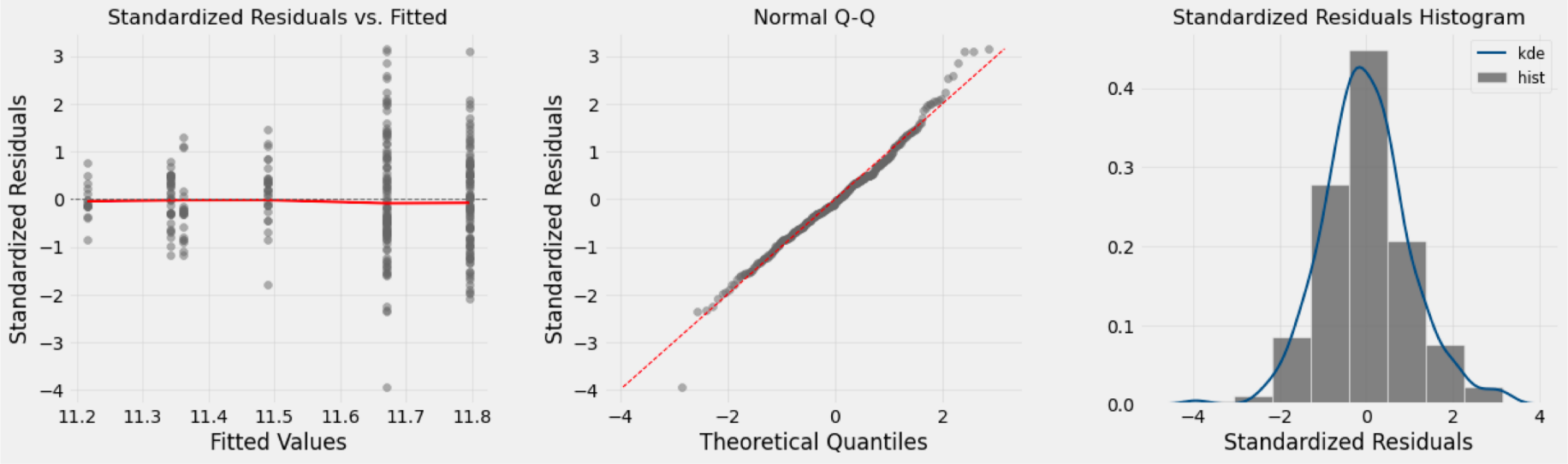




Residuals does not have normal distribution (Jarque-Bera test: p-value < 0.05) and homoscedasticity (Breusch-Pagan test: p-value < 0.05) as well.

Model Building – 1st Approach

Model 2 Residuals – Transformation: Log of Salary:



Residuals does not have normal distribution (Jarque-Bera test: p-value < 0.05) and homoscedasticity (Breusch-Pagan test: p-value < 0.05) as well.

Model Building – 1st Approach

Model 2

- Both transformations (inverse square root and logarithm) does not solve residuals problems.
- Analysis of Salary by Rand and Discipline shows that Applied departments has higher salaries than Theoretical departments for Assistant and Associate Professors.
- So far, Model 2 is the best model (lower AIC=7262.49).
- Keep in mind this results the model could be inefficient and unstable, but we will use the Model 2 to predict in our test data.
- For future models to predict Salary level, we can try non-linear models or transformation in predictors variables.

Model 2 – Test data results

Using K-fold cross-validation in train dataset.

RMSE
Train: 22,609.45 (+/- 5,135.72)
Test: 22,540.10

RMSE statistic is lower in test set than train set.

Model Building – 2nd Approach

Using Logistic Regression to model the probability of Salary being higher than median.

$$Logit[P(Y=1)] = \ln \frac{P(Y=1)}{1-P(Y=1)} = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where:

$P(Y=1)$: is the probability that salary is higher than median.

$X_1, X_2, \ldots, X_n$: are the explanatory variables.

$\beta_1, \beta_2, \ldots, \beta_n$: are the coefficients (effects) of variables.

