STEM Salary Prediction

Group 2: Neo, Geli, Mey

The goal of this project is to answer the question: ***What is the best model to predict the salaries of STEM employees?***

1. Literature Review

We found an article on Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy-Wide Activities and Occupations. The key takeaways are:

- Salary prediction models in literature are mostly concerned with the problem of unequal pay based on gender, race, or other biases that are not related to job content or job performance
- The performance of each regression model is given based on root-mean-square error (RMSE), R-squared (R^2), and mean absolute error (MAE).
- In this study, when cover the broader salary estimates, minor groups can also be featured in the prediction model to capture the occupational characteristics.

2. Data set Information

- We downloaded our data set from Kaggle, where a survey was conducted among Data Science and STEM professionals worldwide from 2017 to 2019. We initially started with about 60,000 rows of data, which was later reduced to 20,000 after removing missing data.
- Source: https://www.kaggle.com/jackogozaly/data-science-and-STEM2-salaries

```
# Read data set
STEM2 <- read.csv("Levels_Fyi_Salary_Data.csv")
attach(STEM2)
```

```
# Transform columns
STEM2$timestamp <- as.POSIXct(STEM2$timestamp, format = "%m/%d/%Y %H:%M:%S")

numerical_columns <- c("totalyearlycompensation", "yearsofexperience", "yearsatcompany",
                       "basesalary", "stockgrantvalue", "bonus", "cityid", "dmaid", "rowNumber")

categorical_columns <- c("company", "level", "title", "location", "tag", "gender",
                         "otherdetails", "Masters_Degree", "Bachelors_Degree",
                         "Doctorate_Degree", "Highschool", "Some_College", "Race_Asian",
                         "Race_White", "Race_Two_Or_More", "Race_Black", "Race_Hispanic", "Race",
                         "Education")

# Transform columns to numerical and categorical
STEM2 <- STEM2 %>%
  mutate(across(all_of(numerical_columns), as.integer)) %>%
  mutate(across(all_of(categorical_columns), as.factor))
```

The dataset consists of **12 variables**.

Half of these are **numerical**, including:

- **Total yearly compensation, base salary, stock grant value, and bonus**—all measured in US dollars. Base salary, stock grant value, and bonus are components of total yearly compensation.
- **Years of experience and years at company**—measured in years, where years at company is a part of years of experience.

For our analysis, we are considering either total yearly compensation or base salary as our potential outcome variables.

For our **categorical** variables, we have a total of 6, comprising gender, race, country, education, Fortune 500 status, and job title. All of these variables have been transformed to simplify the levels, which originally included more than 5 categories (for race and education) or even 15 (for job title).

- **Gender**: male or female.
- **Race**: non-White or White.
- **Country**: only the top 4 highest-response countries were retained to ensure adequate representation.
- **Education**: PhD, Masters, and Bachelor's/College/High School (grouped together).
- **Fortune 500**: a new variable created to determine whether a company ranks among the 2,000 high-revenue companies, which could potentially impact salary.
- **Job title**: Management versus non-managerial roles.

```
library(summarytools)
print(dfSummary(STEM2), method = 'render')
```

Missing Data

Drop columns/rows:

- otherdetails: 35.9% missing (not needed)
- tag - 1.3% missing (not needed)
- cityid - not needed
- dmaid - not needed
- rowNumber - not needed
- level - too many levels (not needed)

```
# List of columns to drop
columns_to_drop <- c("timestamp", "otherdetails", "cityid", "dmaid", "rowNumber", "level")

# Drop the specified columns
STEM2 <- STEM2 %>%
  dplyr::select(-all_of(columns_to_drop))
```

```
# List of columns to drop
columns_to_drop <- c("Masters_Degree", "Bachelors_Degree", "Doctorate_Degree", "Highschool",
                     "Some_College", "Race_Asian", "Race_White", "Race_Two_Or_More", "Race_Black", "Rac

# Drop the specified columns
STEM2 <- STEM2 %>%
  dplyr::select(-all_of(columns_to_drop))
```

```r
# Check missing data
colSums(is.na(STEM2))
```

```
##                  company                    title totalyearlycompensation
##                        0                        0                        0
##                 location       yearsofexperience          yearsatcompany
##                        0                        0                        0
##                      tag             basesalary         stockgrantvalue
##                      808                        0                        0
##                    bonus                   gender                     Race
##                        0                    19540                    40215
##                Education
##                    32272
```

```r
# Remove rows with NA
STEM2 <- STEM2 %>%
  filter(!is.na(Race) | !is.na(Education))
```

```r
# Remove rows with NA
STEM2 <- STEM2 %>%
  filter(!is.na(gender) | !is.na(Race) | !is.na(Education))
```

```r
# Remove rows with NA
STEM2 <- STEM2 %>%
  filter(!is.na(gender) & !is.na(Race) & !is.na(Education))
```

```r
# Check missing data
colSums(is.na(STEM2))
```

```
##                  company                    title totalyearlycompensation
##                        0                        0                        0
##                 location       yearsofexperience          yearsatcompany
##                        0                        0                        0
##                      tag             basesalary         stockgrantvalue
##                        8                        0                        0
##                    bonus                   gender                     Race
##                        0                        0                        0
##                Education
##                        0
```

Create New or Recode Columns

- **Country**: Create column based on 'location'
- **Level**: Simplify
- **Title**: Simplify
- **Company**: Simplify
- **Education**: Simplify

```r
# Country

# Ensure the 'location' column is character type
STEM2$location <- as.character(STEM2$location)

# Create a 'country' column by splitting the 'location' column and extracting the last part
STEM2$country <- sapply(strsplit(STEM2$location, ", "), function(x) trimws(x[length(x)]))

# Replace country codes (2-letter, e.g., "WA", "CA") with "US"
STEM2$country[nchar(STEM2$country) == 2] <- "US"

# Convert to factor
STEM2$country <- as.factor(STEM2$country)

# Replace 'United States' with 'US' in the specified column
STEM2$country[STEM2$country == "United States"] <- "US"
STEM2$country <- droplevels(STEM2$country)
```

```r
# Retain only the high-response countries
STEM2 <- STEM2[STEM2$country %in% c("United Kingdom", "Canada", "India", "US"), ]
```

Create 'Country' column

```r
library(stringr)

# Create new column 'Fortune_500'
fortune_500_list <- c("3M", "ABB", "Accenture", "Adidas", "Adobe", "ADP", "Aetna", "AIG", "Airbnb", "Ai
cleaned_fortune_500 <- str_trim(tolower(fortune_500_list))

STEM2 <- STEM2 %>%
  mutate(
    Fortune_500 = if_else(
      str_trim(tolower(company)) %in% cleaned_fortune_500,
      "Yes",
      "No"
    )
  )

STEM2$Fortune_500 <- as.factor(STEM2$Fortune_500)
```

```r
# Recode the 'title' column and store as 'title4'
STEM2 <- STEM2 %>%
  mutate(title4 = case_when(
    title == "Business Analyst" ~ "Non-Management",
    title == "Sales" ~ "Sales/HR/Marketing",
```

```
    title == "Mechanical Engineer" ~ "Non-Management",
    title == "Recruiter" ~ "Sales/HR/Marketing",
    title == "Human Resources" ~ "Sales/HR/Marketing",
    title == "Management Consultant" ~ "Non-Management",
    title == "Software Engineer" ~ "Non-Management",
    title == "Data Scientist" ~ "Non-Management",
    title == "Marketing" ~ "Sales/HR/Marketing",
    title == "Product Designer" ~ "Non-Management",
    title == "Hardware Engineer" ~ "Non-Management",
    title == "Solution Architect" ~ "Management",
    title == "Product Manager" ~ "Management",
    title == "Technical Program Manager" ~ "Management",
    title == "Software Engineering Manager" ~ "Management",
    TRUE ~ title # Retain the original value if it doesn't match any condition
  ) %>% as.factor())

# Remove rows where title2 is "Sales/HR/Marketing" (not enough data)
STEM2 <- STEM2 %>%
  filter(title4 != "Sales/HR/Marketing") %>% droplevels()
```

```
# Recode the 'Race' column and store as 'Race2'
STEM2 <- STEM2 %>%
  mutate(
    Race2 = ifelse(Race == "White", "White", "Non-White")
  %>% as.factor())
```

```
# Recode the 'Gender: Other'
STEM2 <- STEM2 %>%
  mutate(gender = as.character(gender)) %>%
  mutate(gender = ifelse(gender == "Other", "Female", gender))

STEM2 <- STEM2 %>%
  mutate(gender = as.factor(gender))
```

```
# Recode the 'Education: Highschool/Some College' column and store as 'educ3'
STEM2 <- STEM2 %>%
  mutate(educ3 = case_when(
    Education == "Bachelor's Degree" ~ "College or below",
    Education == "Highschool" ~ "College or below",
    Education == "Some College" ~ "College or below",
    Education == "Master's Degree" ~ "Master's Degree",
    Education == "PhD" ~ "PhD",
    TRUE ~ Education # Retain the original value if it doesn't match any condition
  ) %>% as.factor())
```

Simplify the categorical variable levels

Set the base levels

```
# Explicitly set factor levels with labels
STEM2$gender <- factor(STEM2$gender, levels = c("Male", "Female"))
STEM2$Race2 <- factor(STEM2$Race2, levels = c("Non-White", "White"))
STEM2$educ3 <- factor(STEM2$educ3, levels = c("College or below", "Master's Degree", "PhD"))
STEM2$Fortune_500 <- factor(STEM2$Fortune_500, levels = c("No", "Yes"))
STEM2$title4 <- factor(STEM2$title4, levels = c("Non-Management", "Management"))
STEM2$country <- factor(STEM2$country, levels = c("Canada", "India", "United Kingdom", "US"))
```

Check if all combinations exists: Parameter Estimatability (Chapter 7.3)

Some parameter combinations cannot be estimated due to missing data or sparse representation, limiting the model's ability to generalize for these cases.

```
# Generate all possible combinations
all_combinations <- expand.grid(
  gender  = unique(STEM2$gender),
  Race2  = unique(STEM2$Race2),
  educ3  = unique(STEM2$educ3),
  country  = unique(STEM2$country),
  title4  = unique(STEM2$title4),
  Fortune_500  = unique(STEM2$Fortune_500)
)

# Check for missing combinations
missing_combinations <- anti_join(all_combinations, STEM2, by = c("gender", "educ3", "country", "title4

missing_combinations
```

```
##    gender     Race2 educ3        country         title4 Fortune_500
## 1  Female Non-White   PhD         Canada Non-Management         Yes
## 2  Female     White   PhD         Canada Non-Management         Yes
## 3  Female Non-White   PhD          India Non-Management         Yes
## 4  Female     White   PhD          India Non-Management         Yes
## 5  Female Non-White   PhD         Canada     Management         Yes
## 6  Female     White   PhD         Canada     Management         Yes
## 7    Male Non-White   PhD          India     Management         Yes
## 8  Female Non-White   PhD          India     Management         Yes
## 9    Male     White   PhD          India     Management         Yes
## 10 Female     White   PhD          India     Management         Yes
## 11 Female Non-White   PhD United Kingdom     Management         Yes
## 12 Female     White   PhD United Kingdom     Management         Yes
## 13 Female Non-White   PhD          India Non-Management          No
## 14 Female     White   PhD          India Non-Management          No
## 15   Male Non-White   PhD         Canada     Management          No
## 16 Female Non-White   PhD         Canada     Management          No
## 17   Male     White   PhD         Canada     Management          No
## 18 Female     White   PhD         Canada     Management          No
## 19 Female Non-White   PhD          India     Management          No
## 20 Female     White   PhD          India     Management          No
## 21   Male Non-White   PhD United Kingdom     Management          No
```

