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.

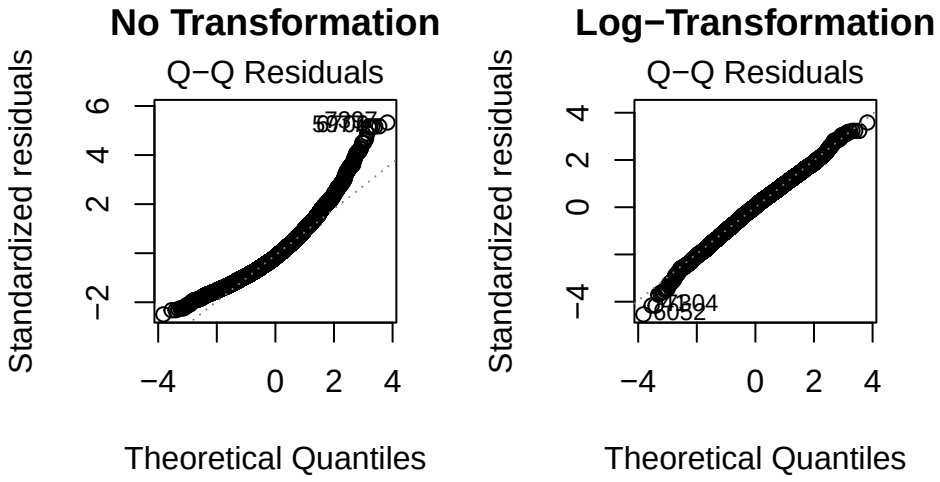


Figure 1: Normality check for Experience and Cost of Living Index Interaction variable on Salary, comparison between no transformation and log-transformation

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.

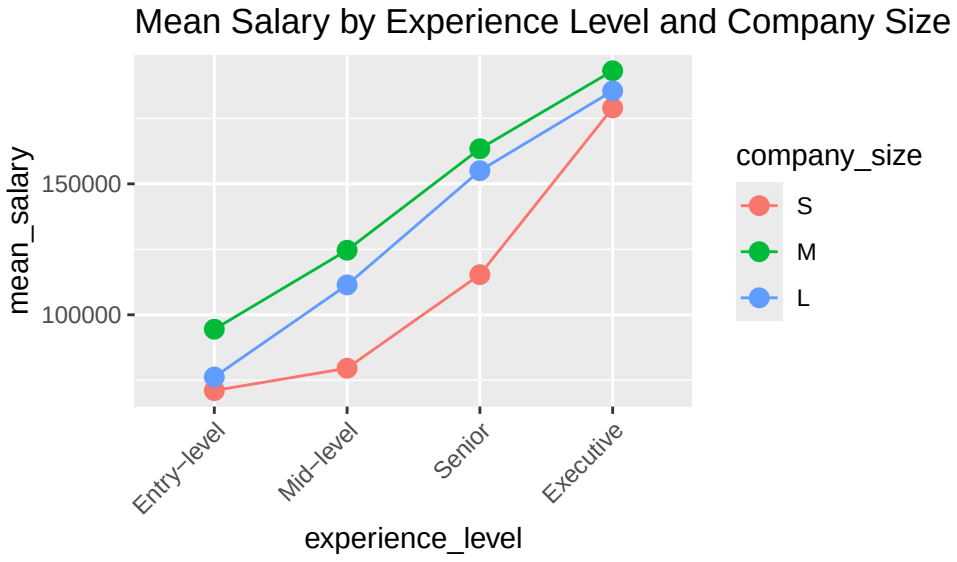


Figure 2: Interaction between Experience Level and Company Size on Salary

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:

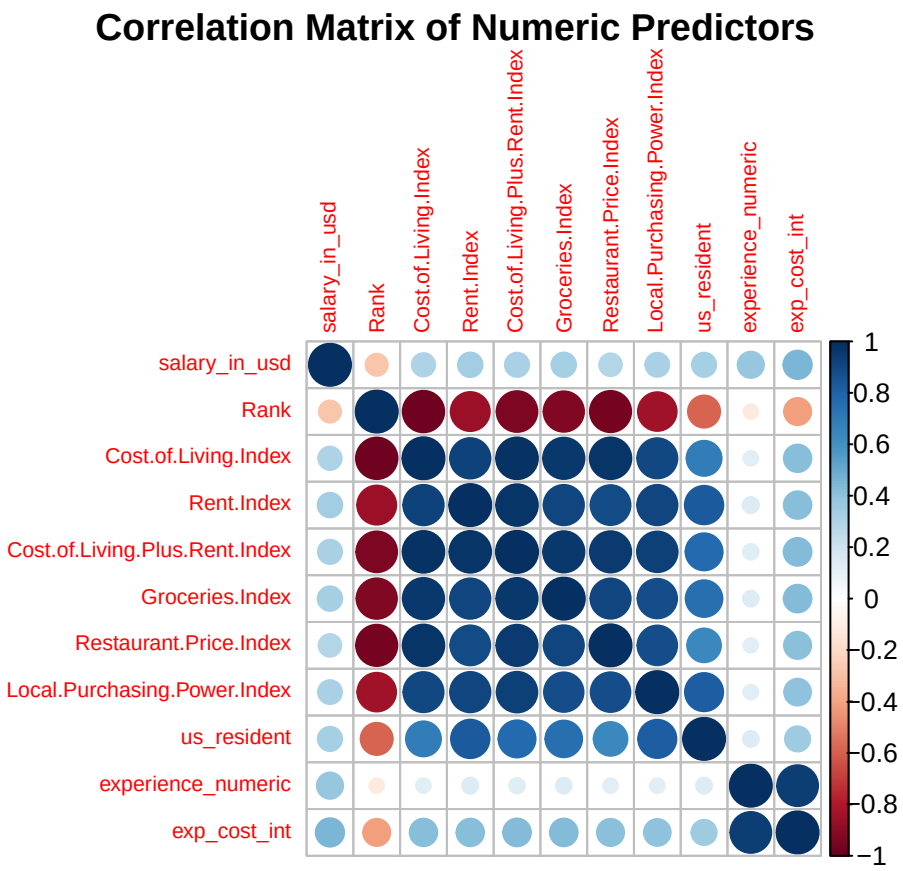
```
base_model <- lm(salary_in_usd~ experience_level + company_location+
                                company_size + job_category + Rank,
                 data = model_df)

vif(base_model)
```

```
##                       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
```

```
## 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.

2.4 Correlation and Multicollinearity Analysis



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.

