Modern Salary Modeling Project

Justin Mai, Troy Russo, Isaac Muhlestein, Conan Li, Jian Kang

Contents

1. Introduction

Description: The job market can be a hard place to navigate, especially with the search of data roles in the recent years. As statistics students, many of us are leaning towards opportunities within data roles. To understand the recent market we will be analyzing data job logistics to investigate the factors and predictors that most impacts the salary of these roles. Within this report, we will be using Salary Index data reported by real people in the industry to (1) discover the factors and variables within the job description that may influence a person’s job salary the most to help students like us navigate the market and (2) …

Disclaimer: We have pivoted from using this dataset https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights which is comprised of synthetic data based off the current job market regarding AI jobs to this dataset https://www.kaggle.com/datasets/murilozangari/jobs-and-salaries-in-data-field-2024/data which consists of real survey data from various people in data roles, reporting through this website https://aijobs.net/salaries/2024/. We decided to make this change because we believe that variables such as `experience_level` and `job_category` which can be found in our current dataset would be strong predictors for `salary`. We also believe that using real survey data as supposed to synthetic data would give us results that are more related to real-life circumstances, making the report more applicable for all.

---

2. Methods

2.1 Data Description

The first dataset was collected through https://aijobs.net/salaries/2024/, it consists of 14199 different observations, with each observation representing a person in their role in 2024. The **response variable** we are measuring is `salary_in_usd` which measures a person's annual gross salary. The **8 predictors** are `experience_level`, `employment_type`, `job_title`, `employee_residence`, `work_setting`, `company_location`, `company_size`, `job_category`. All of these variables are categorical where `company_size` is categorized as S for small, M for medium, and L for large.

The second dataset consists of cost of living index by country where an index of 100 represents the living cost of NYC, United States, so all the indices are relative to that. We will merge the two datasets by `country`. The predictors we're looking at in this dataset are Cost of Living Index, Rent Index, Cost of Living Plus Rent Index, and Local Purchasing Power Index. We believe that the cost of living could be indicative of `salary_usd`.

2.2 Data Processing

The primary dataset will be comprised of the two datasets described in (2.1). We are joining the two datasets on `employee_residence` which is in form of country. Now each row will consist of a specified job description along with the cost indexes for each respective resident. Having all of these predictors in one dataset will allow us to utilize the lm() function to uncover linear trends for all predictor variables in response to `salary`. It will also allow us to compare models easily which we will do using ANOVA tests and by calculating the F-statistic. The primary dataset consists of 14199 observations after joining

Data Manipulation: Rows that consisted of NAs were in countries that weren't listed in the `cost_of_living` data, this demonstrates that their rank is low when ordering by index and there weren't a sufficient number of samples for those countries. Therefore we removed those observations (14161 observations). We also removed exact duplicate rows from the dataset (7575 observations)

Mutations in the data were also made to create new predictors `us_resident` which is a binary variable that denotes if the job is in the U.S. or not, and `experience_numeric` which turns `experience_level` into numerical values (i.e. 1 - "Entry-Level", 2 - Mid-level", 3 - "Senior", 4 - "Executive"), this transformation will support our use of linear modeling and allow us to easily check assumptions such as linearity assumptions. Also, because we have too many different job titles, we decided to aggregate these job titles by keywords into 8 categories (Data Scientist, Data Analyst, Machine Learning, Data Engineer, Leadership, Business Intelligence, Research, Other). This will consolidate our data and make linear models more interpretable

We are also reordering values to ensure that our **baseline term** is what we want it to be (i.e. releveling small companies to be the first type and entry-level jobs to be the first job types).

2.3 Model Diagnostics

Linear Modeling Assumptions:

- **Linearity**: The relationship between the predictor and response is linear.

- **Independence**: All observations are independent of one another (pair-wise independence).
- **Homoscedasticity**: The variance of residuals is constant across predictor levels.
- **Residual Normality**: The residuals follow a normal distribution.

We know that the reported job descriptions are all independent of one another through the data description.

The model that we are using seems highly categorical, even after our data transformation process, which would result in very discrete predictions. To fix this issue we want to create a new interaction term and turn it into a predictor variable, adding a continuous predictor for `salary_usd`. We hypothesize that salary growth will vary by location, therefore, we are creating an interaction term between `experience_numeric` and the Cost of Living Index to test this (experience combined with living costs could impact salary growth differently).

The residuals when using a non-transformed model is skewed due to the deviations within the tails in our qq-plot, to fix this issue, we would have to use a **log-transformation** on the data. The variance of the residuals is also constant and there is a clear linear and positive relationship between the predictor and response.