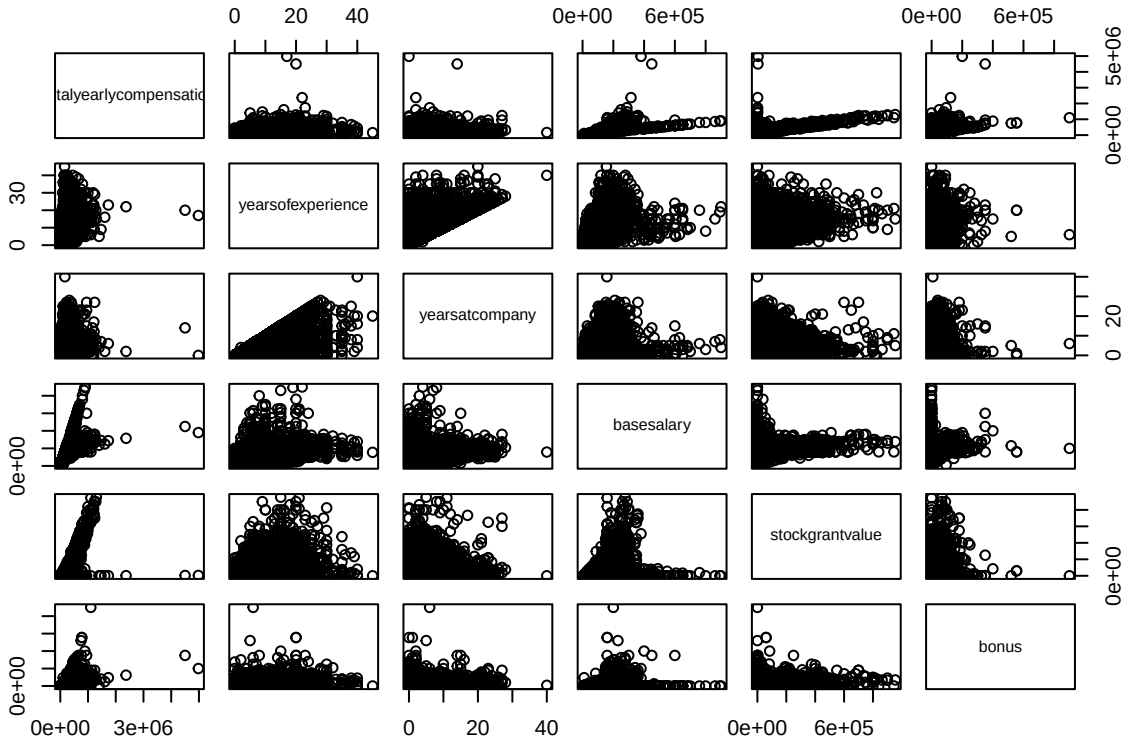
```
## 22 Female Non-White   PhD United Kingdom     Management          No
## 23   Male     White   PhD United Kingdom     Management          No
## 24 Female     White   PhD United Kingdom     Management          No
```

3. Variables Analysis

Scatterplot Matrix

The scatterplot matrix below was used to determine the pairwise correlations of each of our numerical variables. From this graph, we can clearly see a strong positive linear correlation, which suggests multicollinearity (as expected), since some of these variables are derived from each other. There are also some relationships that are not immediately clear, as they appear to be clustered together (such as total yearly compensation and years of experience), but we will explore these in greater detail in the next sections.

```
selected_columns <- STEM2[, c("totalyearlycompensation", "yearsofexperience", "yearsatcompany", "basesal
pairs(selected_columns)
```

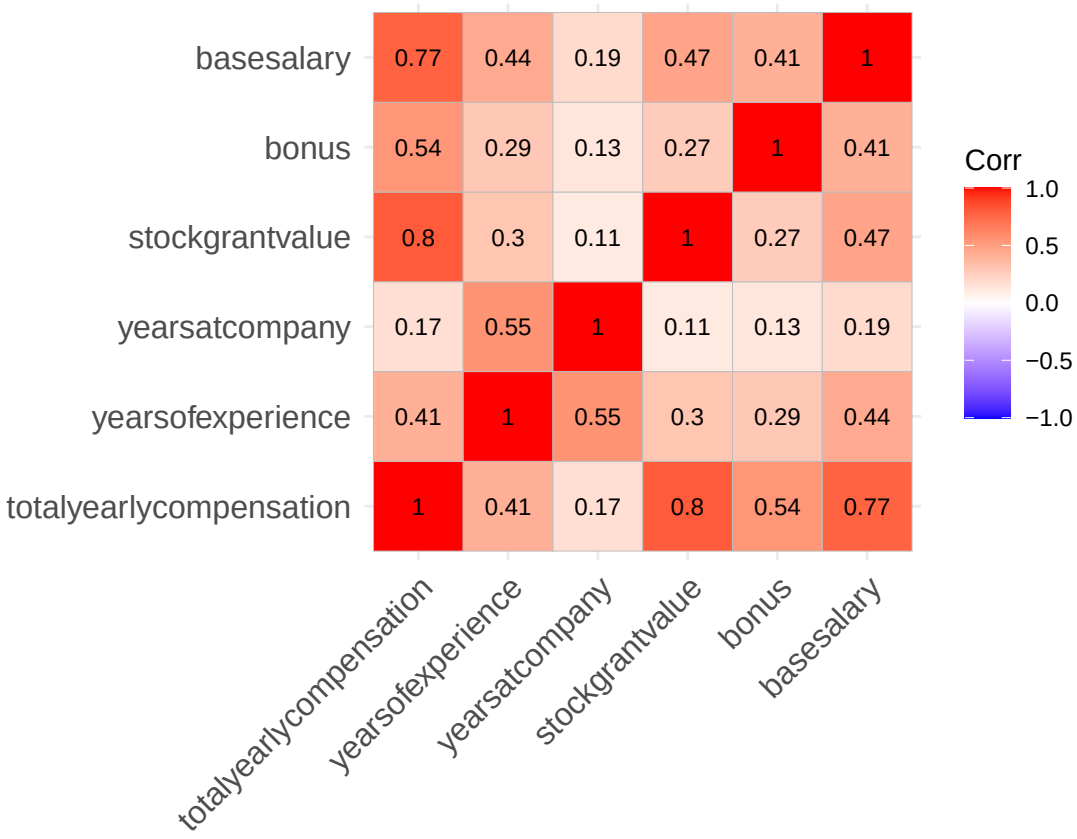


Correlation Matrix

The heat map below shows the strength of correlations between our numerical variables. This further highlights the multicollinearity among the predictors, particularly with base salary, bonus, and stock grant value, all having correlations greater than 50% with total yearly compensation.

Another notable correlation is between years of experience and years at company, which is at 55%. This makes sense, as years at the company is part of the total years of experience. Ultimately, we decided to retain both variables, as our goal is to estimate and predict salary.

```
library(ggcorrplot)
selected_columns <- STEM2[, c("totalyearlycompensation", "yearsofexperience", "yearsatcompany", "stockg
ggcorrplot(round(cor(selected_columns), 2), lab=TRUE,lab_size = 3)
```