```
## [1] 8
```

```
## [1] 8
```

```
## [1] 8
```

```
## 8
## 8
```

The AIC, BIC, CP, and Adjusted R^2, all tells me that `experience numeric`, `us_resident`, `Rank`, `company_size`, and `job_category`

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.

```
##
## Attaching package: 'MASS'
```

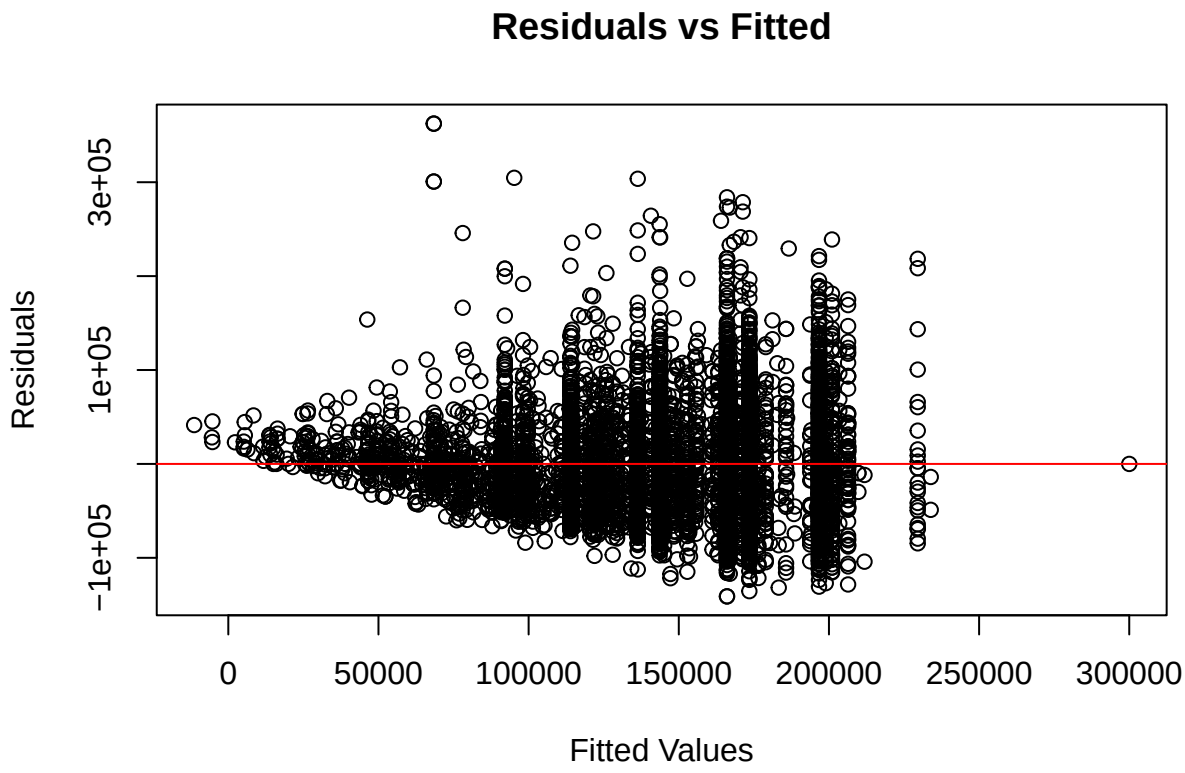
```
## The following object is masked from 'package:dplyr':
##
##     select


##
## Call:
## lm(formula = salary_in_usd ~ experience_level + employee_residence +
##     company_size + job_category, data = model_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141057  -38265   -7070   29943  362544
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             67954      57054   1.191 0.233672
## experience_levelMid-level               21853       2499   8.745  < 2e-16 ***
## experience_levelSenior                  51562       2345  21.991  < 2e-16 ***
## experience_levelExecutive               84443       3894  21.684  < 2e-16 ***
## employee_residenceArgentina            -61751      58738  -1.051 0.293157
## employee_residenceArmenia              -73672      68952  -1.068 0.285346
## employee_residenceAustralia             11326      56869   0.199 0.842146
## employee_residenceAustria              -50437      58985  -0.855 0.392536
## employee_residenceBelgium              -29800      60745  -0.491 0.623741
## employee_residenceBrazil               -67204      57564  -1.167 0.243062
## employee_residenceBulgaria             -21932      79695  -0.275 0.783170
## employee_residenceCanada                 4205      56322   0.075 0.940485
## employee_residenceChile                 51998      79685   0.653 0.514069
## employee_residenceChina                 75034      79640   0.942 0.346140
## employee_residenceColombia             -60625      58215  -1.041 0.297726
## employee_residenceCosta Rica           -50132      79476  -0.631 0.528206
## employee_residenceCroatia              -51785      61636  -0.840 0.400837
## employee_residenceCyprus               -84938      79540  -1.068 0.285614
## employee_residenceDenmark              -57426      64995  -0.884 0.376966
## employee_residenceDominican Republic    10000      79453   0.126 0.899846
## employee_residenceEcuador              -81077      79498  -1.020 0.307831
## employee_residenceEgypt                -10882      58529  -0.186 0.852510
## employee_residenceEstonia              -66530      59284  -1.122 0.261802
## employee_residenceFinland              -85389      64981  -1.314 0.188863
## employee_residenceFrance               -43632      56681  -0.770 0.441461
## employee_residenceGeorgia              -61770      79576  -0.776 0.437630
## employee_residenceGermany              -27436      56554  -0.485 0.627602
## employee_residenceGhana                -66725      62943  -1.060 0.289138
## employee_residenceGreece               -59783      58371  -1.024 0.305783
## employee_residenceHungary              -24208      79525  -0.304 0.760824
## employee_residenceIndia                -41070      57496  -0.714 0.475058
## employee_residenceIndonesia            -52208      79525  -0.656 0.511523
## employee_residenceIraq                  -3430      79639  -0.043 0.965649
## employee_residenceIreland              -35030      58534  -0.598 0.549554
## employee_residenceItaly                -73870      57569  -1.283 0.199478
## employee_residenceJapan                  8212      62974   0.130 0.896252
## employee_residenceKenya                -11231      69021  -0.163 0.870747
## employee_residenceKuwait               -17208      79525  -0.216 0.828692
```

