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) if there's a difference between the criterion (AIC and BIC), we will split our data into training and testing data to compare the two models optimized by the criterion, if not, we will still split the data to test the our optimized model.

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.

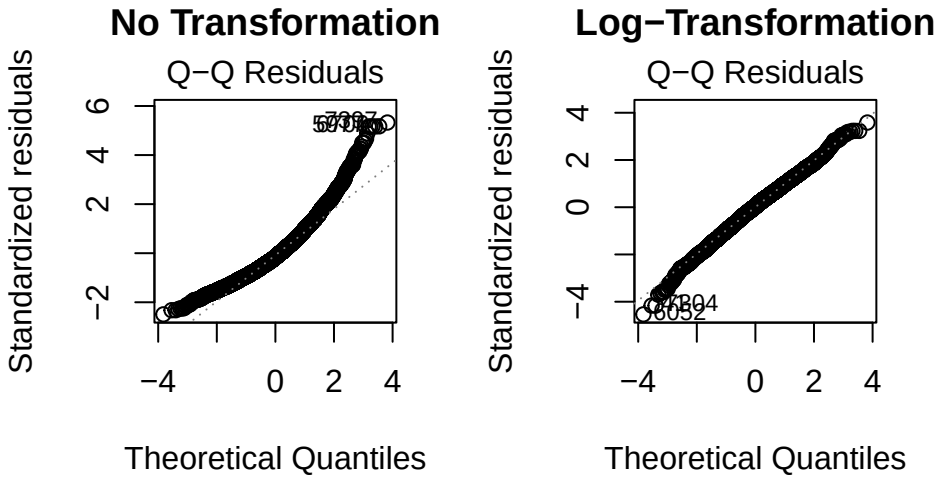


Figure 1: Normality check for Experience and Cost of Living Index Interaction variable on Salary, comparison between no transformation and log-transformation

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.

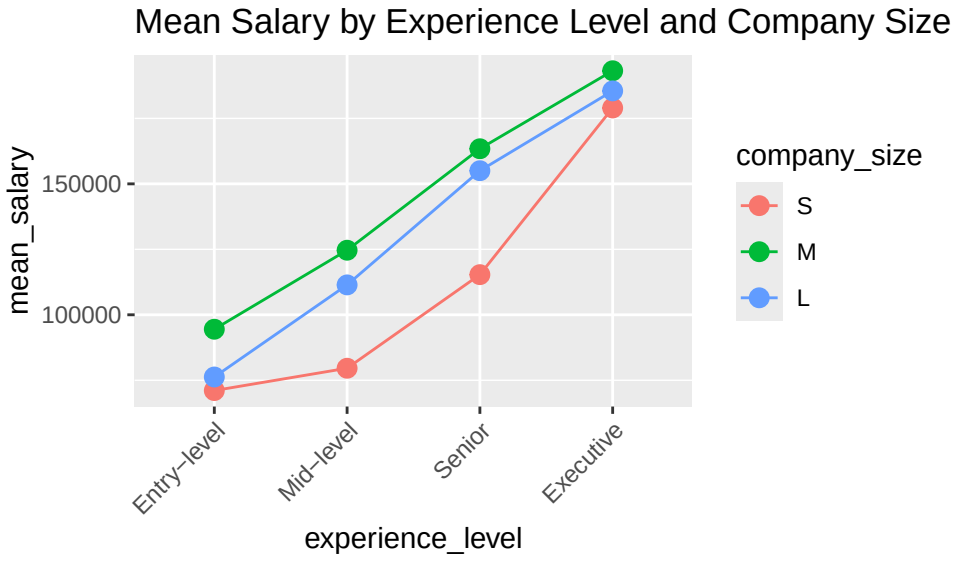
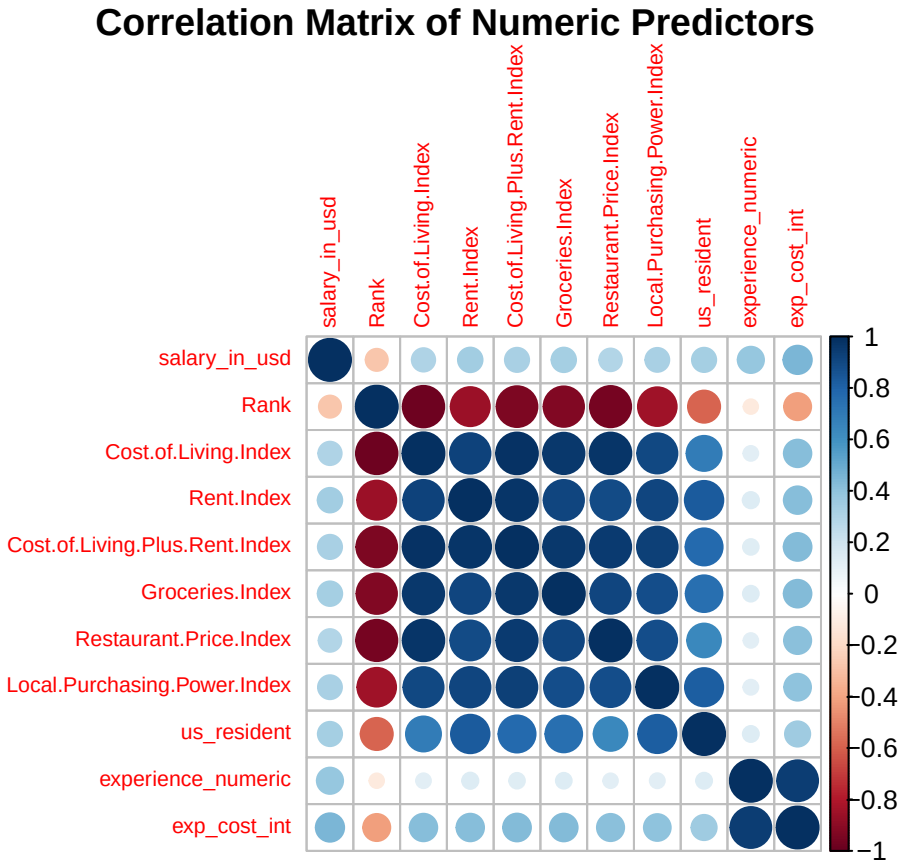


Figure 2: Interaction between Experience Level and Company Size on Salary

Correlation and Multicollinearity Analysis

One of the assumptions we are making in our model is that the predictors we are using are independent. Thus, it becomes expedient to test the multicollinearity of our predictors, to ensure that they are each independent of one another and add new information. Because a large majority of our predictors are categorical, The VIF (Variance Inflation Factor) will be most effective in calculating that multicollinearity:

```
##                       GVIF Df GVIF^(1/(2*Df))
## experience_level 1.214952  3        1.032983
## company_location 6.527462 62        1.015244
## company_size     1.322118  2        1.072303
## job_category     1.171177  7        1.011350
## Rank             4.778094  1        2.185885
```



Examining our correlation matrix, we can see that the index variables are highly correlated with one another since the cost of living indices are joined by country, so there are several repeated values. To satisfy the multicollinearity assumption, we will be removing many of these highly correlated variables (only using `Cost.of.Living.Index`) for our MLR model.

From our data processing and model diagnostics step, to avoid repeating variables, satisfy the assumptions necessary for conducting a linear model, and to consolidate the number of predictors we are using, we'll be using a dataframe with the following variables: experience_level, salary_in_usd, employee_residence, work_setting, company_location, company_size, job_category, Rank, Cost.of.Living.Index, us_resident, experience_numeric, exp_cost_int.

```
set.seed(123)  # Ensure reproducibility

# Calculate the total number of rows in model_df
n <- nrow(model_df)

# Randomly sample indices for the 20% test set
```

```r
test_indices <- sample(seq_len(n), size = floor(0.2 * n))

# Create the test set and candidate (training) set
test_set <- model_df[test_indices, ]
candidate_set <- model_df[-test_indices, ]
```

2.4 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.