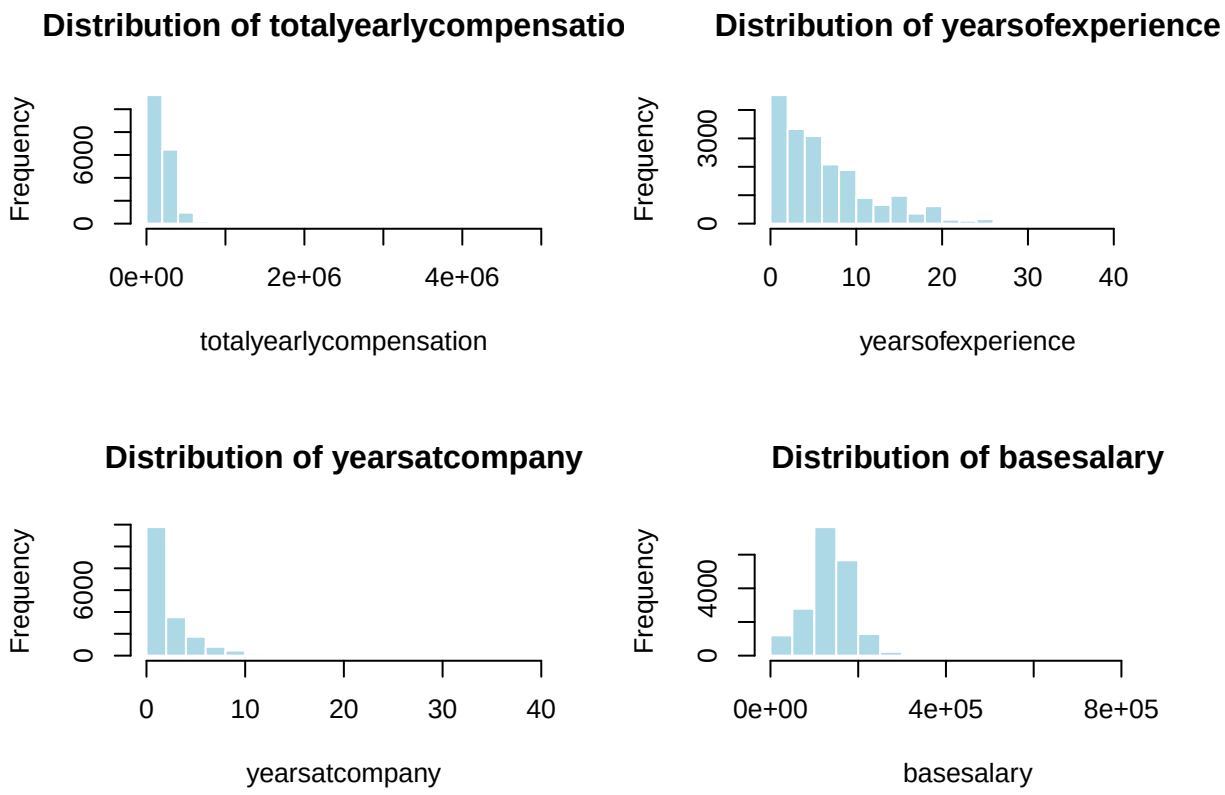


Numerical Variable Distribution

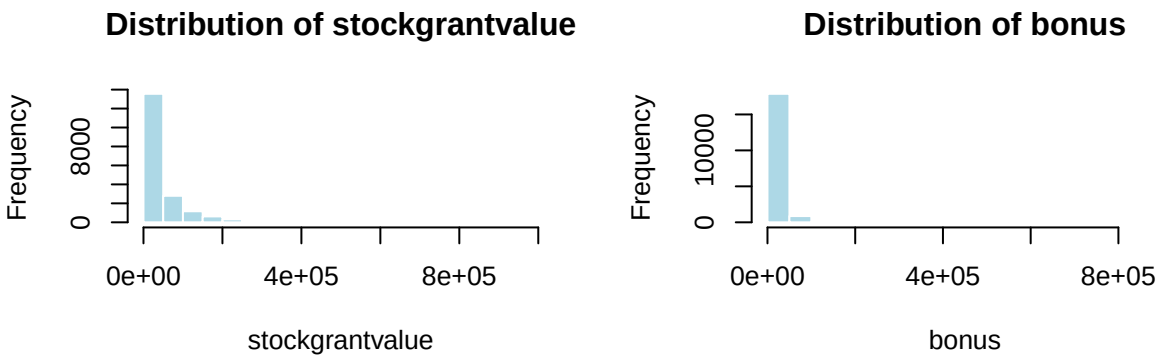
Next, we checked the distribution of each variable, starting with the numerical ones. We found that all of them are right-skewed with extreme outliers on the right, except for **base salary**, which has an almost bell-shaped curve with some outliers.

```
# Exclude specific columns
exclude_columns <- c("company", "Education")

# Identify numerical columns, excluding the specified ones
numerical_columns <- setdiff(names(STEM2)[sapply(STEM2, is.numeric)], exclude_columns)

# Set options to disable scientific notation
options(scipen = 10)  # Increase 'n' to make scientific notation less likely

# Set up the plotting area for a 2x2 grid
par(mfrow = c(2, 2))  # Set the plotting layout to 2 rows and 2 columns
```

```r
# Iterate through numerical columns and plot histograms
for (col in numerical_columns) {
  hist(
    STEM2[[col]],
    main = paste("Distribution of", col),
    xlab = col,
    col = "lightblue",
    border = "white",
    breaks = 20
  )
}
```




```r
# Reset the plotting parameters to default
par(mfrow = c(1, 1))  # Reset to a single plot layout
```




```
# Reset options to default after plotting
options(scipen = 0)
```

Categorical Variable Distribution

For categorical variables, we observed very similar median base salaries for gender, race, and Fortune 500. However, for country, title, and education, there is a noticeable gap in the median base salaries.

```
library(ggplot2)
library(dplyr)
library(gridExtra)
```

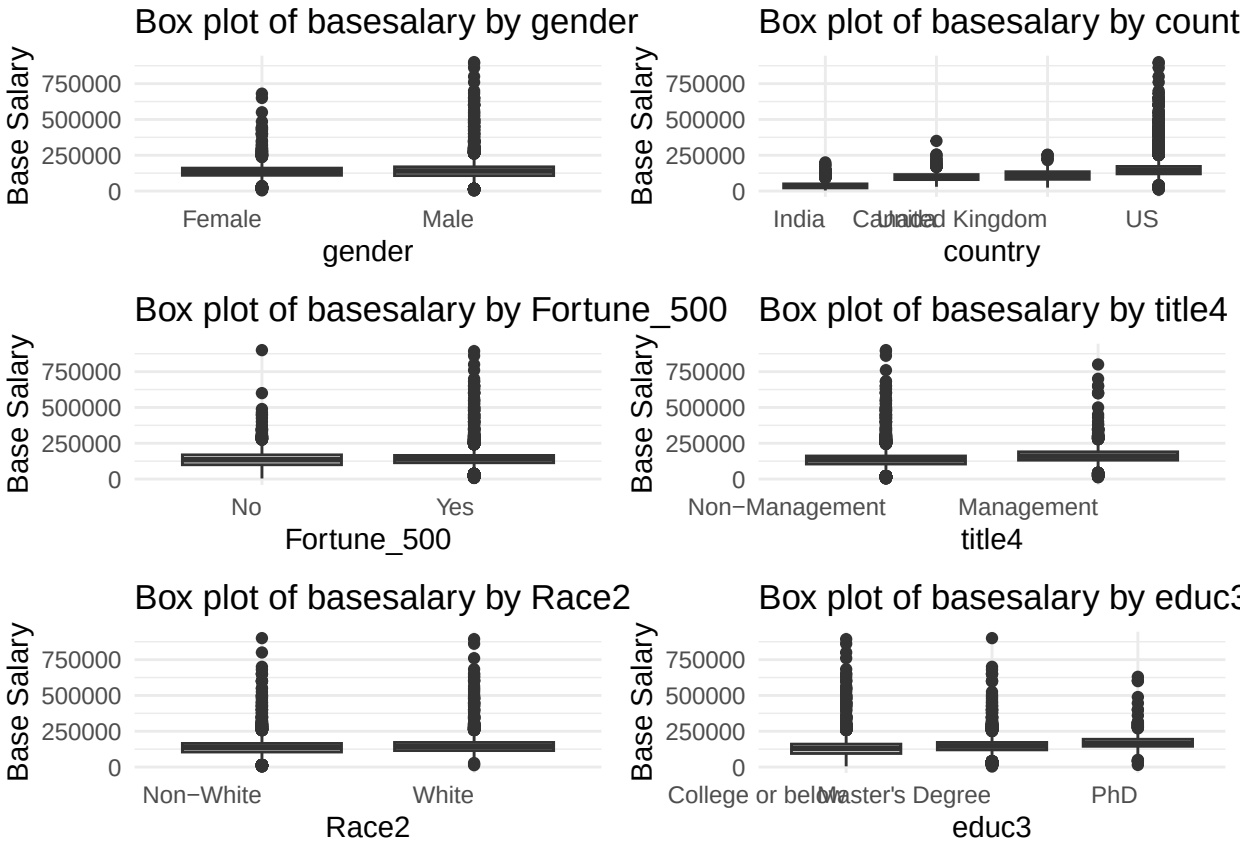
```
## Warning: package 'gridExtra' was built under R version 4.4.2
```

```
# List of columns to exclude
exclude_columns <- c("company", "Education", "tag", "title", "Race")

# Identify categorical columns
categorical_columns <- STEM2 %>%
  select(where(is.factor)) %>%
  select(-all_of(exclude_columns)) %>%
  colnames()

# Create a list to store the plots
```

```r
plot_list <- list()

# Loop through each categorical column and create a box plot
for (cat_col in categorical_columns) {
  # Calculate median salary for each category
  median_data <- STEM2 %>%
    group_by(.data[[cat_col]]) %>%
    summarise(median_salary = median(basesalary, na.rm = TRUE)) %>%
    arrange(median_salary)

  # Reorder the factor levels based on the median salary
  STEM2[[cat_col]] <- factor(STEM2[[cat_col]],
                             levels = median_data[[cat_col]][order(median_data$median_salary)])

  # Create the box plot with ordered x-axis based on the median salary
  p <- ggplot(STEM2, aes_string(x = cat_col, y = "basesalary")) +
    geom_boxplot() +
    labs(title = paste("Box plot of basesalary by", cat_col),
         x = cat_col,
         y = "Base Salary") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 0, hjust = 1))

  # Append the plot to the list
  plot_list[[cat_col]] <- p
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
# Arrange the plots in a 2x2 grid and display them
do.call(grid.arrange, c(plot_list, ncol = 2))
```



Multicollinearity

Before proceeding with model building, we need to decide which variables to keep. Based on our pair plots and heatmaps, there is clear evidence of multicollinearity. This was further confirmed by fitting a simple model with total yearly compensation as the dependent variable.

1. Global utility of the model is highly significant, but t-tests for individual beta's are insignificant.
2. Negative values for education (PhD, Masters), country (US) even when we expect a positive relationship against the totalyearlycompensation (Y).
3. Although VIF values did not exceed 10, suggesting that multicollinearity is not a significant concern, it is important to note that basesalary, stockgrantvalue, and bonus together form the totalyearlycompensation.

With this, we will be using basesalary as our Y as we build our model.

- **Remove**: totalyearlycompensation, stockgrantvalue, bonus
- **Keep**: basesalary

```
# Fit the model with selected variables (Multicollinearity)
M=lm(totalyearlycompensation~basesalary+yearsofexperience+yearsatcompany+
       stockgrantvalue+bonus+gender+Race2+educ3+country+Fortune_500+title4,
     data = STEM2)
summary(M)
```

```
##
## Call:
## lm(formula = totalyearlycompensation ~ basesalary + yearsofexperience +
##     yearsatcompany + stockgrantvalue + bonus + gender + Race2 +
##     educ3 + country + Fortune_500 + title4, data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185329   -2878     141    2560 4339340
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.075e+03  1.566e+03  -1.964  0.04958 *
## basesalary                1.072e+00  9.485e-03 113.056  < 2e-16 ***
## yearsofexperience        -2.282e+02  8.348e+01  -2.733  0.00628 **
## yearsatcompany            5.077e+01  1.234e+02   0.412  0.68070
## stockgrantvalue           9.592e-01  5.267e-03 182.124  < 2e-16 ***
## bonus                     1.203e+00  1.562e-02  77.055  < 2e-16 ***
## genderMale                3.890e+02  9.085e+02   0.428  0.66853
## Race2White                1.858e+01  7.813e+02   0.024  0.98103
## educ3Master's Degree     -4.331e+01  7.446e+02  -0.058  0.95362
## educ3PhD                 -3.373e+03  1.712e+03  -1.970  0.04883 *
## countryCanada            -4.479e+03  2.182e+03  -2.052  0.04015 *
## countryUnited Kingdom    -5.936e+03  2.537e+03  -2.340  0.01931 *
## countryUS                -8.321e+03  1.614e+03  -5.154 2.57e-07 ***
## Fortune_500Yes            3.245e+02  7.171e+02   0.452  0.65094
## title4Management          2.900e+03  9.799e+02   2.960  0.00308 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47160 on 18910 degrees of freedom
## Multiple R-squared:  0.8793, Adjusted R-squared:  0.8792
## F-statistic:  9842 on 14 and 18910 DF,  p-value: < 2.2e-16
```

```
# Check VIF
car::vif(M)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## basesalary        2.440887  1        1.562334
## yearsofexperience 2.058451  1        1.434730
## yearsatcompany    1.453117  1        1.205453
## stockgrantvalue   1.347932  1        1.161005
## bonus             1.255090  1        1.120308
## gender            1.025798  1        1.012817
## Race2             1.148987  1        1.071908
## educ3             1.155364  2        1.036764
## country           1.703846  3        1.092878
## Fortune_500       1.041051  1        1.020319
## title4            1.168629  1        1.081031
```

4. Model Building

The next phase is model building. In this phase, we develop a model relating our dependent variable (Y) to the independent variables we've identified. We also explore potential transformations for our quantitative predictors (X) and interactions between quantitative and categorical variables. We then compare models with only the main effects terms to those with transformations and interactions, and use automated variable selection methods to identify the most important predictors.

a) Variable Transformations

We start by writing the model with our quantitative independent variables, beginning with years of experience. Based on our scatterplot, which shows that the relationship between years of experience and base salary is right-skewed, we aim to test for a linear pattern. We try transforming the variable into its quadratic, logarithmic, and square root forms to examine any improvements.