```
## employee_residenceLatvia                -77834      58205  -1.337 0.181185
## employee_residenceLebanon                 1618      68835   0.024 0.981241
## employee_residenceLithuania             -56430      57942  -0.974 0.330136
## employee_residenceLuxembourg             -8106      79525  -0.102 0.918813
## employee_residenceMalaysia               69097      79597   0.868 0.385374
## employee_residenceMalta                 -33493      64910  -0.516 0.605871
## employee_residenceMauritius              27206      79638   0.342 0.732651
## employee_residenceMexico                -37332      58528  -0.638 0.523588
## employee_residenceNetherlands           -43748      57235  -0.764 0.444677
## employee_residenceNew Zealand            16358      60766   0.269 0.787784
## employee_residenceNigeria               -55663      58557  -0.951 0.341848
## employee_residenceOman                 -127906      79513  -1.609 0.107743
## employee_residencePakistan              -61596      60708  -1.015 0.310313
## employee_residencePeru                  -67648      79683  -0.849 0.395928
## employee_residencePhilippines           -56191      59673  -0.942 0.346402
## employee_residencePoland                -45740      58105  -0.787 0.431188
## employee_residencePortugal              -75594      57164  -1.322 0.186068
## employee_residencePuerto Rico            13580      61619   0.220 0.825578
## employee_residenceQatar                 154955      79519   1.949 0.051375 .
## employee_residenceRomania               -80983      61617  -1.314 0.188790
## employee_residenceSaudi Arabia            9325      64966   0.144 0.885866
## employee_residenceSerbia                -76560      79661  -0.961 0.336547
## employee_residenceSingapore             -22196      61618  -0.360 0.718689
## employee_residenceSlovenia              -57629      60765  -0.948 0.342962
## employee_residenceSouth Africa          -63237      58532  -1.080 0.280009
## employee_residenceSpain                 -64989      56613  -1.148 0.251029
## employee_residenceSweden                  7007      68875   0.102 0.918969
## employee_residenceSwitzerland            -3814      60759  -0.063 0.949947
## employee_residenceThailand             -108549      68906  -1.575 0.115225
## employee_residenceTunisia               -43884      68851  -0.637 0.523904
## employee_residenceUganda                -87192      79481  -1.097 0.272670
## employee_residenceUkraine               -50699      58730  -0.863 0.388026
## employee_residenceUnited Arab Emirates  -52686      62953  -0.837 0.402665
## employee_residenceUnited Kingdom        -23562      56293  -0.419 0.675550
## employee_residenceUnited States          21939      56239   0.390 0.696476
## employee_residenceUzbekistan            -34243      64918  -0.527 0.597881
## company_sizeM                            19762       5080   3.890 0.000101 ***
## company_sizeL                            16839       5474   3.076 0.002104 **
## job_categoryData Analyst                -17584       8500  -2.069 0.038598 *
## job_categoryData Engineer                 4841       8479   0.571 0.568088
## job_categoryData Scientist               12284       8488   1.447 0.147883
## job_categoryLeadership                   -8375       9269  -0.904 0.366255
## job_categoryMachine Learning             35476       8552   4.148 3.39e-05 ***
## job_categoryOther                       -31790       9866  -3.222 0.001278 **
## job_categoryResearch                     39801       9138   4.356 1.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56180 on 7492 degrees of freedom
## Multiple R-squared:  0.3293, Adjusted R-squared:  0.3219
## F-statistic: 44.85 on 82 and 7492 DF,  p-value: < 2.2e-16
```

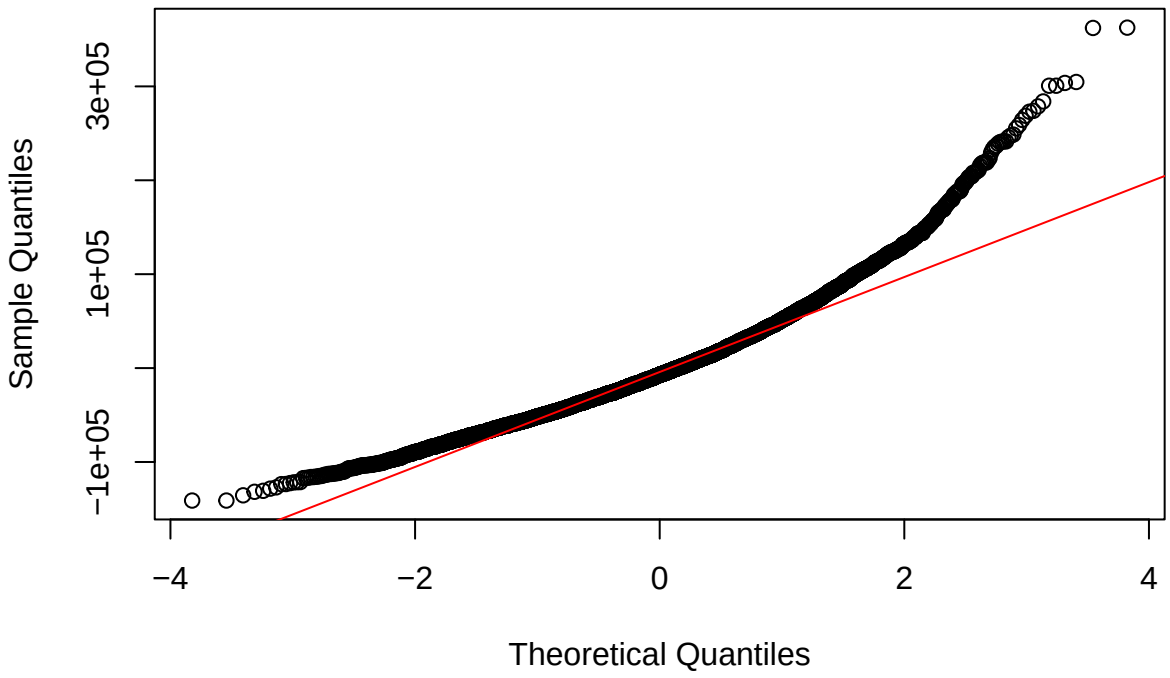
Now that we have the model that we have found from AIC we can test the assumptions of the model as well as performing

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 4.4.2
```

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

```
## Warning: package 'zoo' was built under R version 4.4.2
```

```
##
## 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 = 188.59, df = 82, p-value = 2.237e-10
```

```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 4.4.3

## 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
```

```r
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
```

```r
# 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+12 565.8493 < 2e-16 ***
## employee_residence    70 3.7700e+12 5.3857e+10  17.0627 < 2e-16 ***
## company_size           2 2.2718e+10 1.1359e+10   3.5987 0.02741 *
## job_category           7 2.4583e+12 3.5119e+11 111.2635 < 2e-16 ***
## Residuals           7492 2.3648e+13 3.1564e+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.