We can use a step function from the MASS package that computes the best model purely based on the AIC by comparing every potential model combination and returning the model with the lowest AIC. This model may be overfit however, since AIC does not account for model complexity.

```
##
## Call:
## lm(formula = salary_in_usd ~ experience_level + employee_residence +
##     work_setting + company_size + job_category, data = model_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -142595  -38564   -7188   29821  360710
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           69798.2    56986.9   1.225 0.220685
## experience_levelMid-level             21429.7     2501.9   8.565  < 2e-16
## experience_levelSenior                51315.2     2351.9  21.819  < 2e-16
## experience_levelExecutive             84397.5     3895.6  21.665  < 2e-16
## employee_residenceArgentina          -74787.8    58862.5  -1.271 0.203928
## employee_residenceArmenia            -90891.3    69065.1  -1.316 0.188207
## employee_residenceAustralia           -2893.2    56956.4  -0.051 0.959488
## employee_residenceAustria            -62583.4    59016.5  -1.060 0.288979
## employee_residenceBelgium            -39863.3    60771.6  -0.656 0.511875
## employee_residenceBrazil             -79194.1    57666.5  -1.373 0.169696
## employee_residenceBulgaria           -36775.5    79751.7  -0.461 0.644722
## employee_residenceCanada              -9690.0    56422.8  -0.172 0.863647
## employee_residenceChile               37046.8    79782.2   0.464 0.642411
## employee_residenceChina               71948.3    79547.4   0.904 0.365775
## employee_residenceColombia           -73714.6    58298.8  -1.264 0.206116
## employee_residenceCosta Rica         -50249.3    79380.1  -0.633 0.526739
## employee_residenceCroatia            -66606.2    61744.0  -1.079 0.280735
## employee_residenceCyprus             -88175.6    79447.4  -1.110 0.267094
## employee_residenceDenmark            -60559.0    64920.3  -0.933 0.350943
## employee_residenceDominican Republic  -2004.5    79499.1  -0.025 0.979885
## employee_residenceEcuador            -96565.9    79598.1  -1.213 0.225104
## employee_residenceEgypt              -26474.5    58636.1  -0.452 0.651638
## employee_residenceEstonia            -82755.7    59392.8  -1.393 0.163551
## employee_residenceFinland            -92816.3    64930.6  -1.429 0.152911
```

```
## employee_residenceFrance                -52376.4    56684.9  -0.924 0.355520
## employee_residenceGeorgia               -77130.4    79658.5  -0.968 0.332944
## employee_residenceGermany               -39565.6    56604.9  -0.699 0.484588
## employee_residenceGhana                 -82192.0    63046.9  -1.304 0.192388
## employee_residenceGreece                -72201.1    58471.1  -1.235 0.216937
## employee_residenceHungary               -44502.4    79628.9  -0.559 0.576266
## employee_residenceIndia                 -53798.5    57588.5  -0.934 0.350236
## employee_residenceIndonesia             -72502.4    79628.9  -0.911 0.362586
## employee_residenceIraq                   -6416.2    79545.5  -0.081 0.935714
## employee_residenceIreland               -48922.3    58621.8  -0.835 0.404003
## employee_residenceItaly                 -85285.0    57630.4  -1.480 0.138952
## employee_residenceJapan                  -6012.1    63026.6  -0.095 0.924008
## employee_residenceKenya                 -26050.5    69124.5  -0.377 0.706286
## employee_residenceKuwait                -20810.6    79434.3  -0.262 0.793340
## employee_residenceLatvia                -94193.8    58315.7  -1.615 0.106301
## employee_residenceLebanon               -15191.0    68914.7  -0.220 0.825541
## employee_residenceLithuania             -72973.6    58058.0  -1.257 0.208825
## employee_residenceLuxembourg            -23713.2    79621.2  -0.298 0.765845
## employee_residenceMalaysia               57680.8    79627.6   0.724 0.468854
## employee_residenceMalta                 -45713.2    64920.4  -0.704 0.481366
## employee_residenceMauritius               7566.4    79718.9   0.095 0.924387
## employee_residenceMexico                -48844.0    58598.2  -0.834 0.404566
## employee_residenceNetherlands           -56283.1    57298.1  -0.982 0.325991
## employee_residenceNew Zealand             4508.1    60776.4   0.074 0.940873
## employee_residenceNigeria               -68540.4    58632.1  -1.169 0.242445
## employee_residenceOman                 -144368.7    79550.0  -1.815 0.069592
## employee_residencePakistan              -75405.8    60831.2  -1.240 0.215166
## employee_residencePeru                  -82176.2    79742.3  -1.031 0.302799
## employee_residencePhilippines           -69934.3    59745.0  -1.171 0.241818
## employee_residencePoland                -59479.2    58248.1  -1.021 0.307223
## employee_residencePortugal              -87961.6    57232.8  -1.537 0.124358
## employee_residencePuerto Rico             3634.7    61653.4   0.059 0.952991
## employee_residenceQatar                 143356.0    79555.0   1.802 0.071590
## employee_residenceRomania               -92471.8    61632.1  -1.500 0.133557
## employee_residenceSaudi Arabia            646.3    64962.2   0.010 0.992062
## employee_residenceSerbia                -91108.7    79722.8  -1.143 0.253150
## employee_residenceSingapore             -28760.9    61583.6  -0.467 0.640497
## employee_residenceSlovenia              -69940.7    60823.1  -1.150 0.250220
## employee_residenceSouth Africa          -79612.5    58641.4  -1.358 0.174627
## employee_residenceSpain                 -78194.0    56710.9  -1.379 0.167992
## employee_residenceSweden                  -478.1    68839.9  -0.007 0.994459
## employee_residenceSwitzerland           -13704.3    60760.2  -0.226 0.821560
## employee_residenceThailand             -123933.2    69040.5  -1.795 0.072681
## employee_residenceTunisia               -55694.6    68925.7  -0.808 0.419094
## employee_residenceUganda                -99214.3    79526.3  -1.248 0.212229
## employee_residenceUkraine               -64253.7    58807.2  -1.093 0.274597
## employee_residenceUnited Arab Emirates  -69983.2    63092.4  -1.109 0.267372
## employee_residenceUnited Kingdom        -38113.9    56395.2  -0.676 0.499166
## employee_residenceUnited States           7052.6    56355.2   0.125 0.900412
## employee_residenceUzbekistan            -50156.8    65048.8  -0.771 0.440694
## work_settingIn-person                    16691.7     4771.5   3.498 0.000471
## work_settingRemote                       12004.5     4746.9   2.529 0.011462
## company_sizeM                            16793.5     5117.2   3.282 0.001036
```

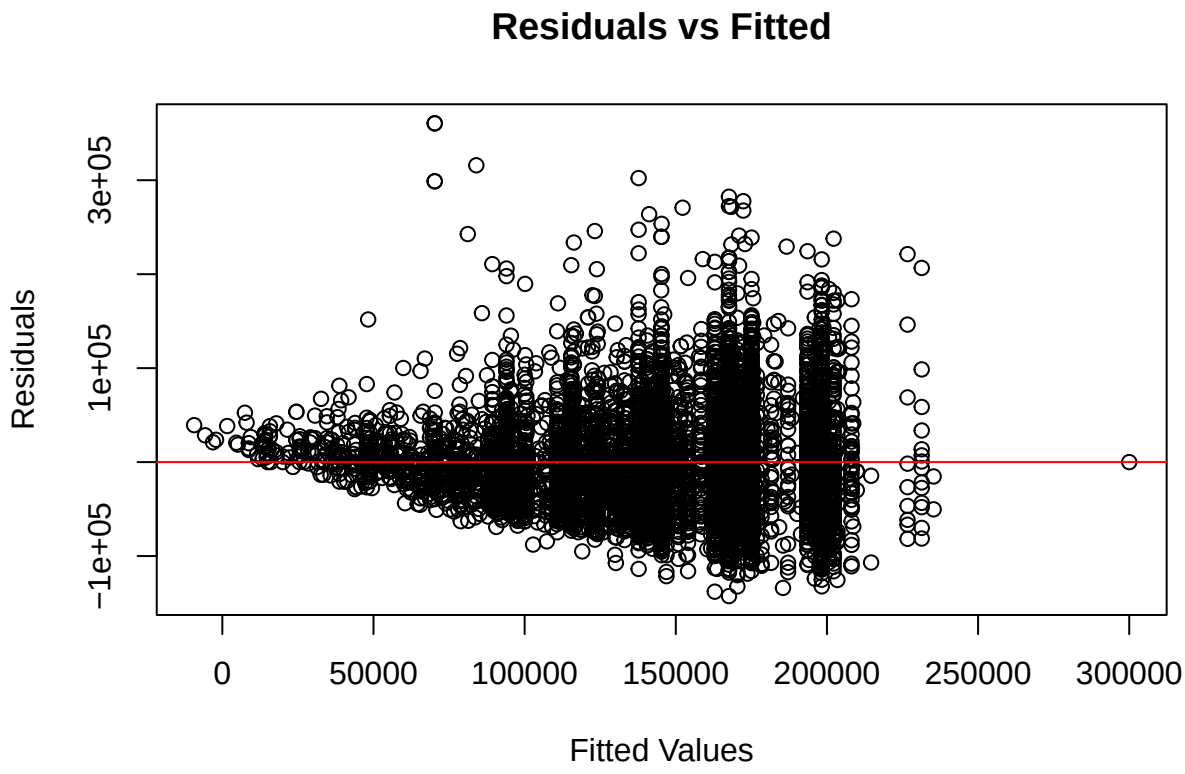
```
## company_sizeL                            17354.9     5483.7   3.165 0.001558
## job_categoryData Analyst                -16342.5     8498.4  -1.923 0.054515
## job_categoryData Engineer                 5943.7     8475.5   0.701 0.483150
## job_categoryData Scientist               13408.2     8484.0   1.580 0.114054
## job_categoryLeadership                   -7592.4     9261.3  -0.820 0.412357
## job_categoryMachine Learning             36617.9     8548.0   4.284 1.86e-05
## job_categoryOther                       -31204.5     9855.8  -3.166 0.001551
## job_categoryResearch                     40554.6     9128.8   4.442 9.02e-06
##
## (Intercept)
## experience_levelMid-level              ***
## experience_levelSenior                 ***
## experience_levelExecutive              ***
## employee_residenceArgentina
## employee_residenceArmenia
## employee_residenceAustralia
## employee_residenceAustria
## employee_residenceBelgium
## employee_residenceBrazil
## employee_residenceBulgaria
## employee_residenceCanada
## employee_residenceChile
## employee_residenceChina
## employee_residenceColombia
## employee_residenceCosta Rica
## employee_residenceCroatia
## employee_residenceCyprus
## employee_residenceDenmark
## employee_residenceDominican Republic
## employee_residenceEcuador
## employee_residenceEgypt
## employee_residenceEstonia
## employee_residenceFinland
## employee_residenceFrance
## employee_residenceGeorgia
## employee_residenceGermany
## employee_residenceGhana
## employee_residenceGreece
## employee_residenceHungary
## employee_residenceIndia
## employee_residenceIndonesia
## employee_residenceIraq
## employee_residenceIreland
## employee_residenceItaly
## employee_residenceJapan
## employee_residenceKenya
## employee_residenceKuwait
## employee_residenceLatvia
## employee_residenceLebanon
## employee_residenceLithuania
## employee_residenceLuxembourg
## employee_residenceMalaysia
## employee_residenceMalta
```

```
## employee_residenceMauritius
## employee_residenceMexico
## employee_residenceNetherlands
## employee_residenceNew Zealand
## employee_residenceNigeria
## employee_residenceOman                 .
## employee_residencePakistan
## employee_residencePeru
## employee_residencePhilippines
## employee_residencePoland
## employee_residencePortugal
## employee_residencePuerto Rico
## employee_residenceQatar                .
## employee_residenceRomania
## employee_residenceSaudi Arabia
## employee_residenceSerbia
## employee_residenceSingapore
## employee_residenceSlovenia
## employee_residenceSouth Africa
## employee_residenceSpain
## employee_residenceSweden
## employee_residenceSwitzerland
## employee_residenceThailand             .
## employee_residenceTunisia
## employee_residenceUganda
## employee_residenceUkraine
## employee_residenceUnited Arab Emirates
## employee_residenceUnited Kingdom
## employee_residenceUnited States
## employee_residenceUzbekistan
## work_settingIn-person                  ***
## work_settingRemote                     *
## company_sizeM                          **
## company_sizeL                          **
## job_categoryData Analyst               .
## job_categoryData Engineer
## job_categoryData Scientist
## job_categoryLeadership
## job_categoryMachine Learning           ***
## job_categoryOther                      **
## job_categoryResearch                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56110 on 7490 degrees of freedom
## Multiple R-squared:  0.3311, Adjusted R-squared:  0.3236
## F-statistic: 44.13 on 84 and 7490 DF,  p-value: < 2.2e-16
```