As shown, all individual t-tests for each transformation are significant, and the ANOVA tests indicate that these transformations contribute significantly to predicting Y. However, there is minimal improvement in their RSEs and adjusted R-squared values compared to the main effects model.

Examining closely at their graphs, the linear and quadratic fits appear almost identical, except for differences in their residuals. In contrast, transformations such as the logarithm and square root of X effectively redistribute the data points in the scatterplot, reducing skewness and yielding a more normalized distribution. However, their residuals are still very similar to the linear fit, which is our main effects model.

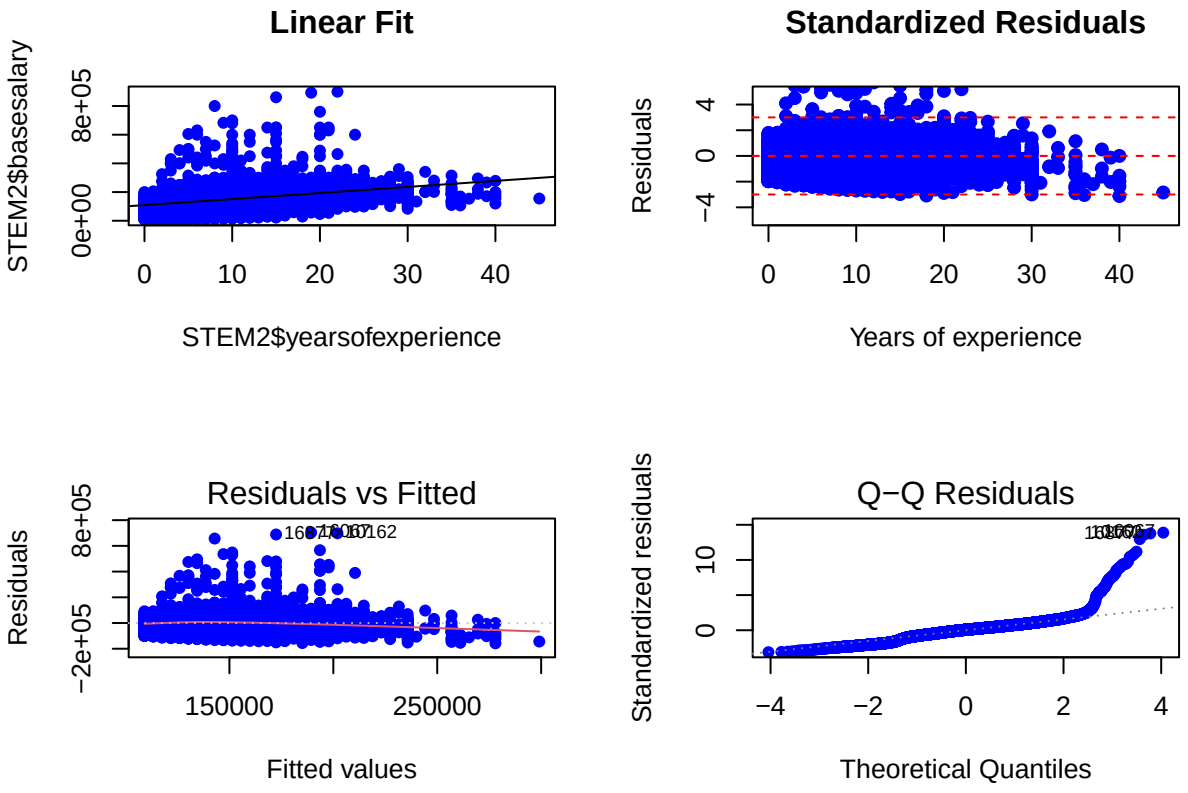
i) yearsofexperience

```
M00=lm(basesalary~yearsofexperience, data = STEM2)
summary(M00)
```

Linear Fit

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience, data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -157996  -25889    3667   26333  703668
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       109111.95     572.42   190.6   <2e-16 ***
## yearsofexperience   4222.10      62.55    67.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50700 on 18923 degrees of freedom
## Multiple R-squared:  0.1941, Adjusted R-squared:  0.194
## F-statistic:  4557 on 1 and 18923 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))

# Scatterplot with regression line
plot(STEM2$basesalary~STEM2$yearsofexperience, pch=16, col="blue",
     main="Linear Fit")
abline(M00)

# Plot the standardized residuals against x
plot(STEM2$yearsofexperience, rstandard(M00), pch=19, col="blue",
     xlab = "Years of experience",
     ylab = "Residuals", ylim=c(-5,5),
     main = "Standardized Residuals")
abline(h = 0, col = "red", lty = 2)
abline(h = c(-3, 0, 3), col = "red", lty = 2)

# Residuals vs. Fitted
plot(M00, 1, pch=16, col="blue")

# Check normality
plot(M00, 2, pch=16, col="blue")
```

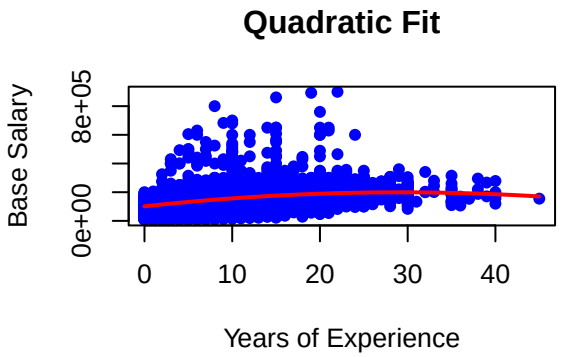


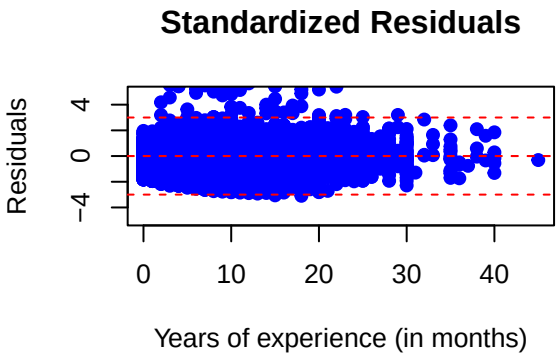

```r
M01=lm(basesalary~yearsofexperience+I(yearsofexperience^2), data = STEM2)
summary(M01)
```

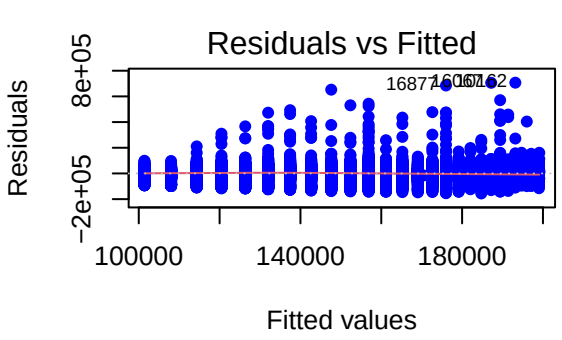
Quadratic transformation on X

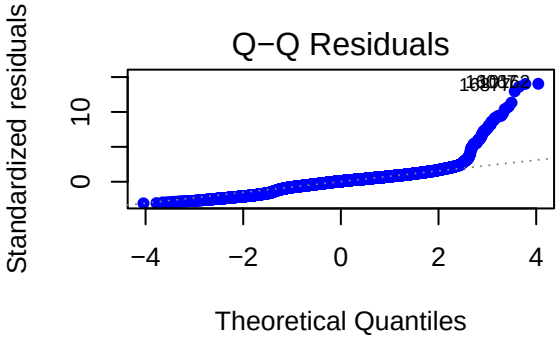
```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + I(yearsofexperience^2),
##     data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -155740  -25457    3383   27012  706831
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             101409.798    744.475  136.22   <2e-16 ***
## yearsofexperience         6693.065    166.222   40.27   <2e-16 ***
## I(yearsofexperience^2)    -114.646      7.153  -16.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50360 on 18922 degrees of freedom
## Multiple R-squared:  0.2049, Adjusted R-squared:  0.2048
## F-statistic:  2438 on 2 and 18922 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))

# Scatterplot with regression line
experience_seq <- seq(min(STEM2$yearsofexperience), max(STEM2$yearsofexperience), length.out = 100)
predicted_salary <- predict(M01, newdata = data.frame(yearsofexperience = experience_seq))
plot(STEM2$basesalary ~ STEM2$yearsofexperience, pch = 16, col = "blue",
     xlab = "Years of Experience", ylab = "Base Salary",
     main = "Quadratic Fit")
lines(experience_seq, predicted_salary, col = "red", lwd = 2)

# Plot the standardized residuals
plot(STEM2$yearsofexperience, rstandard(M01), pch=19, col="blue",
     xlab = "Years of experience (in months)",
     ylab = "Residuals", ylim=c(-5,5),
     main = "Standardized Residuals")
abline(h = 0, col = "red", lty = 2)
abline(h = c(-3, 0, 3), col = "red", lty = 2)