```r
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 +
##     company_location + company_size + job_category + exp_cost_int,
##     data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82942 -0.25093  0.00587  0.25599  1.83497
##
## Coefficients: (22 not defined because of singularities)
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         11.4124073  0.3875078  29.451  < 2e-16
## experience_levelMid-level            0.6219937  0.0614264  10.126  < 2e-16
## experience_levelSenior               1.2273860  0.1200227  10.226  < 2e-16
## experience_levelExecutive            1.8104911  0.1815259   9.974  < 2e-16
## employee_residenceArgentina         -0.7565578  0.4410650  -1.715 0.086332
## employee_residenceArmenia           -0.8989460  0.7197093  -1.249 0.211690
## employee_residenceAustralia          0.0246922  0.4352666   0.057 0.954763
## employee_residenceAustria           -0.2419745  0.4904162  -0.493 0.621740
## employee_residenceBelgium           -0.0107982  0.4675839  -0.023 0.981576
## employee_residenceBrazil            -0.5156816  0.4713949  -1.094 0.274013
## employee_residenceBulgaria          -0.5437693  0.5417990  -1.004 0.315586
## employee_residenceCanada             0.2418462  0.4067804   0.595 0.552171
## employee_residenceChile              0.4892909  0.5401046   0.906 0.365008
## employee_residenceChina              0.6668938  0.5398236   1.235 0.216724
## employee_residenceColombia          -1.0414703  0.3981221  -2.616 0.008916
## employee_residenceCosta Rica        -1.2594037  0.5390246  -2.336 0.019494
## employee_residenceCroatia           -1.0172277  0.5554677  -1.831 0.067096
## employee_residenceCyprus            -0.8830622  0.5390568  -1.638 0.101430
## employee_residenceDenmark           -1.1175392  0.5637242  -1.982 0.047469
## employee_residenceDominican Republic  0.3882445  0.5591474   0.694 0.487484
## employee_residenceEcuador           -1.7741885  0.5387779  -3.293 0.000996
## employee_residenceEgypt             -0.6863536  0.5409884  -1.269 0.204587
## employee_residenceEstonia           -0.9854945  0.4017864  -2.453 0.014198
## employee_residenceFinland           -0.0093182  0.6186291  -0.015 0.987983
## employee_residenceFrance            -0.1865461  0.4094331  -0.456 0.648676
## employee_residenceGeorgia           -0.1925110  0.7150378  -0.269 0.787759
## employee_residenceGermany           -0.1569256  0.4056277  -0.387 0.698863
## employee_residenceGhana             -0.7985417  0.5397535  -1.479 0.139061
## employee_residenceGreece            -0.7543091  0.4982723  -1.514 0.130106
## employee_residenceHungary           -0.8697469  0.7754183  -1.122 0.262048
## employee_residenceIndia             -0.5842793  0.4061974  -1.438 0.150359
## employee_residenceIndonesia         -1.6528843  0.5389598  -3.067 0.002171
## employee_residenceIraq               0.0698442  0.5397996   0.129 0.897053
## employee_residenceIreland           -0.1906015  0.3975945  -0.479 0.631678
## employee_residenceItaly             -0.8282781  0.4166315  -1.988 0.046844
## employee_residenceJapan             -0.0583999  0.5833882  -0.100 0.920264
## employee_residenceKenya             -0.3618596  0.4680617  -0.773 0.439486
## employee_residenceKuwait            -0.3626129  0.5390777  -0.673 0.501188
## employee_residenceLatvia            -1.0164893  0.3944931  -2.577 0.009994
## employee_residenceLebanon           -0.0581369  0.4666472  -0.125 0.900856
## employee_residenceLithuania         -0.7174892  0.3926953  -1.827 0.067726
```

```
## employee_residenceLuxembourg            0.4717622  0.6210577   0.760 0.447511
## employee_residenceMalaysia              -0.0419990  0.5432257  -0.077 0.938376
## employee_residenceMalta                 -0.5699079  0.4402290  -1.295 0.195508
## employee_residenceMauritius              0.3078464  0.5398584   0.570 0.568535
## employee_residenceMexico                -0.1829170  0.5936400  -0.308 0.757994
## employee_residenceNetherlands           -0.4705887  0.4530756  -1.039 0.298999
## employee_residenceNew Zealand            0.2410575  0.4127255   0.584 0.559196
## employee_residenceNigeria               -2.1565367  0.4525896  -4.765 1.93e-06
## employee_residenceOman                  -1.5731184  0.5398369  -2.914 0.003578
## employee_residencePakistan              -0.9651477  0.4367250  -2.210 0.027138
## employee_residencePeru                  -0.9923337  0.5905765  -1.680 0.092945
## employee_residencePhilippines           -0.9233932  0.4575139  -2.018 0.043597
## employee_residencePoland                -0.8373824  0.4474612  -1.871 0.061328
## employee_residencePortugal              -0.6008459  0.4433567  -1.355 0.175388
## employee_residencePuerto Rico            0.4067260  0.5406179   0.752 0.451873
## employee_residenceQatar                  0.7512944  0.5389438   1.394 0.163355
## employee_residenceRomania               -1.1312622  0.4542324  -2.490 0.012778
## employee_residenceSaudi Arabia          -0.0480643  0.4404302  -0.109 0.913102
## employee_residenceSerbia                -1.4428304  0.5566965  -2.592 0.009567
## employee_residenceSingapore             -0.0588884  0.4290096  -0.137 0.890824
## employee_residenceSlovenia              -0.8528111  0.4118136  -2.071 0.038406
## employee_residenceSouth Africa          -1.0836476  0.3980816  -2.722 0.006500
## employee_residenceSpain                 -0.3908190  0.4138915  -0.944 0.345070
## employee_residenceSweden                 0.4913481  0.6184141   0.795 0.426913
## employee_residenceSwitzerland            0.4517858  0.4191857   1.078 0.281171
## employee_residenceThailand              -1.4275581  0.5420204  -2.634 0.008462
## employee_residenceTunisia               -0.6348474  0.4992128  -1.272 0.203520
## employee_residenceUganda                -1.1588554  0.5386457  -2.151 0.031475
## employee_residenceUkraine               -1.3190670  0.5927127  -2.225 0.026079
## employee_residenceUnited Arab Emirates  -0.4135490  0.4267058  -0.969 0.332494
## employee_residenceUnited Kingdom        -0.2047027  0.4203019  -0.487 0.626246
## employee_residenceUnited States          0.2932845  0.3833142   0.765 0.444219
## employee_residenceUzbekistan            -0.3438729  0.4664674  -0.737 0.461033
## company_locationAmerican Samoa          -0.2223757  0.3031412  -0.734 0.463233
## company_locationArgentina               -0.3719258  0.2665054  -1.396 0.162887
## company_locationArmenia                 -0.0510300  0.7221536  -0.071 0.943667
## company_locationAustralia                0.1714596  0.2014230   0.851 0.394663
## company_locationAustria                 -0.1789617  0.3071879  -0.583 0.560193
## company_locationBelgium                 -0.3069112  0.3302022  -0.929 0.352679
## company_locationBrazil                  -0.7607095  0.2841482  -2.677 0.007441
## company_locationCanada                  -0.1511078  0.1382693  -1.093 0.274494
## company_locationColombia                        NA         NA      NA       NA
## company_locationCroatia                  0.4661034  0.4611166   1.011 0.312138
## company_locationCzechia                 -0.2634430  0.3530077  -0.746 0.455521
## company_locationDenmark                  0.2727401  0.3481303   0.783 0.433392
## company_locationEcuador                         NA         NA      NA       NA
## company_locationEgypt                   -0.0361132  0.3985508  -0.091 0.927804
## company_locationEstonia                         NA         NA      NA       NA
## company_locationFinland                 -0.6249108  0.4343665  -1.439 0.150286
## company_locationFrance                  -0.2425001  0.1502510  -1.614 0.106577
## company_locationGermany                 -0.0778961  0.1351875  -0.576 0.564492
## company_locationGhana                   -0.8305336  0.4403507  -1.886 0.059324
## company_locationGibraltar               -0.1702162  0.4139240  -0.411 0.680919
```