Model Building – 2nd Approach

Logistic Regression results:

Parameter/Variable	Coef.	Std. Err.	z statistic	p-value	CI 95%
β_1: C(rank)[1 Asst Prof]	24.27	17514.44	0.0014	0.9989	[-34303.39, 34351.94]
β_2: C(rank)[2 Assoc Prof]	1.47	0.41	3.5107	0.0004	[0.65, 2.29]
β_3: C(rank)[3 Prof]	-1.73	0.26	-6.5448	0.0000	[-2.2538, -1.21]
β_4: C(discipline)[T.Theoretical dept.]	1.21	0.32	3.8013	0.0000	[0.59, 1.84]

Null deviance:	439.20 on 317 degrees of freedom
Residual deviance:	270.02 on 313 degrees of freedom
AIC:	278.02
Pseudo-R2:	0.39

- Residual Deviance is lower then Null Deviance indicates the model is appropriate.
- AIC is lower than Linear Regression models.
- Based on this results, we will evaluate results of this model comparing train and test accuracy.

Model Building – 2nd Approach

Logistic Regression Model

Train			
	Precision	Recall	F1-Score
$Y = 0$	0.74	0.96	0.83
$Y = 1$	0.93	0.64	0.76
Macro avg	0.83	0.80	0.79
Accuracy			0.80

Test			
	Precision	Recall	F1-Score
$Y = 0$	0.33	0.41	0.37
$Y = 1$	0.00	0.00	0.00
Macro avg	0.17	0.20	0.18
Accuracy			0.20

- The difference between train and test accuracy indicates the model is not generalizing well.
- Precision, Recall and F1-Score are zero indicating this model is not the best.

3 DATA SET ENHANCEMENT

Data Set Enhancement

(1) State at least three research questions you would like to address and describe your thought process behind how you formulated these research questions.

1. Is research productivity impact on salaries?

In Mittal et. all (2008), they suggests that being full professor and from higher ranked research university are associated with higher salary.

2. Is race/ethnicity impact on salaries?

In Guillory (2001), the author suggests that salaries does have significant impact by Race/ethnicity.

3. Will professors leave college?

In Gofman and Jin (2019), they investigates the impact of AI human capital from research universities to the private sector could affected negatively in domain-specific transfer of knowledge to students.

References:

- Gofman, M., & Jin, Z. (2019). Artificial Intelligence, Human Capital, and Innovation. Human Capital, and Innovation (August 20, 2019).
- Guillory, E. A. (2001). The Black professoriate: Explaining the salary gap for African-American female professors. Race Ethnicity and Education, 4(3), 225-244.
- Mittal, V., Feick, L., & Murshed, F. (2008). Publish and prosper: The financial impact of publishing by marketing faculty. Marketing Science, 27(3), 430-442.

Data Set Enhancement

(2) Prepare a list of 5-7 additional attributes you would like to add to the data set. Prepare a brief explanation for each attribute.

Variable	Description
Research Productivity	Number of publications in different types of journals (Tier 1, Tier 2 and Tier 3).
Race:	Following U.S. Census Bureau with categories: (a) White, (b) Black or African American, (c) American Indian or Alaska Native, (d) Asian, (e) Native Hawaiian or Other Pacific Islander.
Type of University	Private, four-year colleges and two-year colleges.
Research Activity	Number of hours spent on research.
Teaching Activity	Number of hours spent on teaching.
Field:	Sciences, social sciences, technical, semi-technical, humanities, other.
Turnover	Professor leave college to private sector.

Data Set Enhancement

(3) Estimate and justify the appropriate sample size (and sampling technique, if desired) that would be required to address the research questions you defined.

In 2017, National Center for Education Statistics counted 182,388 Professors, 158,082 Associate professors, 173,409 Assistant professors (Total: 513,879).

For first sampling, Yamane (1967) suggests, for a 95% confidence level and p = 0.5 (maximum variability), sample size:

$$n = \frac{N}{1 + N(e^2)}$$

where N is the population size, and e is the desired level of precision.

$$n = \frac{500000}{1 + 50000(0.05^2)} = 399.68$$

So, we will get sample size as 400.

References:

- Snyder, T.D., de Brey, C., and Dillow, S.A. (2019). Digest of Education Statistics 2018 (NCES 2020-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Yamane, Taro. 1967. Statistics, An Introductory Analysis, 2nd Ed., New York: Harper and Row.

Thanks

GUILHERME G LIMA