# Residuals vs. Fitted
plot(M01, 1, pch=16, col="blue")

# Check normality
plot(M01, 2, pch=16, col="blue")
```









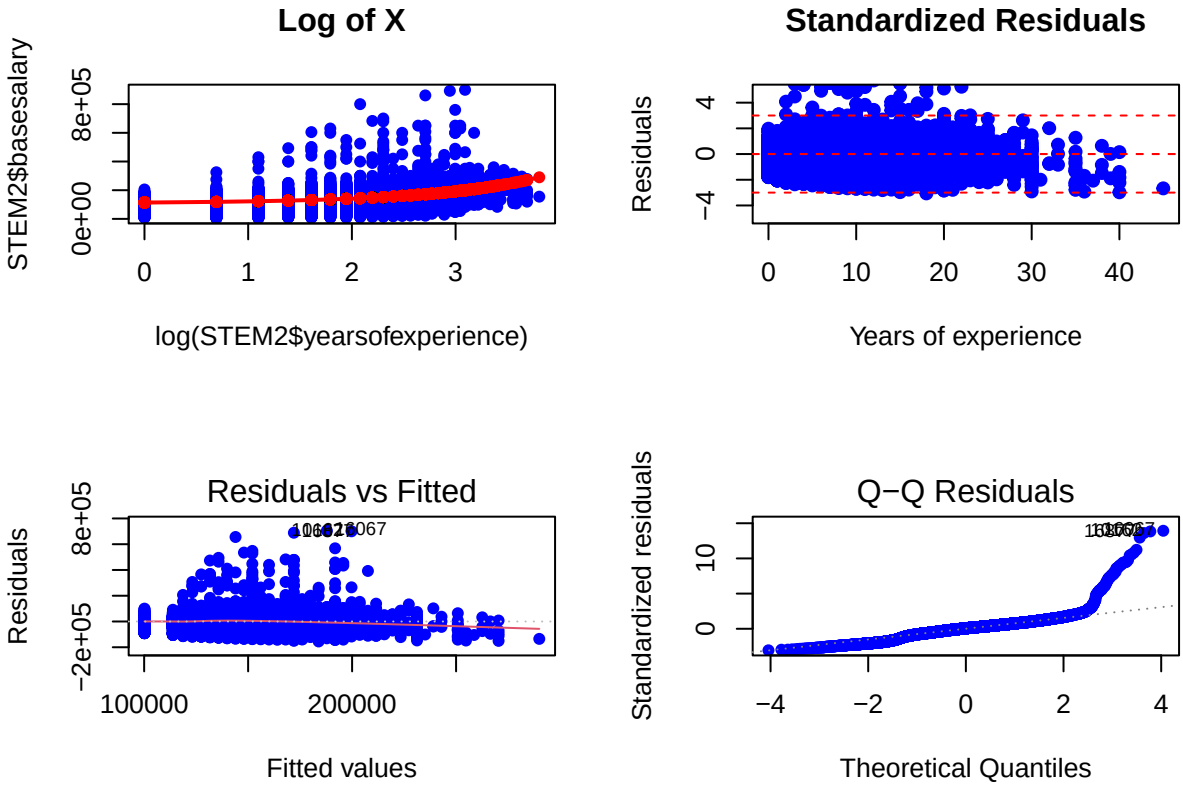

```
# Add a small constant to avoid log(0) or remove the rows with 0 exp. completely
M020_log=lm(basesalary~yearsofexperience+log(yearsofexperience+0.001), data = STEM2)
summary(M020_log)
```

Log transformation on X

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + log(yearsofexperience +
##     0.001), data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154908  -25433    3957   26856  705134
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  109941.35     582.86 188.623  < 2e-16 ***
## yearsofexperience              3881.63      78.01  49.756  < 2e-16 ***
## log(yearsofexperience + 0.001)  1417.46     194.59   7.284 3.36e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 50630 on 18922 degrees of freedom
## Multiple R-squared:  0.1963, Adjusted R-squared:  0.1962
## F-statistic:  2311 on 2 and 18922 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))

# Scatterplot with regression line
plot(STEM2$basesalary~log(STEM2$yearsofexperience), main="Log of X", pch=16, col="blue")
predicted <- predict(M020_log, newdata = STEM2)
points(log(STEM2$yearsofexperience), predicted, col = "red", pch = 16)
lines(sort(log(STEM2$yearsofexperience)),
      predicted[order(log(STEM2$yearsofexperience))],
      col = "red", lwd = 2)

# Plot the standardized residuals
plot(STEM2$yearsofexperience, rstandard(M020_log), pch=19, col="blue",
     xlab = "Years of experience",
     ylab = "Residuals", ylim=c(-5,5),
     main = "Standardized Residuals")
abline(h = 0, col = "red", lty = 2)
abline(h = c(-3, 0, 3), col = "red", lty = 2)

# Residuals vs. Fitted
plot(M020_log, 1, pch=16, col="blue")

# Check normality
plot(M020_log, 2, pch=16, col="blue")
```



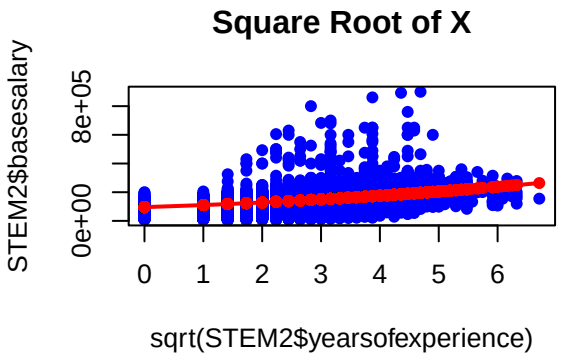

```
M020_sqrt=lm(basesalary~yearsofexperience+sqrt(yearsofexperience), data = STEM2)
summary(M020_sqrt)
```

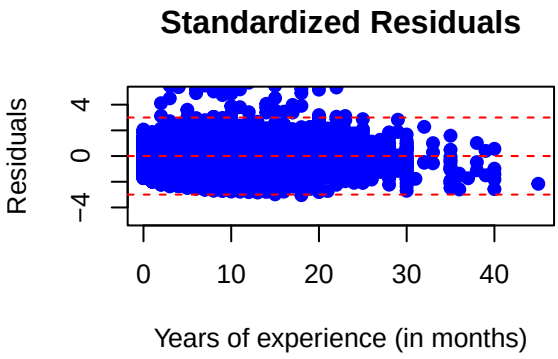
Square root of X

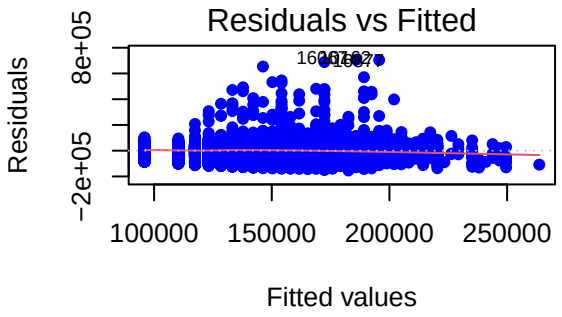
```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience),
##     data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153565  -25252    3649   26764  707119
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               95894.1     1187.1  80.779   <2e-16 ***
## yearsofexperience          1858.3      196.4   9.464   <2e-16 ***
## sqrt(yearsofexperience)   12544.4      988.3  12.694   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50480 on 18922 degrees of freedom
```

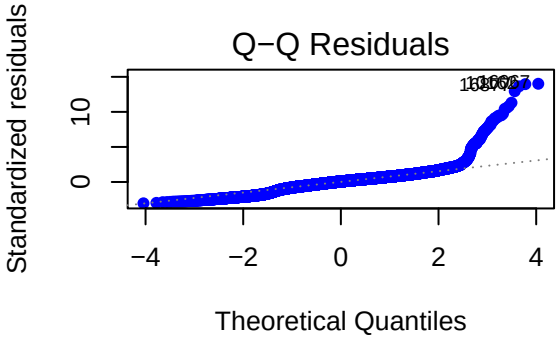
```
## Multiple R-squared:  0.2009, Adjusted R-squared:  0.2008
## F-statistic:  2378 on 2 and 18922 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))

# Scatterplot with regression line
plot(STEM2$basesalary~sqrt(STEM2$yearsofexperience), main="Square Root of X", pch=16, col="blue")
predicted <- predict(M020_sqrt, newdata = STEM2)
points(sqrt(STEM2$yearsofexperience), predicted, col = "red", pch = 16)
lines(sort(sqrt(STEM2$yearsofexperience)),
      predicted[order(sqrt(STEM2$yearsofexperience))],
      col = "red", lwd = 2)

# Plot the standardized residuals
plot(STEM2$yearsofexperience, rstandard(M020_sqrt), pch=19, col="blue",
     xlab = "Years of experience (in months)",
     ylab = "Residuals", ylim=c(-5,5),
     main = "Standardized Residuals")
abline(h = 0, col = "red", lty = 2)
abline(h = c(-3, 0, 3), col = "red", lty = 2)

# Residuals vs. Fitted
plot(M020_sqrt, 1, pch=16, col="blue")

# Check normality
plot(M020_sqrt, 2, pch=16, col="blue")
```