```
## company_locationGreece                -0.1706106  0.3410357  -0.500 0.616898
## company_locationHungary                0.3540696  0.5570968   0.636 0.525081
## company_locationIndia                 -0.6389206  0.1672971  -3.819 0.000135
## company_locationIndonesia                     NA         NA      NA       NA
## company_locationIraq                          NA         NA      NA       NA
## company_locationIreland                       NA         NA      NA       NA
## company_locationIsrael                 0.3090178  0.4260954   0.725 0.468333
## company_locationItaly                 -0.0979156  0.1954640  -0.501 0.616429
## company_locationJapan                  0.0645339  0.3812750   0.169 0.865598
## 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            -0.5533760  0.3053180  -1.812 0.069956
## company_locationMalaysia              -0.7238464  0.5846435  -1.238 0.215718
## company_locationMalta                         NA         NA      NA       NA
## company_locationMauritius                     NA         NA      NA       NA
## company_locationMexico                -0.6116891  0.4432747  -1.380 0.167649
## company_locationNetherlands            0.1348664  0.2474021   0.545 0.585680
## company_locationNew Zealand                   NA         NA      NA       NA
## company_locationNigeria                1.2303373  0.2702796   4.552 5.40e-06
## company_locationOman                          NA         NA      NA       NA
## company_locationPakistan              -0.5957646  0.3262511  -1.826 0.067876
## company_locationPhilippines           -0.2774219  0.3022801  -0.918 0.358772
## company_locationPoland                 0.0249429  0.2457885   0.101 0.919171
## company_locationPortugal              -0.5028196  0.2382653  -2.110 0.034863
## company_locationPuerto Rico           -0.2764516  0.4272682  -0.647 0.517638
## company_locationQatar                         NA         NA      NA       NA
## company_locationRomania               -0.2425027  0.2747941  -0.882 0.377541
## company_locationRussian Federation    -0.9408783  0.4800524  -1.960 0.050039
## 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                 -0.5658850  0.1642322  -3.446 0.000573
## company_locationSweden                -0.2395968  0.4046254  -0.592 0.553772
## company_locationSwitzerland                   NA         NA      NA       NA
## company_locationThailand              -0.5410881  0.5419112  -0.998 0.318079
## company_locationUkraine                0.2866084  0.4752001   0.603 0.546439
## company_locationUnited Arab Emirates          NA         NA      NA       NA
## company_locationUnited Kingdom        -0.0395771  0.1735627  -0.228 0.819631
## company_locationUnited States                 NA         NA      NA       NA
## company_sizeM                          0.2059276  0.0356946   5.769 8.29e-09
## company_sizeL                          0.1682687  0.0385220   4.368 1.27e-05
## job_categoryData Analyst              -0.1623645  0.0576058  -2.819 0.004837
## job_categoryData Engineer              0.0005731  0.0574659   0.010 0.992044
## job_categoryData Scientist             0.0565406  0.0575331   0.983 0.325762
## job_categoryLeadership                -0.0933065  0.0628196  -1.485 0.137504
## job_categoryMachine Learning           0.1948655  0.0579683   3.362 0.000779
## job_categoryOther                     -0.2891424  0.0668642  -4.324 1.55e-05
## job_categoryResearch                   0.2133163  0.0619331   3.444 0.000576
## exp_cost_int                          -0.0056038  0.0008798  -6.370 2.01e-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
## 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: 0.3807 on 7451 degrees of freedom
## Multiple R-squared:  0.4579, Adjusted R-squared:  0.449
## F-statistic: 51.17 on 123 and 7451 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 +
##     company_location + company_size + job_category + exp_cost_int,
##     data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82942 -0.25093  0.00587  0.25599  1.83497
##
## Coefficients: (22 not defined because of singularities)
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        11.4124073  0.3875078  29.451  < 2e-16
## experience_levelMid-level           0.6219937  0.0614264  10.126  < 2e-16
## experience_levelSenior              1.2273860  0.1200227  10.226  < 2e-16
## experience_levelExecutive           1.8104911  0.1815259   9.974  < 2e-16
## employee_residenceArgentina        -0.7565578  0.4410650  -1.715 0.086332
## employee_residenceArmenia          -0.8989460  0.7197093  -1.249 0.211690
## employee_residenceAustralia         0.0246922  0.4352666   0.057 0.954763
## employee_residenceAustria          -0.2419745  0.4904162  -0.493 0.621740
## employee_residenceBelgium          -0.0107982  0.4675839  -0.023 0.981576
## employee_residenceBrazil           -0.5156816  0.4713949  -1.094 0.274013
## employee_residenceBulgaria         -0.5437693  0.5417990  -1.004 0.315586
## employee_residenceCanada            0.2418462  0.4067804   0.595 0.552171
## employee_residenceChile             0.4892909  0.5401046   0.906 0.365008
## employee_residenceChina             0.6668938  0.5398236   1.235 0.216724
## employee_residenceColombia         -1.0414703  0.3981221  -2.616 0.008916
## employee_residenceCosta Rica       -1.2594037  0.5390246  -2.336 0.019494
## employee_residenceCroatia          -1.0172277  0.5554677  -1.831 0.067096
## employee_residenceCyprus           -0.8830622  0.5390568  -1.638 0.101430
## employee_residenceDenmark          -1.1175392  0.5637242  -1.982 0.047469
## employee_residenceDominican Republic  0.3882445  0.5591474   0.694 0.487484
## employee_residenceEcuador          -1.7741885  0.5387779  -3.293 0.000996
## employee_residenceEgypt            -0.6863536  0.5409884  -1.269 0.204587
## employee_residenceEstonia          -0.9854945  0.4017864  -2.453 0.014198
## employee_residenceFinland          -0.0093182  0.6186291  -0.015 0.987983
## employee_residenceFrance           -0.1865461  0.4094331  -0.456 0.648676
## employee_residenceGeorgia          -0.1925110  0.7150378  -0.269 0.787759
## employee_residenceGermany          -0.1569256  0.4056277  -0.387 0.698863
## employee_residenceGhana            -0.7985417  0.5397535  -1.479 0.139061
## employee_residenceGreece           -0.7543091  0.4982723  -1.514 0.130106
## employee_residenceHungary          -0.8697469  0.7754183  -1.122 0.262048
## employee_residenceIndia            -0.5842793  0.4061974  -1.438 0.150359
## employee_residenceIndonesia        -1.6528843  0.5389598  -3.067 0.002171
## employee_residenceIraq              0.0698442  0.5397996   0.129 0.897053
## employee_residenceIreland          -0.1906015  0.3975945  -0.479 0.631678
## employee_residenceItaly            -0.8282781  0.4166315  -1.988 0.046844
## employee_residenceJapan            -0.0583999  0.5833882  -0.100 0.920264
```