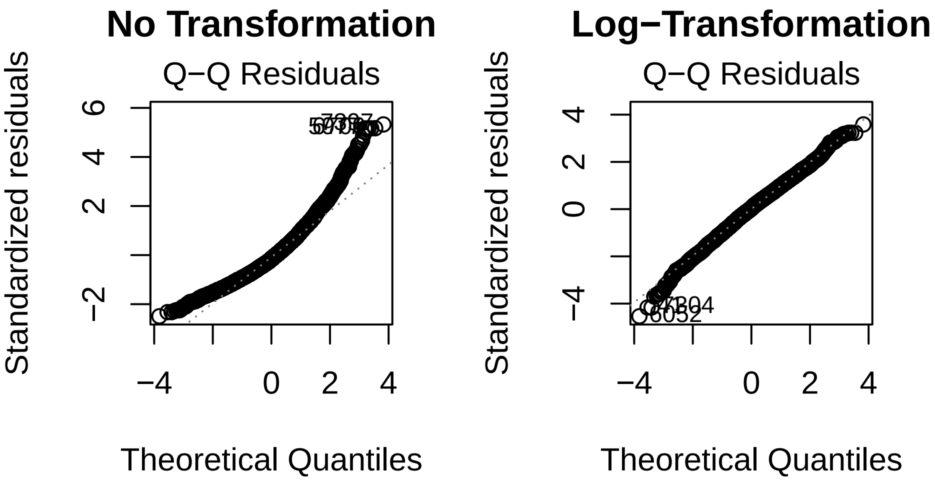


Figure 1: Normality check for Experience and Cost of Living Index Interaction variable on Salary, comparison between no transformation and log-transformation

To test the assumptions for our categorical variables we must look at the average salary by category, to see if there is a clear trend between the different experience levels in respective company size. This checks off the linearity assumption because there is a (somewhat) linear increasing trend for all the experience levels. We can see that we don't need an interaction effect between `experience_level` and `company_size` because the trend is increasing for each company size at around the same rate. We can test this using an ANOVA test at $\alpha = 0.05$ to see if adding an interaction term effects the model.

```
## Analysis of Variance Table
##
## Model 1: salary_in_usd ~ experience_level + company_size
## Model 2: salary_in_usd ~ experience_level * company_size
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1   7569 2.9623e+13
## 2   7563 2.9595e+13  6 2.8265e+10 1.2039 0.3008
```

As we can see from the ANOVA test, our the p-value of 0.3008 is much higher than our $\alpha = 0.05$, therefore, the addition of an interaction isn't impactful to our linear model.