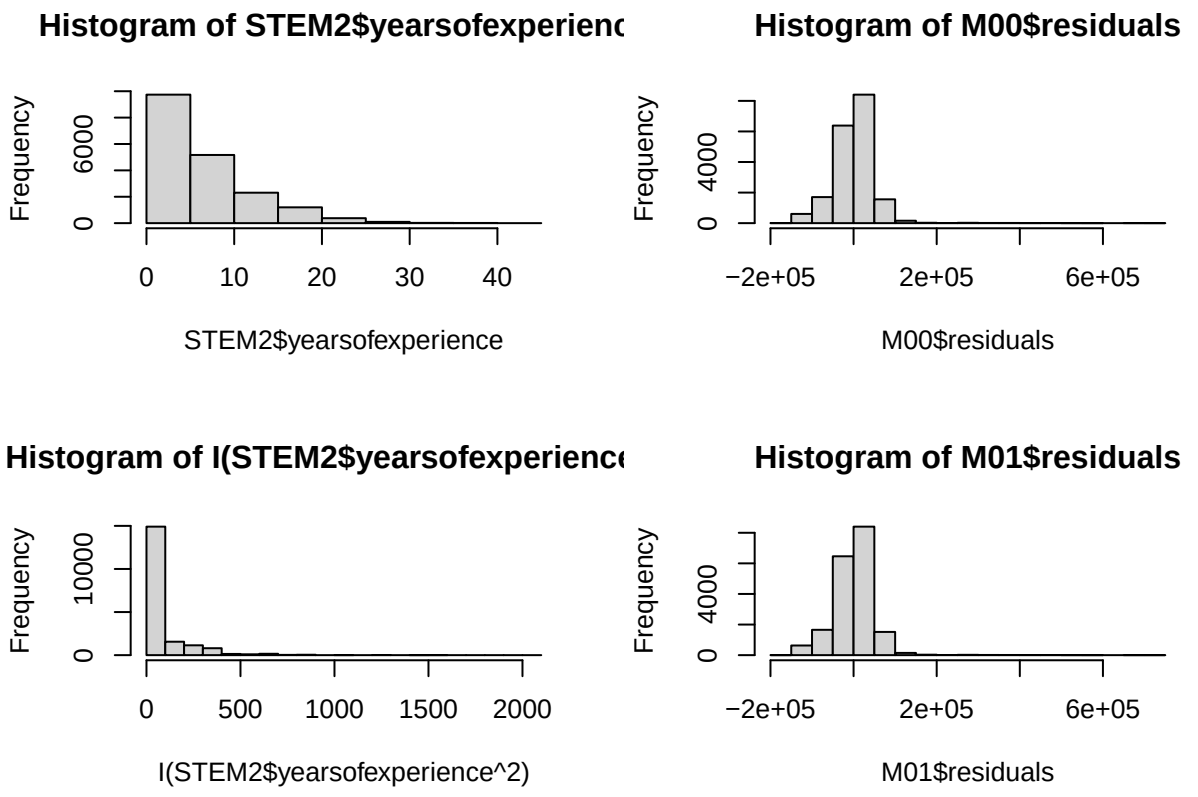




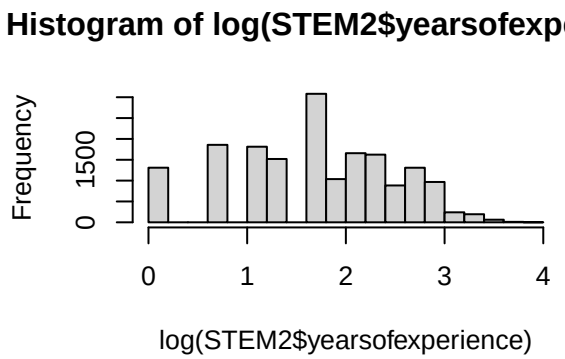


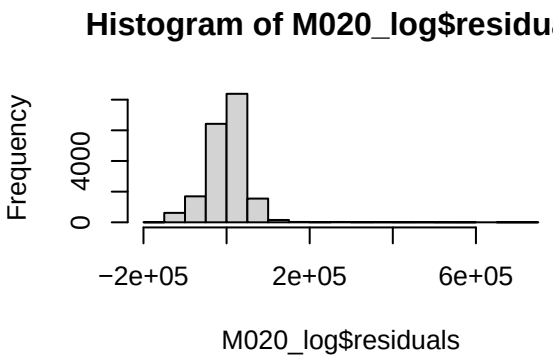
Histogram comparisons: Main Effect vs. Quadratic vs. Log vs. Square Root

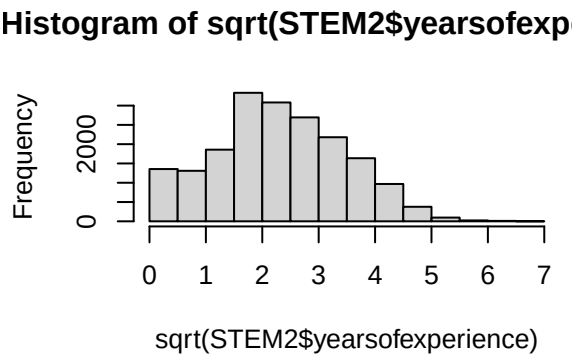
```
par(mfrow = c(2, 2))
hist(STEM2$yearsofexperience)
hist(M00$residuals)
hist(I(STEM2$yearsofexperience^2))
hist(M01$residuals)
```

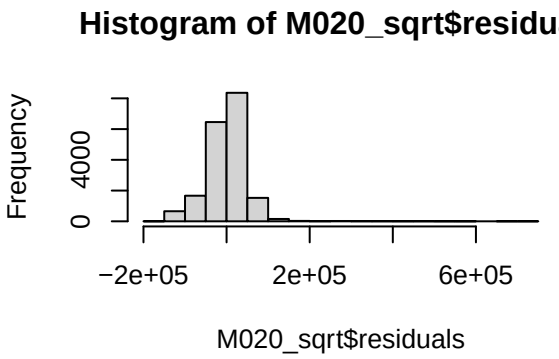



```
par(mfrow = c(2, 2))
hist(log(STEM2$yearsofexperience))
hist(M020_log$residuals)
hist(sqrt(STEM2$yearsofexperience))
hist(M020_sqrt$residuals)
```









ii) yearsatcompany A similar pattern is observed with the second quantitative variable, years at company. Although individual t-tests and ANOVA show that the transformations are statistically significant, they result in only minimal improvements in RSE and adjusted R-squared values compared to the main effects model.

```
M1_comp=lm(basesalary~yearsatcompany, data = STEM2)
summary(M1_comp)
```

Linear Fit

```
##
## Call:
## lm(formula = basesalary ~ yearsatcompany, data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -147628  -29992     289   28529  757049
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    129992.2      516.0  251.90   <2e-16 ***
## yearsatcompany   3239.6      120.3   26.94   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55420 on 18923 degrees of freedom
## Multiple R-squared:  0.03693,    Adjusted R-squared:  0.03688
## F-statistic: 725.6 on 1 and 18923 DF,  p-value: < 2.2e-16
```

```
M2_comp=lm(basesalary~yearsatcompany+I(yearsatcompany^2), data = STEM2)
summary(M2_comp)
```

Quadratic transformation on X

```
##
## Call:
## lm(formula = basesalary ~ yearsatcompany + I(yearsatcompany^2),
##     data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148050  -29320     680   28344  756300
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         129320.17     595.23 217.261   <2e-16 ***
## yearsatcompany        3741.11     252.00  14.846   <2e-16 ***
## I(yearsatcompany^2)    -36.55      16.14  -2.265   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55410 on 18922 degrees of freedom
## Multiple R-squared:  0.03719,    Adjusted R-squared:  0.03709
## F-statistic: 365.4 on 2 and 18922 DF,  p-value: < 2.2e-16
```

```
M3_comp=lm(basesalary~yearsatcompany+log(yearsatcompany+0.001), data = STEM2)
summary(M3_comp)
```

Log transformation on X

```
##
## Call:
## lm(formula = basesalary ~ yearsatcompany + log(yearsatcompany +
##     0.001), data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152383  -29922     919   28144  758876
##
```

```
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  126721.7      677.7 186.993  < 2e-16 ***
## yearsatcompany                 3969.8      155.2  25.583  < 2e-16 ***
## log(yearsatcompany + 0.001)  -1065.4      143.4  -7.431 1.13e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55340 on 18922 degrees of freedom
## Multiple R-squared:  0.03973,    Adjusted R-squared:  0.03963
## F-statistic: 391.4 on 2 and 18922 DF,  p-value: < 2.2e-16
```

```
M4_comp=lm(basesalary~yearsatcompany+sqrt(yearsatcompany), data = STEM2)
summary(M4_comp)
```

Square Root of X

```
##
## Call:
## lm(formula = basesalary ~ yearsatcompany + sqrt(yearsatcompany),
##     data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151412  -30039     532   27625  758676
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           132468.1      730.1 181.434  < 2e-16 ***
## yearsatcompany          4520.9      293.2  15.418  < 2e-16 ***
## sqrt(yearsatcompany)  -4614.0      963.1  -4.791 1.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55390 on 18922 degrees of freedom
## Multiple R-squared:  0.03809,    Adjusted R-squared:  0.03799
## F-statistic: 374.7 on 2 and 18922 DF,  p-value: < 2.2e-16
```

b) Interaction Terms

Next, we introduce the main effects and interaction terms for the qualitative independent variables to examine whether the effect of one predictor on the response variable depends on the level of another predictor. With 8 predictors in total, there are 28 possible two-way combinations, and 18 of these were found to be significant. However, when comparing the model with only the main effects of these two predictors to the model with their two-way interaction, we observed only a slight improvement in both the RSE and adjusted R-squared.

```
response_var <- "basesalary"
predictor_vars <- c("yearsofexperience", "yearsatcompany", "gender", "Race2",
```

```
                    "educ3", "country", "title4", "Fortune_500")

results <- interaction_model_stats(data = STEM2, response = response_var,
                                   predictors = predictor_vars)
print(results)
```

Two-way interaction

```
##                          Interaction      P_value      RSE Adjusted_R_squared
## 26                   country:title4 4.303841e-02 46158.25        0.331838364
## 4           yearsofexperience:educ3 2.628751e-11 49522.39        0.230894148
## 1  yearsofexperience:yearsatcompany 1.901042e-09 50545.89        0.198774745
## 3           yearsofexperience:Race2 1.159191e-02 50574.35        0.197872183
## 6          yearsofexperience:title4 6.033294e-08 50612.72        0.196654661
## 7     yearsofexperience:Fortune_500 2.787597e-03 50638.19        0.195845832
## 10             yearsatcompany:educ3 4.736566e-13 54036.92        0.084277105
## 24                     educ3:title4 4.603539e-07 54056.69        0.083606863
## 19                      Race2:educ3 1.221695e-13 54351.44        0.073586156
## 12            yearsatcompany:title4 1.863536e-09 54635.47        0.063878305
## 25                educ3:Fortune_500 1.169183e-02 55035.64        0.050115130
## 15                     gender:educ3 3.689430e-02 55090.93        0.048205472
## 21                     Race2:title4 4.590862e-03 55091.67        0.048180062
## 13       yearsatcompany:Fortune_500 8.290504e-06 55320.22        0.040266366
## 17                   gender:title4 3.541960e-03 55404.73        0.037331942
## 8             yearsatcompany:gender 2.893273e-02 55413.73        0.037018976
## 14                     gender:Race2 3.402299e-05 56065.30        0.014240058
## 5         yearsofexperience:country 1.536429e-01 40758.09        0.479032664
## 11           yearsatcompany:country 3.051512e-01 46206.02        0.330454565
## 23                     educ3:country 2.646372e-01 46416.90        0.324329218
## 16                   gender:country 1.871869e-01 47013.95        0.306835447
## 27              country:Fortune_500 7.872380e-02 47084.40        0.304756302
## 20                    Race2:country 7.990564e-01 47101.67        0.304246262
## 2          yearsofexperience:gender 4.619018e-01 50680.90        0.194488803
## 9              yearsatcompany:Race2 5.195121e-02 55134.26        0.046707964
## 28               title4:Fortune_500 6.034254e-01 55356.69        0.039000371
## 22                Race2:Fortune_500 6.173549e-01 55950.72        0.018265069
## 18               gender:Fortune_500 4.559360e-01 56350.51        0.004185023
##    Significant
## 26        TRUE
## 4         TRUE
## 1         TRUE
## 3         TRUE
## 6         TRUE
## 7         TRUE
## 10        TRUE
## 24        TRUE
## 19        TRUE
## 12        TRUE
## 25        TRUE
## 15        TRUE
## 21        TRUE
## 13        TRUE
## 17        TRUE
```