```
## employee_residenceKenya                -0.3618596  0.4680617  -0.773 0.439486
## employee_residenceKuwait               -0.3626129  0.5390777  -0.673 0.501188
## employee_residenceLatvia               -1.0164893  0.3944931  -2.577 0.009994
## employee_residenceLebanon              -0.0581369  0.4666472  -0.125 0.900856
## employee_residenceLithuania            -0.7174892  0.3926953  -1.827 0.067726
## employee_residenceLuxembourg            0.4717622  0.6210577   0.760 0.447511
## employee_residenceMalaysia             -0.0419990  0.5432257  -0.077 0.938376
## employee_residenceMalta                -0.5699079  0.4402290  -1.295 0.195508
## employee_residenceMauritius             0.3078464  0.5398584   0.570 0.568535
## employee_residenceMexico               -0.1829170  0.5936400  -0.308 0.757994
## employee_residenceNetherlands          -0.4705887  0.4530756  -1.039 0.298999
## employee_residenceNew Zealand           0.2410575  0.4127255   0.584 0.559196
## employee_residenceNigeria              -2.1565367  0.4525896  -4.765 1.93e-06
## employee_residenceOman                 -1.5731184  0.5398369  -2.914 0.003578
## employee_residencePakistan             -0.9651477  0.4367250  -2.210 0.027138
## employee_residencePeru                 -0.9923337  0.5905765  -1.680 0.092945
## employee_residencePhilippines          -0.9233932  0.4575139  -2.018 0.043597
## employee_residencePoland               -0.8373824  0.4474612  -1.871 0.061328
## employee_residencePortugal             -0.6008459  0.4433567  -1.355 0.175388
## employee_residencePuerto Rico           0.4067260  0.5406179   0.752 0.451873
## employee_residenceQatar                 0.7512944  0.5389438   1.394 0.163355
## employee_residenceRomania              -1.1312622  0.4542324  -2.490 0.012778
## employee_residenceSaudi Arabia         -0.0480643  0.4404302  -0.109 0.913102
## employee_residenceSerbia               -1.4428304  0.5566965  -2.592 0.009567
## employee_residenceSingapore            -0.0588884  0.4290096  -0.137 0.890824
## employee_residenceSlovenia             -0.8528111  0.4118136  -2.071 0.038406
## employee_residenceSouth Africa         -1.0836476  0.3980816  -2.722 0.006500
## employee_residenceSpain                -0.3908190  0.4138915  -0.944 0.345070
## employee_residenceSweden                0.4913481  0.6184141   0.795 0.426913
## employee_residenceSwitzerland           0.4517858  0.4191857   1.078 0.281171
## employee_residenceThailand             -1.4275581  0.5420204  -2.634 0.008462
## employee_residenceTunisia              -0.6348474  0.4992128  -1.272 0.203520
## employee_residenceUganda               -1.1588554  0.5386457  -2.151 0.031475
## employee_residenceUkraine              -1.3190670  0.5927127  -2.225 0.026079
## employee_residenceUnited Arab Emirates -0.4135490  0.4267058  -0.969 0.332494
## employee_residenceUnited Kingdom       -0.2047027  0.4203019  -0.487 0.626246
## employee_residenceUnited States         0.2932845  0.3833142   0.765 0.444219
## employee_residenceUzbekistan           -0.3438729  0.4664674  -0.737 0.461033
## company_locationAmerican Samoa         -0.2223757  0.3031412  -0.734 0.463233
## company_locationArgentina              -0.3719258  0.2665054  -1.396 0.162887
## company_locationArmenia                -0.0510300  0.7221536  -0.071 0.943667
## company_locationAustralia               0.1714596  0.2014230   0.851 0.394663
## company_locationAustria                -0.1789617  0.3071879  -0.583 0.560193
## company_locationBelgium                -0.3069112  0.3302022  -0.929 0.352679
## company_locationBrazil                 -0.7607095  0.2841482  -2.677 0.007441
## company_locationCanada                 -0.1511078  0.1382693  -1.093 0.274494
## company_locationColombia                       NA         NA      NA       NA
## company_locationCroatia                 0.4661034  0.4611166   1.011 0.312138
## company_locationCzechia                -0.2634430  0.3530077  -0.746 0.455521
## company_locationDenmark                 0.2727401  0.3481303   0.783 0.433392
## company_locationEcuador                        NA         NA      NA       NA
## company_locationEgypt                  -0.0361132  0.3985508  -0.091 0.927804
## company_locationEstonia                        NA         NA      NA       NA
```