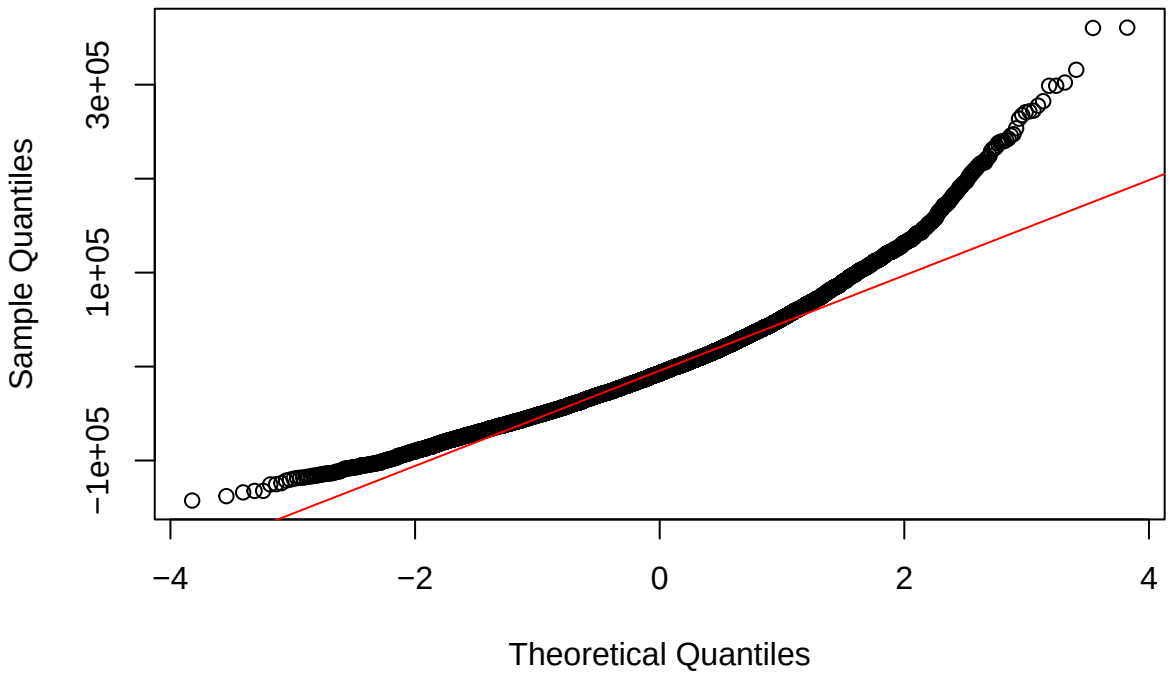
Now that we have the model that we have found from AIC we can test the assumptions of the model as well as performing cross validation to test the valididity of the model. We can also perform an anova test to see if our predictors are significant.

```r
# Assuming your final model is stored in 'step_model' and your data frame is 'joined_df'

# Alternatively, you can plot individual diagnostic plots:
# Residuals vs Fitted
plot(step_model$fitted.values, step_model$residuals,
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```




```r
# Q-Q Plot for Normality of Residuals
qqnorm(step_model$residuals, main = "Q-Q Plot")
qqline(step_model$residuals, col = "red")
```




```
# Testing for heteroscedasticity using the Breusch-Pagan test
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.3.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bp_test <- bptest(step_model)
print(bp_test)  # p-value < 0.05 indicates potential heteroscedasticity
```

```
##
##  studentized Breusch-Pagan test
##
## data:  step_model
## BP = 220.39, df = 84, p-value = 3.508e-14
```

```r
# 2. Model Validation: Cross-Validation -----------------------------

# Using the caret package for 10-fold cross-validation
# Install caret if needed: install.packages("caret")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
set.seed(123)  # for reproducibility

# Define training control for 10-fold cross-validation
train_control <- trainControl(method = "cv", number = 10)

# Refit the model using caret's train() function.
# Here, we use the same predictors that were selected in your step_model.
cv_model <- train(salary_in_usd ~ experience_level + job_title + employee_residence +
                  work_setting + company_size,
                  data = joined_df,
                  method = "lm",
                  trControl = train_control)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases

## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```
print(cv_model)
```

```
## Linear Regression
##
## 7575 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 6816, 6817, 6819, 6817, 6817, 6818, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   54993.62  0.3514908  42142.88
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
# 3. Theoretical Considerations -------------------------------------
# Although not strictly "code," here are some steps to integrate theory:

# - Review each predictor's significance and consider if domain knowledge suggests its retention.
#   For example, even if certain levels of a categorical predictor have high p-values, the variable m
#   important overall. You could run an ANOVA to test the overall significance of categorical variabl
anova_result <- anova(step_model)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Response: salary_in_usd
##                       Df     Sum Sq    Mean Sq  F value    Pr(>F)
## experience_level       3 5.3581e+12 1.7860e+12 567.2167 < 2.2e-16 ***
## employee_residence    70 3.7700e+12 5.3857e+10  17.1040 < 2.2e-16 ***
## work_setting           2 7.3645e+10 3.6822e+10  11.6942 8.496e-06 ***
## company_size           2 2.4152e+10 1.2076e+10   3.8352   0.02164 *
## job_category           7 2.4466e+12 3.4951e+11 110.9984 < 2.2e-16 ***
## Residuals           7490 2.3584e+13 3.1488e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The cross validation tells us that our model is not over fitting since the R^2 is only about 35%, but also suggests that we can improve our model because the RMSE is quite high at 54993, so we may need to add more interactions or do some non linear transformations of the data.

The residual plot shows a funnel shape, which suggests hetereoscadascity of the variance and the residuals do not appear to be centered at zero, which indicates that there is bias. The plots indicate that we may want to use the log of the salary instead of salary as our response.