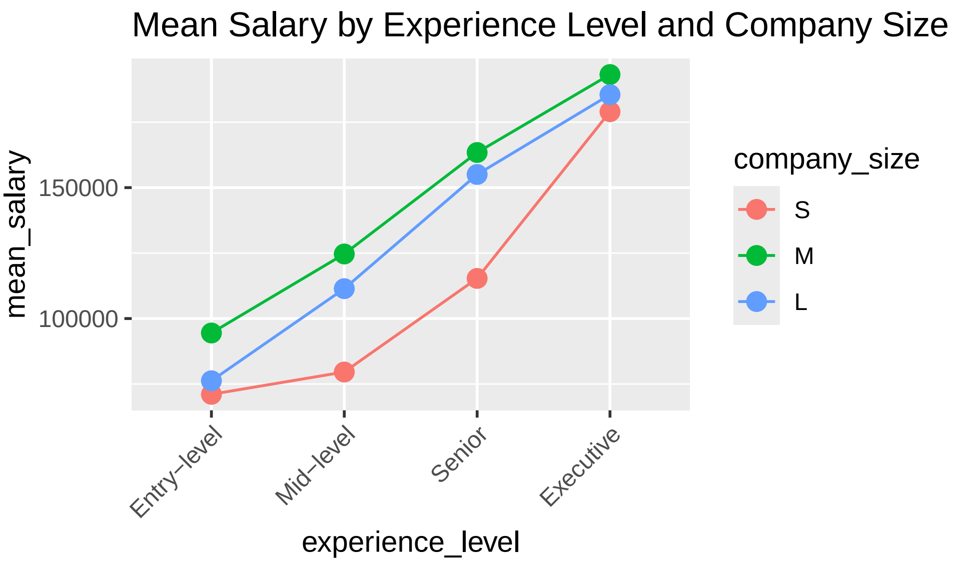
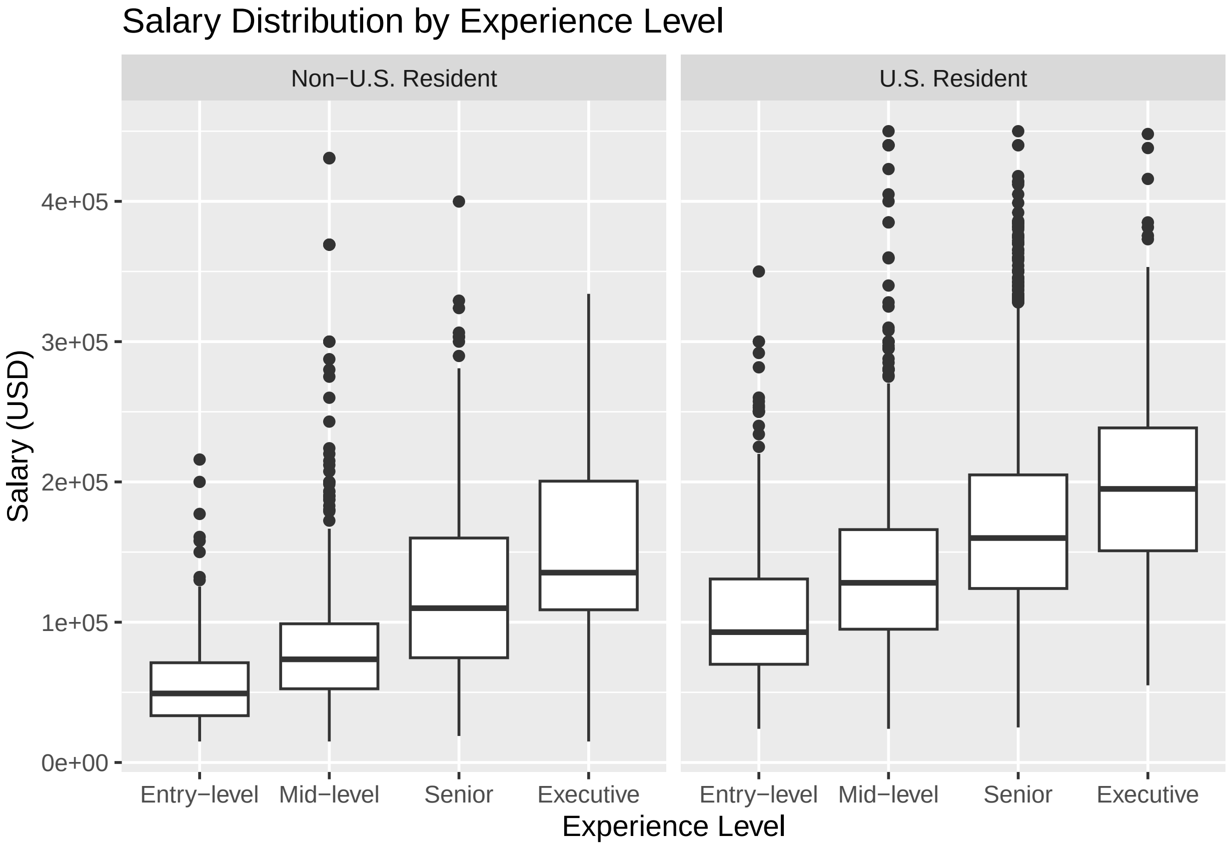


Figure 2: Interaction between Experience Level and Company Size on Salary

3. Results

3.1 Data Exploration



3.2 Model Selection

To determine our model, we will fit multiple linear regression models and find the model that minimizes our **AIC (Akaike Information Criterion) and/or BIC (Bayesian Information Criterion)**. If the AIC and BIC suggest different models, we will favor the model selected by lowest AIC because BIC penalizes models with a large number of observations and tends to predict less than the AIC.

```
full_fit <- lm(salary_in_usd ~ ., data=model_df)
summary(full_fit)$r.sq
```

```
## [1] 0.3319171
```

```
fit1 <- lm(salary_in_usd ~ experience_level + company_size, data=model_df)
summary(fit1)$r.sq
```

```
## [1] 0.1597929
```

```
fit2 <- lm(salary_in_usd ~ experience_level + company_size, data=model_df)
summary(fit2)$r.sq
```

```
## [1] 0.1597929
```

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: salary_in_usd ~ experience_level + company_size
## Model 2: salary_in_usd ~ experience_level + company_size
##   Res.Df        RSS Df Sum of Sq F Pr(>F)
## 1   7569 2.9623e+13
## 2   7569 2.9623e+13  0         0
```

3.3 Modeling

3.4 Model Descriptions

The initial model was a multiple linear regression model with categorical and numerical predictors to predict 'salary_in_usd'. We found Heteroscedasticity and non-normally distributed residuals, hence through application of log transformation to minimize AIC we can improve the model significantly.

The final model uses 'experience_level', 'job_title','employee_residence','work_setting','company_size' and an interaction term 'experience_level:company_size' as predictors to predict 'log_salary'. We can get the actual salary by taking the exponent of the log. This model assumes linear relationship, and there seems to have residual bias which suggests some possible non-linear relationships.

4. Conclusion