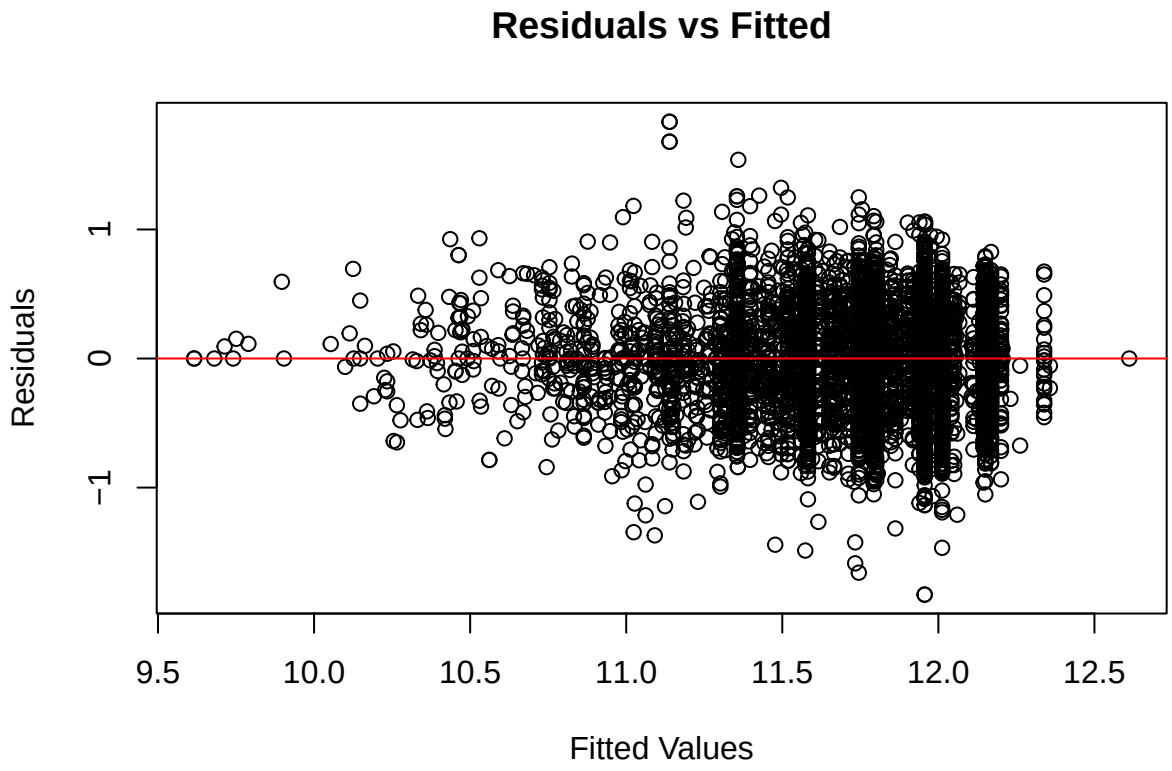
```
## company_locationFinland               -0.6249108  0.4343665  -1.439 0.150286
## company_locationFrance                -0.2425001  0.1502510  -1.614 0.106577
## company_locationGermany               -0.0778961  0.1351875  -0.576 0.564492
## company_locationGhana                 -0.8305336  0.4403507  -1.886 0.059324
## company_locationGibraltar             -0.1702162  0.4139240  -0.411 0.680919
## company_locationGreece                -0.1706106  0.3410357  -0.500 0.616898
## company_locationHungary                0.3540696  0.5570968   0.636 0.525081
## company_locationIndia                 -0.6389206  0.1672971  -3.819 0.000135
## company_locationIndonesia                     NA         NA      NA       NA
## company_locationIraq                          NA         NA      NA       NA
## company_locationIreland                       NA         NA      NA       NA
## company_locationIsrael                 0.3090178  0.4260954   0.725 0.468333
## company_locationItaly                 -0.0979156  0.1954640  -0.501 0.616429
## company_locationJapan                  0.0645339  0.3812750   0.169 0.865598
## 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            -0.5533760  0.3053180  -1.812 0.069956
## company_locationMalaysia              -0.7238464  0.5846435  -1.238 0.215718
## company_locationMalta                         NA         NA      NA       NA
## company_locationMauritius                     NA         NA      NA       NA
## company_locationMexico                -0.6116891  0.4432747  -1.380 0.167649
## company_locationNetherlands            0.1348664  0.2474021   0.545 0.585680
## company_locationNew Zealand                   NA         NA      NA       NA
## company_locationNigeria                1.2303373  0.2702796   4.552 5.40e-06
## company_locationOman                          NA         NA      NA       NA
## company_locationPakistan              -0.5957646  0.3262511  -1.826 0.067876
## company_locationPhilippines           -0.2774219  0.3022801  -0.918 0.358772
## company_locationPoland                 0.0249429  0.2457885   0.101 0.919171
## company_locationPortugal              -0.5028196  0.2382653  -2.110 0.034863
## company_locationPuerto Rico           -0.2764516  0.4272682  -0.647 0.517638
## company_locationQatar                         NA         NA      NA       NA
## company_locationRomania               -0.2425027  0.2747941  -0.882 0.377541
## company_locationRussian Federation    -0.9408783  0.4800524  -1.960 0.050039
## 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                 -0.5658850  0.1642322  -3.446 0.000573
## company_locationSweden                -0.2395968  0.4046254  -0.592 0.553772
## company_locationSwitzerland                   NA         NA      NA       NA
## company_locationThailand              -0.5410881  0.5419112  -0.998 0.318079
## company_locationUkraine                0.2866084  0.4752001   0.603 0.546439
## company_locationUnited Arab Emirates          NA         NA      NA       NA
## company_locationUnited Kingdom        -0.0395771  0.1735627  -0.228 0.819631
## company_locationUnited States                 NA         NA      NA       NA
## company_sizeM                          0.2059276  0.0356946   5.769 8.29e-09
## company_sizeL                          0.1682687  0.0385220   4.368 1.27e-05
## job_categoryData Analyst              -0.1623645  0.0576058  -2.819 0.004837
## job_categoryData Engineer              0.0005731  0.0574659   0.010 0.992044
## job_categoryData Scientist             0.0565406  0.0575331   0.983 0.325762
## job_categoryLeadership                -0.0933065  0.0628196  -1.485 0.137504
```