we can now run the step AIC function again using the log(salary) as our response and perform the same model diagnostic tests as above. We also add an interaction term between experience_level and company size.

```
model_df$log_salary <- log(model_df$salary_in_usd)
log_full_model <- lm(log_salary ~ . - salary_in_usd -log_salary, data = model_df)
step_log_model <- stepAIC(log_full_model, direction = "both", trace = FALSE)
summary(step_log_model)
```

```
##
## Call:
## lm(formula = log_salary ~ experience_level + employee_residence +
##     work_setting + company_location + company_size + job_category +
##     exp_cost_int, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83514 -0.24654  0.00905  0.25671  1.82446
##
## Coefficients: (22 not defined because of singularities)
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          11.4239475  0.3869696  29.522  < 2e-16
## experience_levelMid-level             0.6051081  0.0614443   9.848  < 2e-16
## experience_levelSenior                1.1984545  0.1200107   9.986  < 2e-16
## experience_levelExecutive             1.7708632  0.1814623   9.759  < 2e-16
## employee_residenceArgentina          -0.8966058  0.4417041  -2.030 0.042404
## employee_residenceArmenia            -1.0432066  0.7198251  -1.449 0.147310
## employee_residenceAustralia          -0.1442537  0.4363268  -0.331 0.740949
## employee_residenceAustria            -0.4424280  0.4918523  -0.900 0.368408
## employee_residenceBelgium            -0.1709004  0.4684190  -0.365 0.715237
## employee_residenceBrazil             -0.6401061  0.4717178  -1.357 0.174832
## employee_residenceBulgaria           -0.6838161  0.5420249  -1.262 0.207134
## employee_residenceCanada              0.0916615  0.4076604   0.225 0.822104
## employee_residenceChile               0.3288985  0.5407421   0.608 0.543050
## employee_residenceChina               0.6445711  0.5390874   1.196 0.231863
## employee_residenceColombia           -1.1585325  0.3984919  -2.907 0.003656
## employee_residenceCosta Rica         -1.2636224  0.5382665  -2.348 0.018922
## employee_residenceCroatia            -1.1718682  0.5559874  -2.108 0.035088
## employee_residenceCyprus             -0.9083805  0.5383234  -1.687 0.091563
## employee_residenceDenmark            -1.1572665  0.5629912  -2.056 0.039859
## employee_residenceDominican Republic  0.2158843  0.5599946   0.386 0.699869
## employee_residenceEcuador            -1.9359942  0.5394129  -3.589 0.000334
## employee_residenceEgypt              -0.8033148  0.5410384  -1.485 0.137648
## employee_residenceEstonia            -1.1332369  0.4025025  -2.815 0.004883
## employee_residenceFinland            -0.1110453  0.6181213  -0.180 0.857432
## employee_residenceFrance             -0.3144902  0.4099283  -0.767 0.442997
## employee_residenceGeorgia            -0.3280403  0.7150748  -0.459 0.646427
## employee_residenceGermany            -0.2957378  0.4062340  -0.728 0.466637
## employee_residenceGhana              -0.9191088  0.5398412  -1.703 0.088693
## employee_residenceGreece             -0.9134857  0.4989926  -1.831 0.067191
```

```
## employee_residenceHungary              -1.0597325  0.7754677  -1.367 0.171801
## employee_residenceIndia                -0.7262642  0.4070442  -1.784 0.074425
## employee_residenceIndonesia            -1.8335549  0.5396069  -3.398 0.000682
## employee_residenceIraq                  0.0476417  0.5390613   0.088 0.929578
## employee_residenceIreland              -0.3290022  0.3982367  -0.826 0.408747
## employee_residenceItaly                -1.0011483  0.4179822  -2.395 0.016636
## employee_residenceJapan                -0.1582719  0.5829432  -0.272 0.786010
## employee_residenceKenya                -0.5097414  0.4686901  -1.088 0.276812
## employee_residenceKuwait               -0.3946021  0.5383597  -0.733 0.463599
## employee_residenceLatvia               -1.1623238  0.3951990  -2.941 0.003280
## employee_residenceLebanon              -0.2119294  0.4671761  -0.454 0.650102
## employee_residenceLithuania            -0.8652648  0.3934438  -2.199 0.027894
## employee_residenceLuxembourg            0.2355330  0.6227943   0.378 0.705302
## employee_residenceMalaysia             -0.1533620  0.5432243  -0.282 0.777707
## employee_residenceMalta                -0.6824672  0.4402743  -1.550 0.121161
## employee_residenceMauritius             0.1336356  0.5403748   0.247 0.804682
## employee_residenceMexico               -0.2945634  0.5933759  -0.496 0.619613
## employee_residenceNetherlands          -0.6333319  0.4539924  -1.395 0.163049
## employee_residenceNew Zealand           0.1319548  0.4127948   0.320 0.749234
## employee_residenceNigeria              -2.2570308  0.4525679  -4.987 6.27e-07
## employee_residenceOman                 -1.7138488  0.5399372  -3.174 0.001509
## employee_residencePakistan             -1.1362484  0.4380799  -2.594 0.009513
## employee_residencePeru                 -1.1398351  0.5909854  -1.929 0.053806
## employee_residencePhilippines          -1.0362386  0.4576860  -2.264 0.023598
## employee_residencePoland               -1.0021335  0.4485323  -2.234 0.025496
## employee_residencePortugal             -0.7245526  0.4438321  -1.632 0.102618
## employee_residencePuerto Rico           0.3851056  0.5398837   0.713 0.475675
## employee_residenceQatar                 0.6229802  0.5391288   1.156 0.247910
## employee_residenceRomania              -1.2605074  0.4544713  -2.774 0.005558
## employee_residenceSaudi Arabia         -0.1362726  0.4403185  -0.309 0.756960
## employee_residenceSerbia               -1.6076120  0.5572643  -2.885 0.003928
## employee_residenceSingapore            -0.1200237  0.4286166  -0.280 0.779466
## employee_residenceSlovenia             -0.9739017  0.4121673  -2.363 0.018159
## employee_residenceSouth Africa         -1.2236238  0.3986793  -3.069 0.002154
## employee_residenceSpain                -0.5096022  0.4142663  -1.230 0.218687
## employee_residenceSweden                0.4307907  0.6176737   0.697 0.485549
## employee_residenceSwitzerland           0.3416632  0.4192883   0.815 0.415176
## employee_residenceThailand             -1.5733998  0.5424035  -2.901 0.003733
## employee_residenceTunisia              -0.7921261  0.5000858  -1.584 0.113241
## employee_residenceUganda               -1.2906698  0.5388918  -2.395 0.016643
## employee_residenceUkraine              -1.4889631  0.5934124  -2.509 0.012123
## employee_residenceUnited Arab Emirates -0.5777536  0.4276572  -1.351 0.176745
## employee_residenceUnited Kingdom       -0.3482457  0.4208842  -0.827 0.408029
## employee_residenceUnited States         0.1438742  0.3842045   0.374 0.708063
## employee_residenceUzbekistan           -0.4897141  0.4671600  -1.048 0.294544
## work_settingIn-person                   0.1514859  0.0329827   4.593 4.44e-06
## work_settingRemote                      0.1315038  0.0328646   4.001 6.36e-05
## company_locationAmerican Samoa         -0.2252636  0.3027178  -0.744 0.456817
## company_locationArgentina              -0.3518009  0.2661645  -1.322 0.186294
## company_locationArmenia                -0.0731696  0.7214157  -0.101 0.919216
## company_locationAustralia               0.1999240  0.2012975   0.993 0.320656
## company_locationAustria                -0.0882035  0.3075968  -0.287 0.774311
## company_locationBelgium                -0.2283240  0.3302515  -0.691 0.489358
```

```
## company_locationBrazil               -0.7520056  0.2837539  -2.650 0.008061
## company_locationCanada               -0.1402720  0.1381194  -1.016 0.309860
## company_locationColombia                     NA         NA      NA       NA
## company_locationCroatia               0.4754002  0.4605707   1.032 0.302013
## company_locationCzechia              -0.2003822  0.3528755  -0.568 0.570150
## company_locationDenmark               0.2804419  0.3476440   0.807 0.419869
## company_locationEcuador                      NA         NA      NA       NA
## company_locationEgypt                -0.0499617  0.3980701  -0.126 0.900123
## company_locationEstonia                      NA         NA      NA       NA
## company_locationFinland              -0.5916273  0.4338151  -1.364 0.172679
## company_locationFrance               -0.2032827  0.1503470  -1.352 0.176387
## company_locationGermany              -0.0580567  0.1350768  -0.430 0.667350
## company_locationGhana                -0.8583373  0.4398153  -1.952 0.051025
## company_locationGibraltar            -0.2041695  0.4134195  -0.494 0.621423
## company_locationGreece               -0.1370524  0.3406409  -0.402 0.687448
## company_locationHungary               0.3606529  0.5563304   0.648 0.516830
## company_locationIndia                -0.5992417  0.1673561  -3.581 0.000345
## company_locationIndonesia                    NA         NA      NA       NA
## company_locationIraq                         NA         NA      NA       NA
## company_locationIreland                      NA         NA      NA       NA
## company_locationIsrael                0.2328617  0.4259912   0.547 0.584646
## company_locationItaly                -0.0230530  0.1959674  -0.118 0.906358
## company_locationJapan                 0.0491684  0.3808029   0.129 0.897268
## company_locationKenya                        NA         NA      NA       NA
## company_locationLatvia                       NA         NA      NA       NA
## company_locationLebanon                      NA         NA      NA       NA
## company_locationLithuania                    NA         NA      NA       NA
## company_locationLuxembourg           -0.4843368  0.3053978  -1.586 0.112800
## company_locationMalaysia             -0.7882387  0.5843723  -1.349 0.177422
## company_locationMalta                        NA         NA      NA       NA
## company_locationMauritius                    NA         NA      NA       NA
## company_locationMexico               -0.6146408  0.4426543  -1.389 0.165016
## company_locationNetherlands           0.1738477  0.2472661   0.703 0.482028
## company_locationNew Zealand                  NA         NA      NA       NA
## company_locationNigeria               1.2022978  0.2699684   4.453 8.57e-06
## company_locationOman                         NA         NA      NA       NA
## company_locationPakistan             -0.5700558  0.3258662  -1.749 0.080271
## company_locationPhilippines          -0.3014640  0.3019282  -0.998 0.318087
## company_locationPoland                0.0532493  0.2455165   0.217 0.828302
## company_locationPortugal             -0.4968318  0.2380396  -2.087 0.036906
## company_locationPuerto Rico          -0.3865490  0.4277391  -0.904 0.366182
## company_locationQatar                        NA         NA      NA       NA
## company_locationRomania              -0.1949592  0.2746030  -0.710 0.477747
## company_locationRussian Federation   -0.9504914  0.4794661  -1.982 0.047472
## company_locationSaudi Arabia                 NA         NA      NA       NA
## company_locationSingapore                    NA         NA      NA       NA
## company_locationSlovenia                     NA         NA      NA       NA
## company_locationSouth Africa                 NA         NA      NA       NA
## company_locationSpain                -0.5768745  0.1640186  -3.517 0.000439
## company_locationSweden               -0.2610805  0.4041355  -0.646 0.518285
## company_locationSwitzerland                  NA         NA      NA       NA
## company_locationThailand             -0.5575994  0.5411601  -1.030 0.302866
## company_locationUkraine               0.3420310  0.4748747   0.720 0.471390
```

```
## company_locationUnited Arab Emirates          NA         NA      NA       NA
## company_locationUnited Kingdom         -0.0373958  0.1733210  -0.216 0.829180
## company_locationUnited States                 NA         NA      NA       NA
## company_sizeM                          0.1840721  0.0359652   5.118 3.16e-07
## company_sizeL                          0.1760413  0.0385807   4.563 5.13e-06
## job_categoryData Analyst              -0.1573551  0.0575860  -2.733 0.006300
## job_categoryData Engineer              0.0050013  0.0574307   0.087 0.930607
## job_categoryData Scientist             0.0615928  0.0574908   1.071 0.284046
## job_categoryLeadership                -0.0905907  0.0627574  -1.444 0.148920
## job_categoryMachine Learning           0.2001919  0.0579276   3.456 0.000552
## job_categoryOther                     -0.2866873  0.0667811  -4.293 1.79e-05
## job_categoryResearch                   0.2177361  0.0618577   3.520 0.000434
## exp_cost_int                          -0.0054217  0.0008793  -6.166 7.39e-10
##
## (Intercept)                           ***
## experience_levelMid-level             ***
## experience_levelSenior                ***
## experience_levelExecutive             ***
## employee_residenceArgentina           *
## employee_residenceArmenia
## employee_residenceAustralia
## employee_residenceAustria
## employee_residenceBelgium
## employee_residenceBrazil
## employee_residenceBulgaria
## employee_residenceCanada
## employee_residenceChile
## employee_residenceChina
## employee_residenceColombia            **
## employee_residenceCosta Rica          *
## employee_residenceCroatia             *
## employee_residenceCyprus              .
## employee_residenceDenmark             *
## employee_residenceDominican Republic
## employee_residenceEcuador             ***
## employee_residenceEgypt
## employee_residenceEstonia             **
## employee_residenceFinland
## employee_residenceFrance
## employee_residenceGeorgia
## employee_residenceGermany
## employee_residenceGhana               .
## employee_residenceGreece              .
## employee_residenceHungary
## employee_residenceIndia               .
## employee_residenceIndonesia           ***
## employee_residenceIraq
## employee_residenceIreland
## employee_residenceItaly               *
## employee_residenceJapan
## employee_residenceKenya
## employee_residenceKuwait
## employee_residenceLatvia              **
```

```
## employee_residenceLebanon
## employee_residenceLithuania           *
## employee_residenceLuxembourg
## employee_residenceMalaysia
## employee_residenceMalta
## employee_residenceMauritius
## employee_residenceMexico
## employee_residenceNetherlands
## employee_residenceNew Zealand
## employee_residenceNigeria             ***
## employee_residenceOman                **
## employee_residencePakistan            **
## employee_residencePeru                .
## employee_residencePhilippines         *
## employee_residencePoland              *
## employee_residencePortugal
## employee_residencePuerto Rico
## employee_residenceQatar
## employee_residenceRomania             **
## employee_residenceSaudi Arabia
## employee_residenceSerbia              **
## employee_residenceSingapore
## employee_residenceSlovenia            *
## employee_residenceSouth Africa        **
## employee_residenceSpain
## employee_residenceSweden
## employee_residenceSwitzerland
## employee_residenceThailand            **
## employee_residenceTunisia
## employee_residenceUganda              *
## employee_residenceUkraine             *
## employee_residenceUnited Arab Emirates
## employee_residenceUnited Kingdom
## employee_residenceUnited States
## employee_residenceUzbekistan
## work_settingIn-person                 ***
## work_settingRemote                    ***
## company_locationAmerican Samoa
## company_locationArgentina
## company_locationArmenia
## company_locationAustralia
## company_locationAustria
## company_locationBelgium
## company_locationBrazil                **
## company_locationCanada
## company_locationColombia
## company_locationCroatia
## company_locationCzechia
## company_locationDenmark
## company_locationEcuador
## company_locationEgypt
## company_locationEstonia
## company_locationFinland
```

```
## company_locationFrance
## company_locationGermany
## company_locationGhana                  .
## company_locationGibraltar
## company_locationGreece
## company_locationHungary
## company_locationIndia                  ***
## company_locationIndonesia
## company_locationIraq
## company_locationIreland
## company_locationIsrael
## company_locationItaly
## company_locationJapan
## company_locationKenya
## company_locationLatvia
## company_locationLebanon
## company_locationLithuania
## company_locationLuxembourg
## company_locationMalaysia
## company_locationMalta
## company_locationMauritius
## company_locationMexico
## company_locationNetherlands
## company_locationNew Zealand
## company_locationNigeria                ***
## company_locationOman
## company_locationPakistan               .
## company_locationPhilippines
## company_locationPoland
## company_locationPortugal               *
## company_locationPuerto Rico
## company_locationQatar
## company_locationRomania
## company_locationRussian Federation     *
## company_locationSaudi Arabia
## company_locationSingapore
## company_locationSlovenia
## company_locationSouth Africa
## company_locationSpain                  ***
## company_locationSweden
## company_locationSwitzerland
## company_locationThailand
## company_locationUkraine
## company_locationUnited Arab Emirates
## company_locationUnited Kingdom
## company_locationUnited States
## company_sizeM                          ***
## company_sizeL                          ***
## job_categoryData Analyst               **
## job_categoryData Engineer
## job_categoryData Scientist
## job_categoryLeadership
## job_categoryMachine Learning           ***
```

```
## job_categoryOther                       ***
## job_categoryResearch                    ***
## exp_cost_int                            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3802 on 7449 degrees of freedom
## Multiple R-squared:  0.4596, Adjusted R-squared:  0.4505
## F-statistic: 50.68 on 125 and 7449 DF,  p-value: < 2.2e-16
```

```r
model_with_interaction <- update(step_log_model,
                                 . ~ . + experience_level:company_size)

# 2. Perform stepwise selection again starting from the updated model
step_log_model_interact <- stepAIC(model_with_interaction,
                                   direction = "both",
                                   trace = FALSE)

# 3. Review the summary of the new model
summary(step_log_model_interact)
```

```
##
## Call:
## lm(formula = log_salary ~ experience_level + employee_residence +
##     work_setting + company_location + company_size + job_category +
##     exp_cost_int, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83514 -0.24654  0.00905  0.25671  1.82446
##
## Coefficients: (22 not defined because of singularities)
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           11.4239475  0.3869696  29.522  < 2e-16
## experience_levelMid-level              0.6051081  0.0614443   9.848  < 2e-16
## experience_levelSenior                 1.1984545  0.1200107   9.986  < 2e-16
## experience_levelExecutive              1.7708632  0.1814623   9.759  < 2e-16
## employee_residenceArgentina           -0.8966058  0.4417041  -2.030 0.042404
## employee_residenceArmenia             -1.0432066  0.7198251  -1.449 0.147310
## employee_residenceAustralia           -0.1442537  0.4363268  -0.331 0.740949
## employee_residenceAustria             -0.4424280  0.4918523  -0.900 0.368408
## employee_residenceBelgium             -0.1709004  0.4684190  -0.365 0.715237
## employee_residenceBrazil              -0.6401061  0.4717178  -1.357 0.174832
## employee_residenceBulgaria            -0.6838161  0.5420249  -1.262 0.207134
## employee_residenceCanada               0.0916615  0.4076604   0.225 0.822104
## employee_residenceChile                0.3288985  0.5407421   0.608 0.543050
## employee_residenceChina                0.6445711  0.5390874   1.196 0.231863
## employee_residenceColombia            -1.1585325  0.3984919  -2.907 0.003656
## employee_residenceCosta Rica          -1.2636224  0.5382665  -2.348 0.018922
## employee_residenceCroatia             -1.1718682  0.5559874  -2.108 0.035088
## employee_residenceCyprus              -0.9083805  0.5383234  -1.687 0.091563
## employee_residenceDenmark             -1.1572665  0.5629912  -2.056 0.039859
## employee_residenceDominican Republic   0.2158843  0.5599946   0.386 0.699869
```

```
## employee_residenceEcuador              -1.9359942  0.5394129  -3.589 0.000334
## employee_residenceEgypt                -0.8033148  0.5410384  -1.485 0.137648
## employee_residenceEstonia              -1.1332369  0.4025025  -2.815 0.004883
## employee_residenceFinland              -0.1110453  0.6181213  -0.180 0.857432
## employee_residenceFrance               -0.3144902  0.4099283  -0.767 0.442997
## employee_residenceGeorgia              -0.3280403  0.7150748  -0.459 0.646427
## employee_residenceGermany              -0.2957378  0.4062340  -0.728 0.466637
## employee_residenceGhana                -0.9191088  0.5398412  -1.703 0.088693
## employee_residenceGreece               -0.9134857  0.4989926  -1.831 0.067191
## employee_residenceHungary              -1.0597325  0.7754677  -1.367 0.171801
## employee_residenceIndia                -0.7262642  0.4070442  -1.784 0.074425
## employee_residenceIndonesia            -1.8335549  0.5396069  -3.398 0.000682
## employee_residenceIraq                  0.0476417  0.5390613   0.088 0.929578
## employee_residenceIreland              -0.3290022  0.3982367  -0.826 0.408747
## employee_residenceItaly                -1.0011483  0.4179822  -2.395 0.016636
## employee_residenceJapan                -0.1582719  0.5829432  -0.272 0.786010
## employee_residenceKenya                -0.5097414  0.4686901  -1.088 0.276812
## employee_residenceKuwait               -0.3946021  0.5383597  -0.733 0.463599
## employee_residenceLatvia               -1.1623238  0.3951990  -2.941 0.003280
## employee_residenceLebanon              -0.2119294  0.4671761  -0.454 0.650102
## employee_residenceLithuania            -0.8652648  0.3934438  -2.199 0.027894
## employee_residenceLuxembourg            0.2355330  0.6227943   0.378 0.705302
## employee_residenceMalaysia             -0.1533620  0.5432243  -0.282 0.777707
## employee_residenceMalta                -0.6824672  0.4402743  -1.550 0.121161
## employee_residenceMauritius             0.1336356  0.5403748   0.247 0.804682
## employee_residenceMexico               -0.2945634  0.5933759  -0.496 0.619613
## employee_residenceNetherlands          -0.6333319  0.4539924  -1.395 0.163049
## employee_residenceNew Zealand           0.1319548  0.4127948   0.320 0.749234
## employee_residenceNigeria              -2.2570308  0.4525679  -4.987 6.27e-07
## employee_residenceOman                 -1.7138488  0.5399372  -3.174 0.001509
## employee_residencePakistan             -1.1362484  0.4380799  -2.594 0.009513
## employee_residencePeru                 -1.1398351  0.5909854  -1.929 0.053806
## employee_residencePhilippines          -1.0362386  0.4576860  -2.264 0.023598
## employee_residencePoland               -1.0021335  0.4485323  -2.234 0.025496
## employee_residencePortugal             -0.7245526  0.4438321  -1.632 0.102618
## employee_residencePuerto Rico           0.3851056  0.5398837   0.713 0.475675
## employee_residenceQatar                 0.6229802  0.5391288   1.156 0.247910
## employee_residenceRomania              -1.2605074  0.4544713  -2.774 0.005558
## employee_residenceSaudi Arabia         -0.1362726  0.4403185  -0.309 0.756960
## employee_residenceSerbia               -1.6076120  0.5572643  -2.885 0.003928
## employee_residenceSingapore            -0.1200237  0.4286166  -0.280 0.779466
## employee_residenceSlovenia             -0.9739017  0.4121673  -2.363 0.018159
## employee_residenceSouth Africa         -1.2236238  0.3986793  -3.069 0.002154
## employee_residenceSpain                -0.5096022  0.4142663  -1.230 0.218687
## employee_residenceSweden                0.4307907  0.6176737   0.697 0.485549
## employee_residenceSwitzerland           0.3416632  0.4192883   0.815 0.415176
## employee_residenceThailand             -1.5733998  0.5424035  -2.901 0.003733
## employee_residenceTunisia              -0.7921261  0.5000858  -1.584 0.113241
## employee_residenceUganda               -1.2906698  0.5388918  -2.395 0.016643
## employee_residenceUkraine              -1.4889631  0.5934124  -2.509 0.012123
## employee_residenceUnited Arab Emirates -0.5777536  0.4276572  -1.351 0.176745
## employee_residenceUnited Kingdom       -0.3482457  0.4208842  -0.827 0.408029
## employee_residenceUnited States         0.1438742  0.3842045   0.374 0.708063
```

```
## employee_residenceUzbekistan          -0.4897141  0.4671600  -1.048 0.294544
## work_settingIn-person                  0.1514859  0.0329827   4.593 4.44e-06
## work_settingRemote                     0.1315038  0.0328646   4.001 6.36e-05
## company_locationAmerican Samoa        -0.2252636  0.3027178  -0.744 0.456817
## company_locationArgentina             -0.3518009  0.2661645  -1.322 0.186294
## company_locationArmenia               -0.0731696  0.7214157  -0.101 0.919216
## company_locationAustralia              0.1999240  0.2012975   0.993 0.320656
## company_locationAustria               -0.0882035  0.3075968  -0.287 0.774311
## company_locationBelgium               -0.2283240  0.3302515  -0.691 0.489358
## company_locationBrazil                -0.7520056  0.2837539  -2.650 0.008061
## company_locationCanada                -0.1402720  0.1381194  -1.016 0.309860
## company_locationColombia                      NA         NA      NA       NA
## company_locationCroatia                0.4754002  0.4605707   1.032 0.302013
## company_locationCzechia               -0.2003822  0.3528755  -0.568 0.570150
## company_locationDenmark                0.2804419  0.3476440   0.807 0.419869
## company_locationEcuador                       NA         NA      NA       NA
## company_locationEgypt                 -0.0499617  0.3980701  -0.126 0.900123
## company_locationEstonia                       NA         NA      NA       NA
## company_locationFinland               -0.5916273  0.4338151  -1.364 0.172679
## company_locationFrance                -0.2032827  0.1503470  -1.352 0.176387
## company_locationGermany               -0.0580567  0.1350768  -0.430 0.667350
## company_locationGhana                 -0.8583373  0.4398153  -1.952 0.051025
## company_locationGibraltar             -0.2041695  0.4134195  -0.494 0.621423
## company_locationGreece                -0.1370524  0.3406409  -0.402 0.687448
## company_locationHungary                0.3606529  0.5563304   0.648 0.516830
## company_locationIndia                 -0.5992417  0.1673561  -3.581 0.000345
## company_locationIndonesia                     NA         NA      NA       NA
## company_locationIraq                          NA         NA      NA       NA
## company_locationIreland                       NA         NA      NA       NA
## company_locationIsrael                 0.2328617  0.4259912   0.547 0.584646
## company_locationItaly                 -0.0230530  0.1959674  -0.118 0.906358
## company_locationJapan                  0.0491684  0.3808029   0.129 0.897268
## company_locationKenya                         NA         NA      NA       NA
## company_locationLatvia                        NA         NA      NA       NA
## company_locationLebanon                       NA         NA      NA       NA
## company_locationLithuania                     NA         NA      NA       NA
## company_locationLuxembourg            -0.4843368  0.3053978  -1.586 0.112800
## company_locationMalaysia              -0.7882387  0.5843723  -1.349 0.177422
## company_locationMalta                         NA         NA      NA       NA
## company_locationMauritius                     NA         NA      NA       NA
## company_locationMexico                -0.6146408  0.4426543  -1.389 0.165016
## company_locationNetherlands            0.1738477  0.2472661   0.703 0.482028
## company_locationNew Zealand                   NA         NA      NA       NA
## company_locationNigeria                1.2022978  0.2699684   4.453 8.57e-06
## company_locationOman                          NA         NA      NA       NA
## company_locationPakistan              -0.5700558  0.3258662  -1.749 0.080271
## company_locationPhilippines           -0.3014640  0.3019282  -0.998 0.318087
## company_locationPoland                 0.0532493  0.2455165   0.217 0.828302
## company_locationPortugal              -0.4968318  0.2380396  -2.087 0.036906
## company_locationPuerto Rico           -0.3865490  0.4277391  -0.904 0.366182
## company_locationQatar                         NA         NA      NA       NA
## company_locationRomania               -0.1949592  0.2746030  -0.710 0.477747
## company_locationRussian Federation    -0.9504914  0.4794661  -1.982 0.047472
```