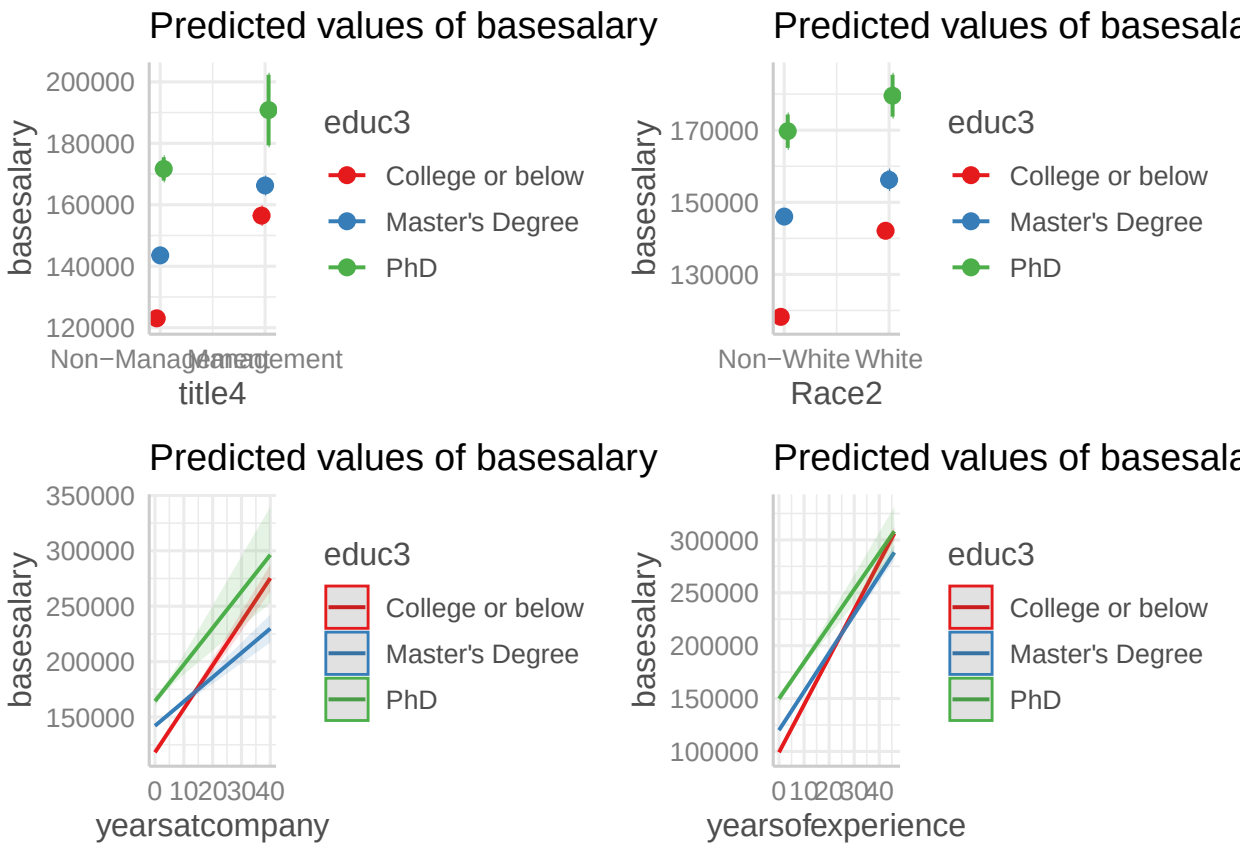
```
## 8          TRUE
## 14         TRUE
## 5         FALSE
## 11        FALSE
## 23        FALSE
## 16        FALSE
## 27        FALSE
## 20        FALSE
## 2         FALSE
## 9         FALSE
## 28        FALSE
## 22        FALSE
## 18        FALSE
```

We visually inspected the interactions of these predictors, focusing on those with a greater than 0.10% improvement in the adjusted R-squared.

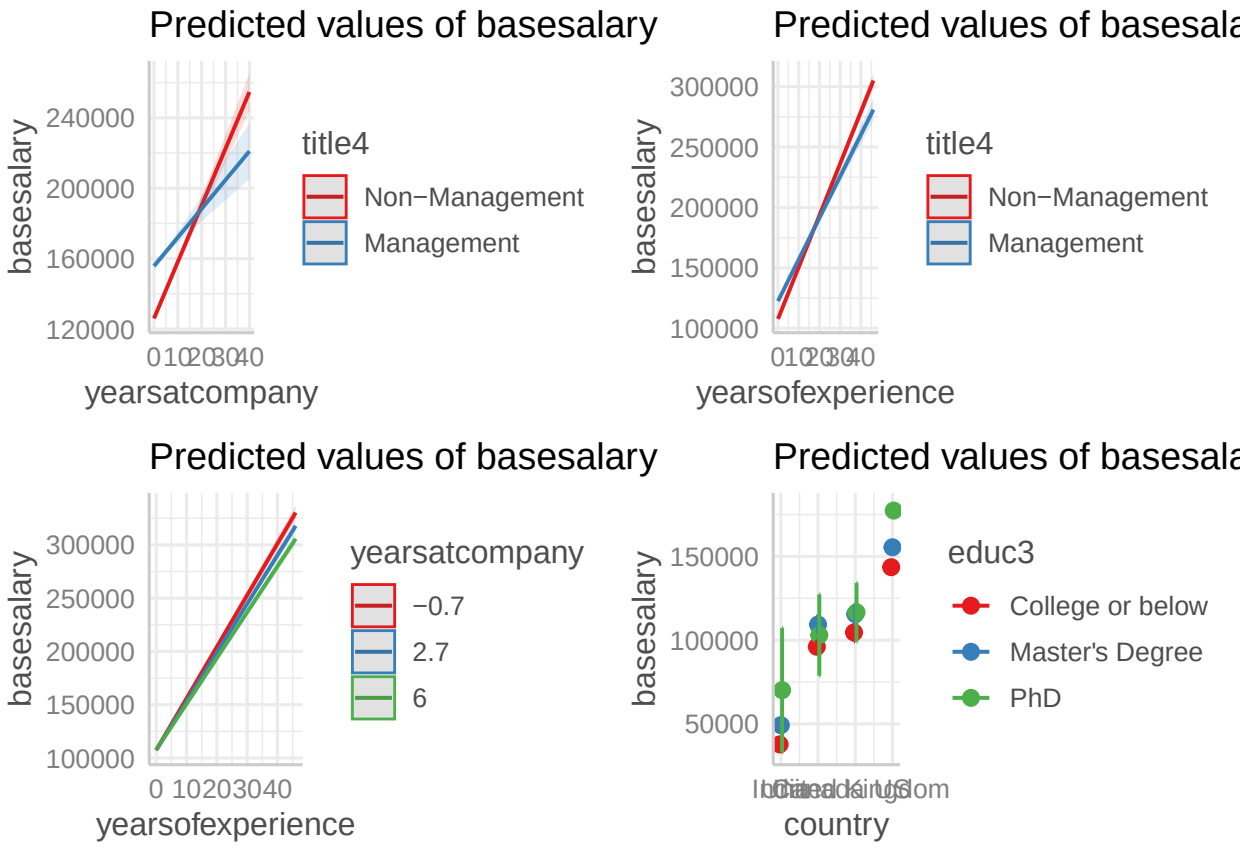
- The base salary increase for College graduates (or lower) with more years at the company or with more experience is slightly higher compared to those with Master's or Doctoral degrees. Individuals with advanced degrees often start with higher salaries, but their salary growth tends to slow as they've already reached a high initial value. On the other hand, College graduates, starting with lower salaries, see faster growth as they gain experience and seniority, particularly if they stay with the company longer. This suggests that companies may prioritize compensating for experience and tenure in lower-educated employees, which results in steeper salary increases over time, aligning with the idea that experience can sometimes substitute for formal education.
- We also see that base salary growth is slightly higher for non-management roles with more years at the company or with more experience compared to management roles. Non-management positions often require specialized technical skills, which are in high demand and continue to grow in value, leading to higher salary increases. In contrast, management roles focus on leadership and strategy, which may not see the same market-driven salary growth as technical skills, which directly impact productivity. As industries, especially tech, increasingly value technical talent, non-management roles often experience faster salary growth, creating a wage premium over management positions.

```
library(grid)
library(sjPlot)
library(sjmisc)
library(ggplot2)
library(gridExtra)

# Ensure consistent theme for all plots
theme_set(theme_sjplot())

# List of models and terms
models <- list(
  list(model = lm(basesalary ~ title4 * educ3, data = STEM2), terms = c("title4", "educ3")),
  list(model = lm(basesalary ~ Race2 * educ3, data = STEM2), terms = c("Race2", "educ3")),
  list(model = lm(basesalary ~ yearsatcompany * educ3, data = STEM2), terms = c("yearsatcompany", "educ
  list(model = lm(basesalary ~ yearsofexperience * educ3, data = STEM2), terms = c("yearsofexperience",
  list(model = lm(basesalary ~ yearsatcompany * title4, data = STEM2), terms = c("yearsatcompany", "tit
  list(model = lm(basesalary ~ yearsofexperience * title4, data = STEM2), terms = c("yearsofexperience"
  list(model = lm(basesalary ~ yearsofexperience * yearsatcompany, data = STEM2), terms = c("yearsofexpe
  list(model = lm(basesalary ~ country * educ3, data = STEM2), terms = c("country", "educ3"))
)
```

```r
# Generate the plots
plots <- lapply(models, function(x) {
  plot_model(x$model, type = "pred", terms = x$terms)
})

# Arrange plots in a 2x2 grid
grid.arrange(grobs = plots[1:4], ncol = 2, nrow = 2)
```




```r
grid.arrange(grobs = plots[5:8], ncol = 2, nrow = 2)
```