```
## job_categoryMachine Learning            0.1948655  0.0579683   3.362 0.000779
## job_categoryOther                       -0.2891424  0.0668642  -4.324 1.55e-05
## job_categoryResearch                     0.2133163  0.0619331   3.444 0.000576
## exp_cost_int                            -0.0056038  0.0008798  -6.370 2.01e-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
```

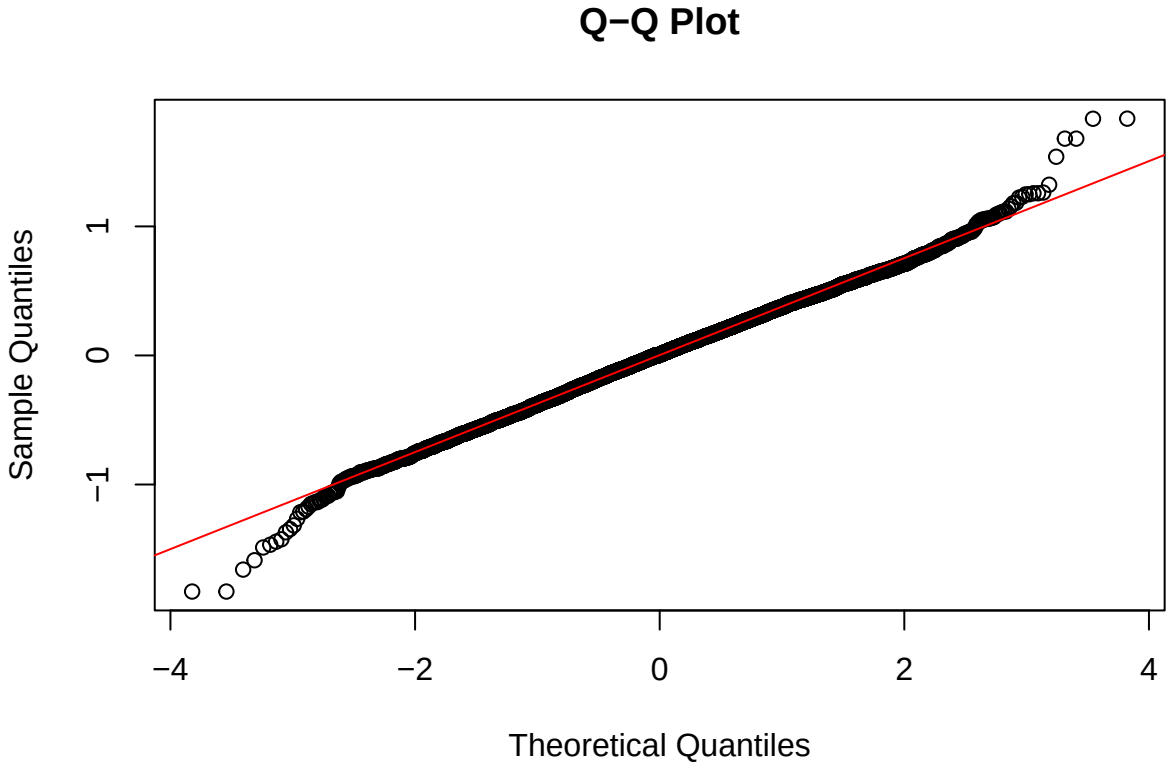
```
## employee_residenceNigeria              ***
## employee_residenceOman                 **
## employee_residencePakistan             *
## employee_residencePeru                 .
## employee_residencePhilippines          *
## employee_residencePoland               .
## employee_residencePortugal
## 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
## 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: 0.3807 on 7451 degrees of freedom
## Multiple R-squared:  0.4579, Adjusted R-squared:  0.449
## F-statistic: 51.17 on 123 and 7451 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
```

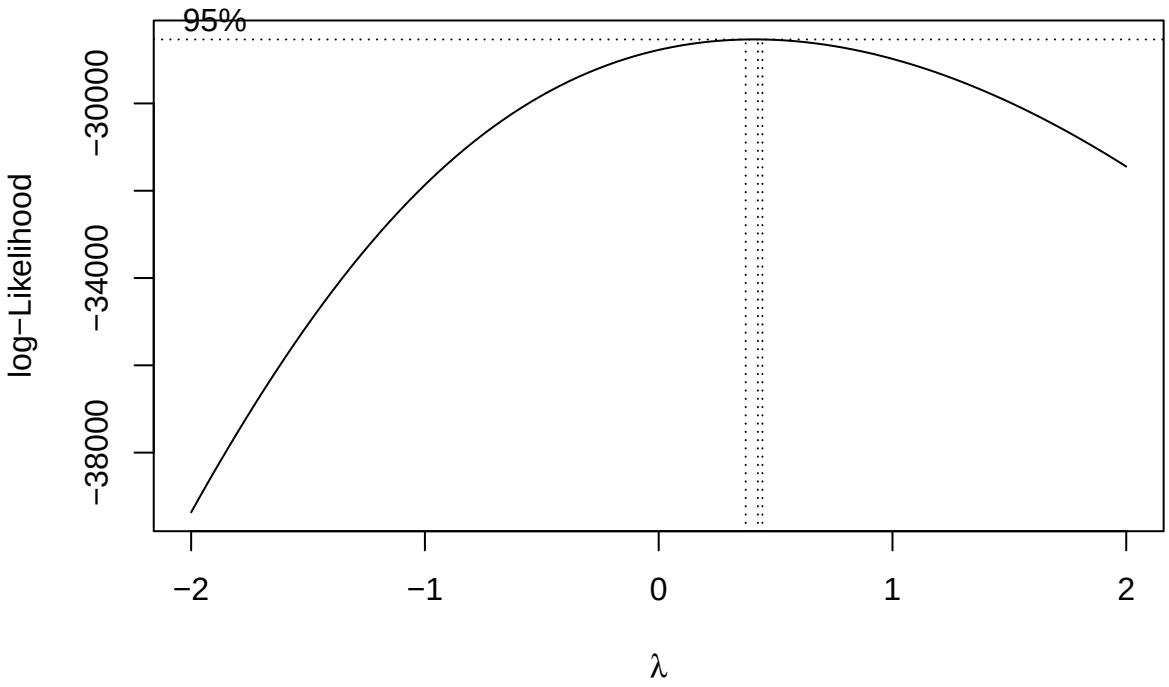
```
##
##  studentized Breusch-Pagan test
##
## data:  step_log_model_interact
## BP = 228.59, df = 123, p-value = 2.398e-08
```

```
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  382.78 127.594 880.1490 < 2.2e-16 ***
## employee_residence   70  396.07   5.658  39.0306 < 2.2e-16 ***
## company_location     40   13.06   0.326   2.2518 1.100e-05 ***
## company_size          2    2.64   1.321   9.1094 0.0001119 ***
## job_category          7  111.98  15.997 110.3507 < 2.2e-16 ***
## exp_cost_int          1    5.88   5.882  40.5727 2.007e-10 ***
## Residuals          7451 1080.16   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.

```
library(MASS)
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_usd
```

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.

3.3 Modeling

3.5 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