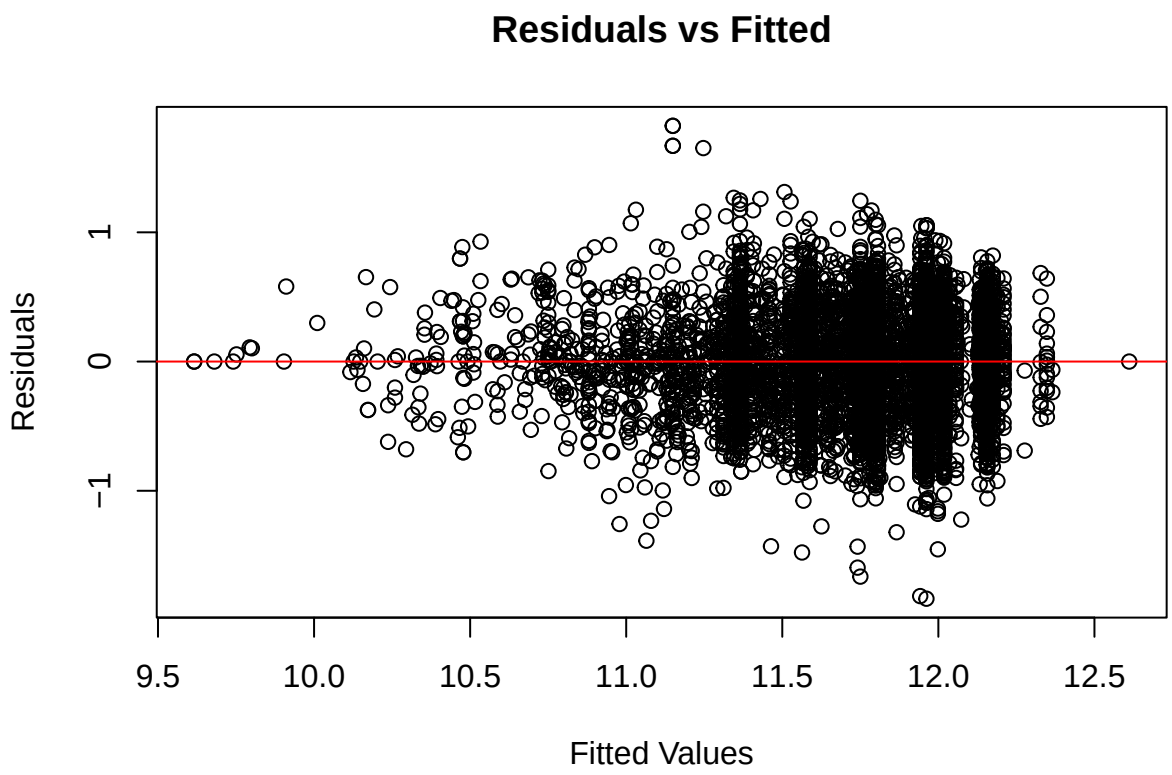
```
## company_locationSaudi Arabia                   NA         NA      NA       NA
## company_locationSingapore                       NA         NA      NA       NA
## company_locationSlovenia                        NA         NA      NA       NA
## company_locationSouth Africa                    NA         NA      NA       NA
## company_locationSpain                   -0.5768745  0.1640186  -3.517 0.000439
## company_locationSweden                  -0.2610805  0.4041355  -0.646 0.518285
## company_locationSwitzerland                     NA         NA      NA       NA
## company_locationThailand                -0.5575994  0.5411601  -1.030 0.302866
## company_locationUkraine                  0.3420310  0.4748747   0.720 0.471390
## company_locationUnited Arab Emirates            NA         NA      NA       NA
## company_locationUnited Kingdom          -0.0373958  0.1733210  -0.216 0.829180
## company_locationUnited States                   NA         NA      NA       NA
## company_sizeM                            0.1840721  0.0359652   5.118 3.16e-07
## company_sizeL                            0.1760413  0.0385807   4.563 5.13e-06
## job_categoryData Analyst                -0.1573551  0.0575860  -2.733 0.006300
## job_categoryData Engineer                0.0050013  0.0574307   0.087 0.930607
## job_categoryData Scientist               0.0615928  0.0574908   1.071 0.284046
## job_categoryLeadership                  -0.0905907  0.0627574  -1.444 0.148920
## job_categoryMachine Learning             0.2001919  0.0579276   3.456 0.000552
## job_categoryOther                       -0.2866873  0.0667811  -4.293 1.79e-05
## job_categoryResearch                     0.2177361  0.0618577   3.520 0.000434
## exp_cost_int                            -0.0054217  0.0008793  -6.166 7.39e-10
##
## (Intercept)                             ***
## experience_levelMid-level               ***
## experience_levelSenior                  ***
## experience_levelExecutive               ***
## employee_residenceArgentina             *
## employee_residenceArmenia
## employee_residenceAustralia
## employee_residenceAustria
## employee_residenceBelgium
## employee_residenceBrazil
## employee_residenceBulgaria
## employee_residenceCanada
## employee_residenceChile
## employee_residenceChina
## employee_residenceColombia              **
## employee_residenceCosta Rica            *
## employee_residenceCroatia               *
## employee_residenceCyprus                .
## employee_residenceDenmark               *
## employee_residenceDominican Republic
## employee_residenceEcuador               ***
## employee_residenceEgypt
## employee_residenceEstonia               **
## employee_residenceFinland
## employee_residenceFrance
## employee_residenceGeorgia
## employee_residenceGermany
## employee_residenceGhana                 .
## employee_residenceGreece                .
## employee_residenceHungary
```

```
## employee_residenceIndia               .
## employee_residenceIndonesia           ***
## employee_residenceIraq
## employee_residenceIreland
## employee_residenceItaly               *
## employee_residenceJapan
## employee_residenceKenya
## employee_residenceKuwait
## employee_residenceLatvia              **
## employee_residenceLebanon
## employee_residenceLithuania           *
## employee_residenceLuxembourg
## employee_residenceMalaysia
## employee_residenceMalta
## employee_residenceMauritius
## employee_residenceMexico
## employee_residenceNetherlands
## employee_residenceNew Zealand
## employee_residenceNigeria             ***
## employee_residenceOman                **
## employee_residencePakistan            **
## employee_residencePeru                .
## employee_residencePhilippines         *
## employee_residencePoland              *
## employee_residencePortugal
## employee_residencePuerto Rico
## employee_residenceQatar
## employee_residenceRomania             **
## employee_residenceSaudi Arabia
## employee_residenceSerbia              **
## employee_residenceSingapore
## employee_residenceSlovenia            *
## employee_residenceSouth Africa        **
## employee_residenceSpain
## employee_residenceSweden
## employee_residenceSwitzerland
## employee_residenceThailand            **
## employee_residenceTunisia
## employee_residenceUganda              *
## employee_residenceUkraine             *
## employee_residenceUnited Arab Emirates
## employee_residenceUnited Kingdom
## employee_residenceUnited States
## employee_residenceUzbekistan
## work_settingIn-person                 ***
## work_settingRemote                    ***
## company_locationAmerican Samoa
## company_locationArgentina
## company_locationArmenia
## company_locationAustralia
## company_locationAustria
## company_locationBelgium
## company_locationBrazil                **
```

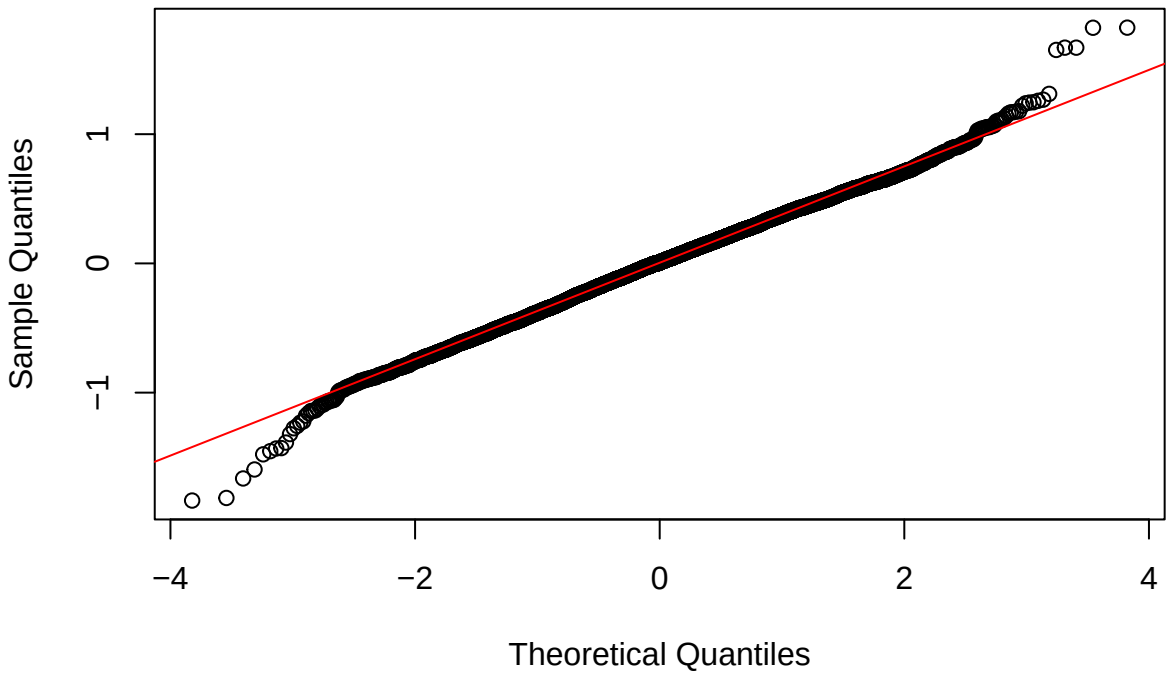
```
## company_locationCanada
## company_locationColombia
## company_locationCroatia
## company_locationCzechia
## company_locationDenmark
## company_locationEcuador
## company_locationEgypt
## company_locationEstonia
## company_locationFinland
## company_locationFrance
## company_locationGermany
## company_locationGhana                .
## company_locationGibraltar
## company_locationGreece
## company_locationHungary
## company_locationIndia                ***
## company_locationIndonesia
## company_locationIraq
## company_locationIreland
## company_locationIsrael
## company_locationItaly
## company_locationJapan
## company_locationKenya
## company_locationLatvia
## company_locationLebanon
## company_locationLithuania
## company_locationLuxembourg
## company_locationMalaysia
## company_locationMalta
## company_locationMauritius
## company_locationMexico
## company_locationNetherlands
## company_locationNew Zealand
## company_locationNigeria              ***
## company_locationOman
## company_locationPakistan             .
## company_locationPhilippines
## company_locationPoland
## company_locationPortugal             *
## company_locationPuerto Rico
## company_locationQatar
## company_locationRomania
## company_locationRussian Federation   *
## company_locationSaudi Arabia
## company_locationSingapore
## company_locationSlovenia
## company_locationSouth Africa
## company_locationSpain                ***
## company_locationSweden
## company_locationSwitzerland
## company_locationThailand
## company_locationUkraine
## company_locationUnited Arab Emirates
```

```
## company_locationUnited Kingdom
## company_locationUnited States
## company_sizeM                          ***
## company_sizeL                          ***
## job_categoryData Analyst               **
## job_categoryData Engineer
## job_categoryData Scientist
## job_categoryLeadership
## job_categoryMachine Learning           ***
## job_categoryOther                      ***
## job_categoryResearch                   ***
## exp_cost_int                           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3802 on 7449 degrees of freedom
## Multiple R-squared:  0.4596, Adjusted R-squared:  0.4505
## F-statistic: 50.68 on 125 and 7449 DF,  p-value: < 2.2e-16
```

```r
plot(step_log_model_interact$fitted.values, step_log_model_interact$residuals,
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```




```r
qqnorm(step_log_model_interact$residuals, main = "Q-Q Plot")
qqline(step_log_model_interact$residuals, col = "red")
```




```
bp_test <- bptest(step_log_model_interact)
print(bp_test)  # p-value < 0.05 indicates potential heteroscedasticity
```

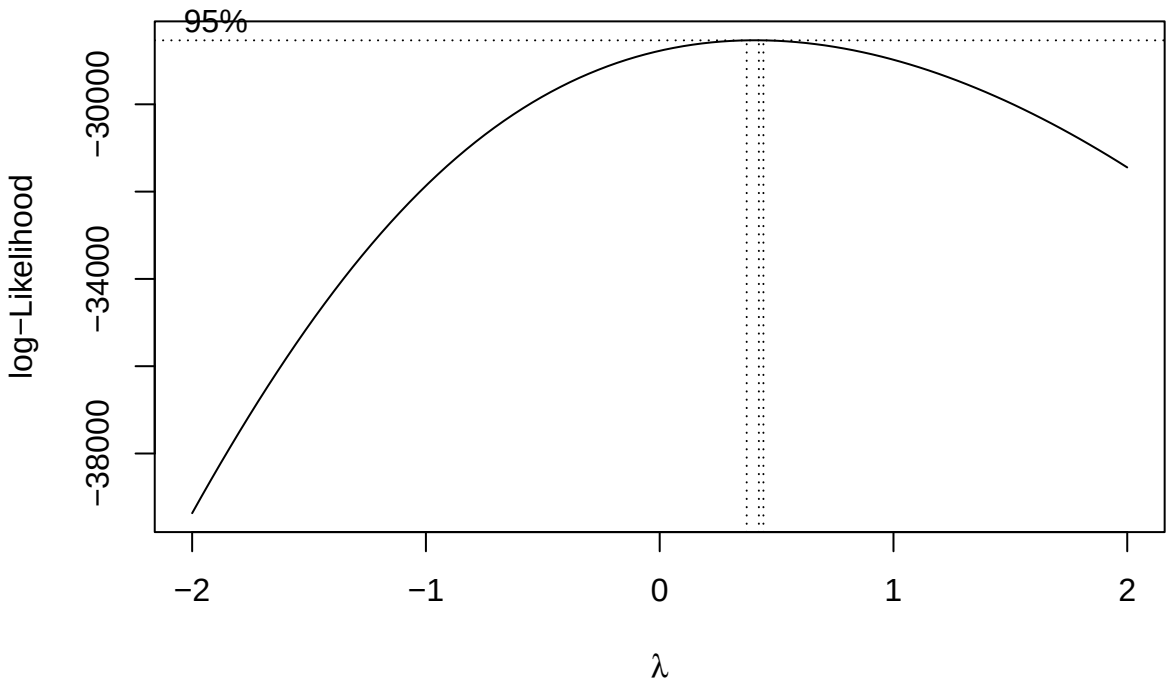
```
##
##  studentized Breusch-Pagan test
##
## data:  step_log_model_interact
## BP = 241.2, df = 125, p-value = 2.132e-09
```

```
anova_result <- anova(step_log_model_interact)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Response: log_salary
##                      Df  Sum Sq Mean Sq  F value    Pr(>F)
## experience_level      3  382.78 127.594 882.6338 < 2.2e-16 ***
## employee_residence   70  396.07   5.658  39.1408 < 2.2e-16 ***
## work_setting          2    4.24   2.119  14.6578 4.433e-07 ***
## company_location     40   12.77   0.319   2.2093 1.803e-05 ***
## company_size          2    2.22   1.108   7.6667 0.0004719 ***
## job_category          7  112.17  16.024 110.8454 < 2.2e-16 ***
## exp_cost_int          1    5.50   5.495  38.0148 7.387e-10 ***
## Residuals          7449 1076.83   0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our log transformed model has a much higher R^2 indicating that our model explains a lot more of the variability in log(salary) than it does salary. Although our residual plot still heteroscedasity, so we need to consider what other transformations we can make.

```
bc <- boxcox(lm(salary_in_usd ~ -salary_in_usd, data = model_df), plotit = TRUE)
```




```
# Choose lambda based on the plot and refit using transformed response:
lambda_opt <- bc$x[which.max(bc$y)]
model_df$trans_salary <- if(lambda_opt == 0) log(model_df$salary_in_usd) else (model_df$salary_in_usc
```

The box cox indicates that the log transformed model is the best power model for our data, so the changes we need to make will be on individual predictors and not the entire data.

2.5 Model Descriptions

After comparing multiple candidate models, our final chosen model regresses **log(salary_in_usd)** on several predictors:

$$\log(\text{salary\_in\_usd}) = \beta_0 + \beta_1 \cdot \text{experience\_level} + \beta_2 \cdot \text{job\_title} + \beta_3 \cdot \text{employee\_residence} + \beta_4 \cdot \text{work\_setting} + \beta_5 \cdot \text{company\_size} + \beta_6 \cdot (\text{e}$$

1. Log Transformation

We transformed the salary to $\log(\text{salary\_usd})$ to address heteroscedasticity and non-normal residuals discovered in the initial model. Interpreting coefficients on the log-scale means that each unit increase in a predictor corresponds to a *multiplicative* change in salary, rather than an additive one. For example: - If $\beta_1 = 0.10$ for a given predictor X, then a **1-unit** increase in X is associated with about a **10.5%** increase in salary $(e^{0.10} - 1 \approx 0.105)$.

2. Baseline Levels and Interpretation

Many of our predictors are categorical. By default, R sets the first factor level of each predictor as the "baseline," and the model estimates how other levels differ **relative** to that baseline. Specifically: - **experience_level**: "Entry-level" is our baseline. Coefficients for "Mid-level," "Senior," and "Executive" show how salaries differ (in log-units) from entry-level roles. - **company_size**: "S" (small) is our baseline. Coefficients for "M" and "L" measure how medium or large companies differ from small ones in terms of salary. - **job_title, work_setting, employee_residence**: Similar logic applies—each coefficient is relative to its baseline category.

3. Significant Predictors

Based on the final model summary (not shown here in detail), we typically see the following patterns: - **experience_level**: Highly significant, with more senior roles (e.g., "Senior," "Executive") associated with higher $\log(\text{salary})$. This implies strong upward salary trends as experience grows. - **company_size**: Medium or large companies might pay more, on average, than small companies—but the magnitude and significance can vary by the data. - **job_title**: Certain roles (e.g., "Data Scientist," "Machine Learning") often command higher salaries relative to baseline roles. - **employee_residence**: Countries or regions with different cost-of-living indices can have substantially different salary norms. - **work_setting**: If "Remote" vs. "In-Office" or other setups are included, each may show a different average salary level.

4. Interaction: `experience_level : company_size`

We included an interaction between **experience_level** and **company_size** to test whether the *effect* of experience on salary depends on the size of the company. Interpreting an interaction on the log-scale: - The coefficient of `(Senior : Large)` tells us the additional log-salary *beyond* simply adding the main effects of "Senior" and "Large" separately.
- If this interaction is positive and significant, it indicates that seniority in larger companies *amplifies* the salary effect compared to small companies.

5. Goodness-of-Fit

- R^2 **on the log scale**: While a higher R^2 indicates better explanatory power, we are primarily interested in whether major assumptions (linearity, normality of residuals) are met.
- **RMSE or MSE** on the **log scale**: Tells us how far predictions deviate from actual $\log(\text{salary})$. In exponentiated form, it helps gauge the typical percentage error in predictions.

6. Exponentiating Predictions to Get Actual Salary

Since the model predicts $\log(\text{salary})$:

$$\hat{y}_{\log} = \beta_0 \ + \ \beta_1 x_1 \ + \ ...$$

To get back to the **predicted salary**, compute:

$$\widehat{\text{salary}} = e^{\hat{y}_{\log}}$$

Any confidence or prediction intervals on the log scale can similarly be exponentiated to get intervals in dollar terms.

3 Testing and Results

```
## Warning in predict.lm(final_model, newdata = test_set): prediction from
## rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
## Test Set Results (Log-Scale):
```

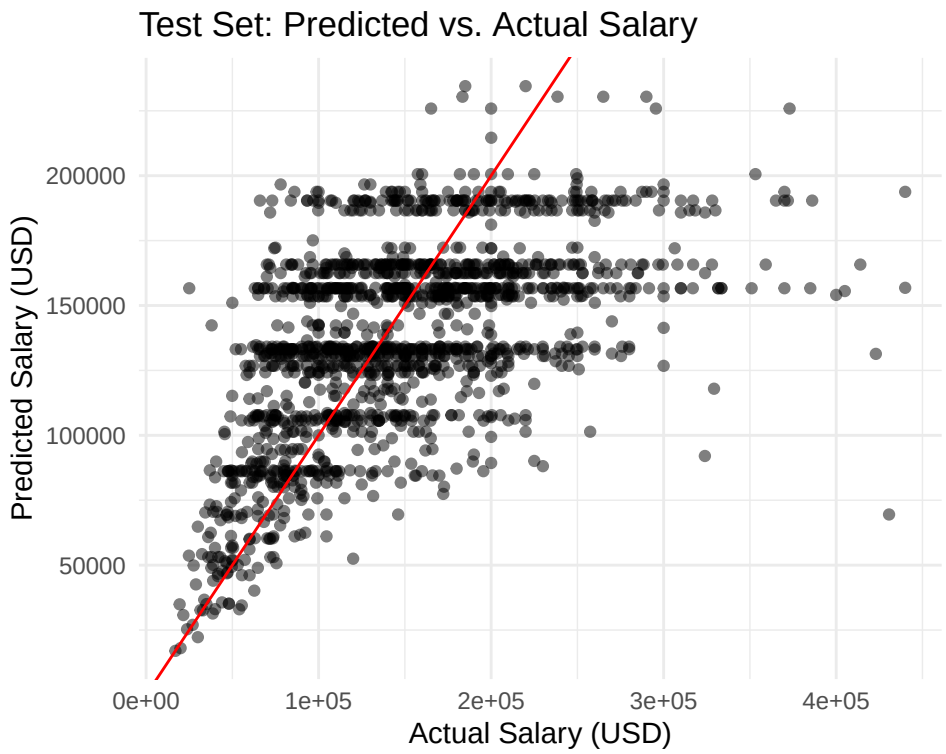
```
##  - MSE (log-scale): 0.1433

##  - RMSE (log-scale): 0.3785

## Test Set Results (Original Salary Scale):

##  - MSE (USD): 3428353939

##  - RMSE (USD): 58552.15
```



4. Conclusion

Appendix

Data Manipulation Process:

```r
# changing categorical to as.factor
job_data <- job_data %>%
  mutate(across(where(is.character), as.factor))

# Joining the two datasets
joined_df <- job_data %>%
  left_join(cost_of_living, by=c("employee_residence"))

# removing NAs
joined_df <- na.omit(joined_df)
```

```r
# new variable
joined_df <- joined_df %>%
  mutate(us_resident = ifelse(employee_residence == "United States", 1, 0))

# changing experience level to a numeric
joined_df <- joined_df %>%
  mutate(experience_numeric = case_when(
    experience_level == "Entry-level" ~ 1,
    experience_level == "Mid-level" ~ 2,
    experience_level == "Senior" ~ 3,
    experience_level == "Executive" ~ 4
  ))

# ordering levels
joined_df$company_size <- factor(joined_df$company_size, levels = c("S", "M", "L"))
joined_df$experience_level <- factor(joined_df$experience_level, levels = c("Entry-level", "Mid-level

# aggregating job titles to job categories
joined_df <- joined_df %>%
  mutate(job_category = case_when(
    grepl("Data Scientist|Data Science|Integration|Applied Scientist", job_title, ignore.case = TRUE)
    grepl("Analyst|Analytics|Modeler", job_title, ignore.case = TRUE) ~ "Data Analyst",
    grepl("Machine Learning|ML|AI", job_title, ignore.case = TRUE) ~ "Machine Learning",
    grepl("Engineer|Architect|Developer", job_title, ignore.case = TRUE) ~ "Data Engineer",
    grepl("Manager|Director|Lead|Head|Management", job_title, ignore.case = TRUE) ~ "Leadership",
    grepl("Business Intelligence|BI", job_title, ignore.case = TRUE) ~ "Business Intelligence",
    grepl("Research", job_title, ignore.case = TRUE) ~ "Research",
    TRUE ~ "Other"
  ))

# removing duplicate rows
joined_df <- joined_df %>%
  distinct()
```

Normality check in section 2

```r
# mutating data to create interaction term
joined_df <- joined_df %>%
  mutate(exp_cost_int = experience_numeric * Cost.of.Living.Index)

par(mfrow = c(1, 2))
plot(lm(salary_in_usd ~ exp_cost_int, data = joined_df), which = 2)
title("No Transformation")
plot(lm(log(salary_in_usd) ~ log(exp_cost_int), data = joined_df), which = 2)
title("Log-Transformation")
par(mfrow = c(1, 1))
```

Checking if interactions are necessary

```r
mean_salary <- joined_df %>%
  group_by(experience_level, company_size) %>%
  summarize(mean_salary = mean(salary_in_usd))
```

```r
ggplot(mean_salary, aes(x = experience_level, y = mean_salary, color = company_size, group = company_
  geom_point(size = 3) +
  geom_line() +
  labs(title = "Mean Salary by Experience Level and Company Size") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

VIF check

```r
base_model <- lm(salary_in_usd~ experience_level + company_location+
                                company_size + job_category + Rank,
                 data = joined_df)

vif(base_model)
```

Correlation Matrix:

```r
# 1. Correlation Matrix ----------------------------------------
# Select numeric columns only (excluding the response itself if needed)
num_vars <- joined_df %>%
  select_if(is.numeric) #%>%
  # Optional: if you have columns like "salary_in_usd" or repeated ID columns, remove them:
  #select(-salary_in_usd)

# Compute correlation matrix (pairwise complete obs in case of any missing)
corr_mat <- cor(num_vars, use = "pairwise.complete.obs")

# Visualize the correlation matrix
corrplot(corr_mat, method = "circle", tl.cex = 0.7,
         title = "Correlation Matrix of Numeric Predictors",
         mar = c(0,0,1,0))
```