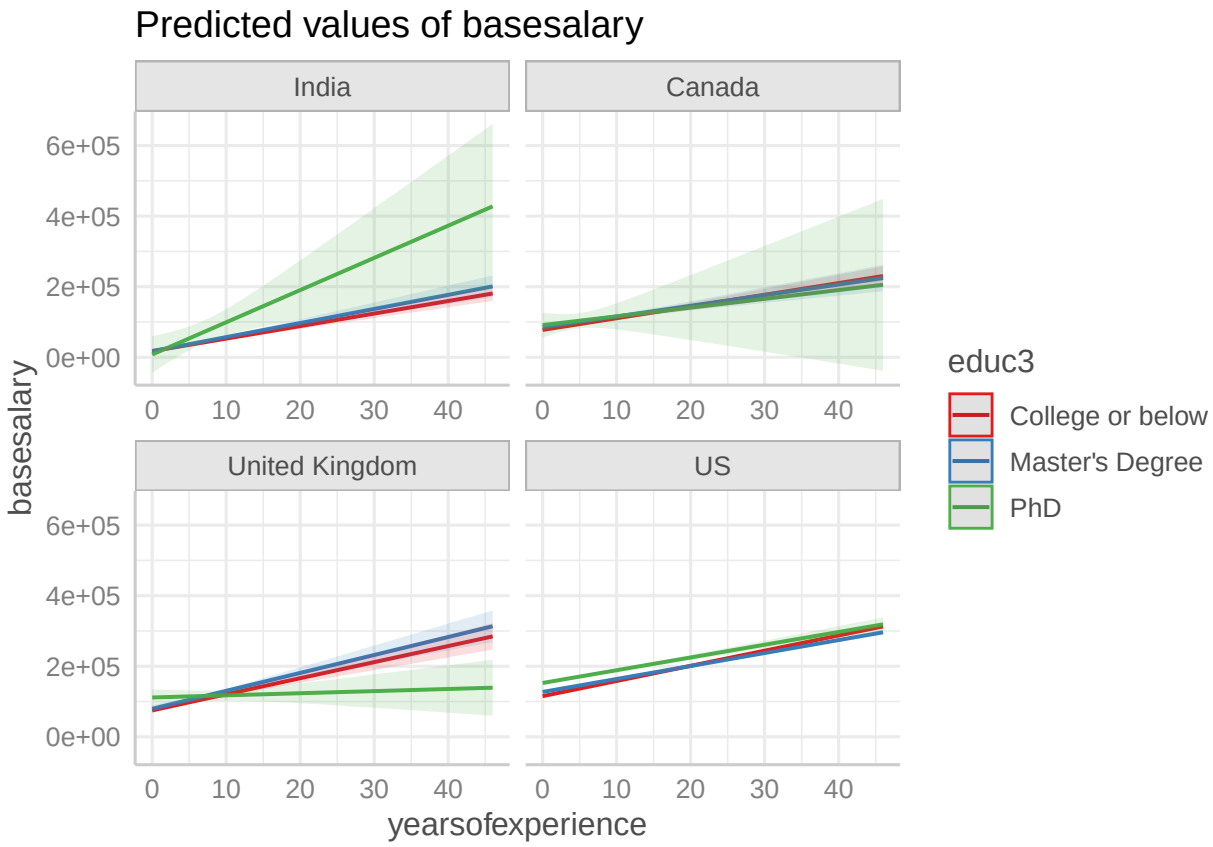


Three-way interaction We also explored 3-way and 4-way interaction plots but encountered a rank deficiency issue, which prevents the estimation of unique coefficients for all predictors. This is likely caused by overfitting, or too many categorical variables with multiple levels, or insufficient data to support higher-order interactions.

As a result, we decided not to pursue these higher-order interaction terms, as they are not theoretically justified and do not contribute significantly to the model.

```
data <- STEM2
response <- "basesalary"
predictors <- c("yearsofexperience", "yearsatcompany", "gender", "Race2",
                "educ3", "country", "title4", "Fortune_500")
result <- three_way_interaction_model_stats(data, response, predictors)
#print(result)
```

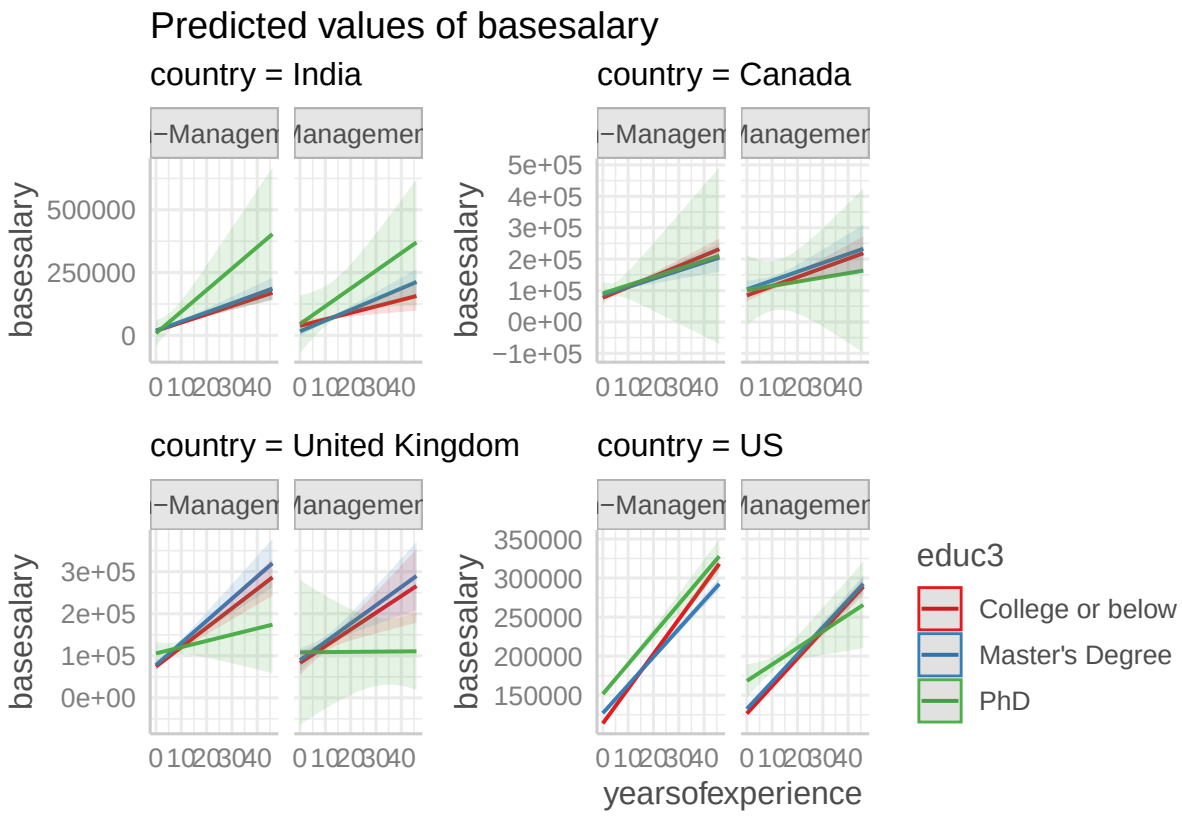
```
fit3 <- lm(basesalary ~ yearsofexperience * educ3 * country, data = STEM2)
plot_model(fit3, type = "pred", terms = c("yearsofexperience", "educ3", "country"))
```




```
fit4 <- lm(basesalary ~ yearsofexperience * educ3 * title4 * country, data = STEM2)
plot_model(fit4, type = "pred", terms = c("yearsofexperience", "educ3", "title4", "country"))
```

Four-way interaction

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



c) Main Effects Model

Now that we have established what transformations or interactions not to include, we are ready to build our model, starting with the main effects. Fitting all 8 predictors against the base salary (Y), we initially identified an insignificant predictor, Fortune 500, which we removed.

```
Main <- lm(formula = basesalary ~ yearsofexperience + yearsatcompany + gender +
             Race2 + educ3 + country + title4 + Fortune_500, data = STEM2)
summary(Main)
```

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500,
##     data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -172208  -21622   -1892   16809  704778
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             74650.07    1534.80  48.638  < 2e-16 ***
## yearsofexperience        4360.86      62.58  69.690  < 2e-16 ***
## yearsatcompany          -1537.64     103.72 -14.824  < 2e-16 ***
```

```
## genderFemale           -3285.68     767.68  -4.280 1.88e-05 ***
## Race2White             -2462.59     660.46  -3.729 0.000193 ***
## educ3Master's Degree    6258.08     628.59   9.956  < 2e-16 ***
## educ3PhD               31353.53    1425.72  21.991  < 2e-16 ***
## countryIndia          -59812.66    1791.92 -33.379  < 2e-16 ***
## countryUnited Kingdom   5644.57    2293.34   2.461 0.013853 *
## countryUS              45787.96    1487.98  30.772  < 2e-16 ***
## title4Management        4051.01     826.15   4.903 9.49e-07 ***
## Fortune_500Yes          -308.92     602.81  -0.512 0.608334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39920 on 18913 degrees of freedom
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.5002
## F-statistic:  1723 on 11 and 18913 DF,  p-value: < 2.2e-16
```

After this adjustment, all remaining predictors were significant, and the model's adjusted R-squared improved to 50%, indicating moderate fit. The overall model is also statistically significant.

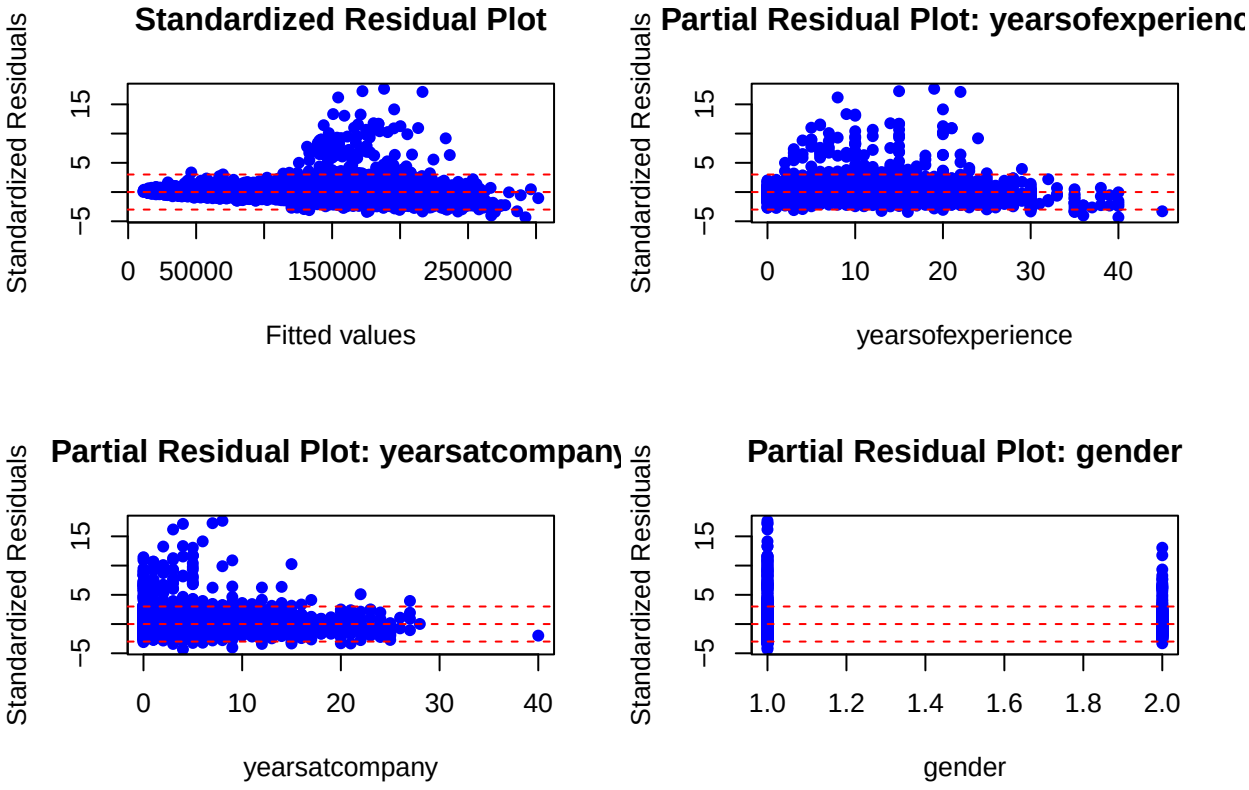
```
Main_Fortune <- lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
                     gender + Race2 + educ3 + country + title4, data = STEM2)
summary(Main_Fortune)
```

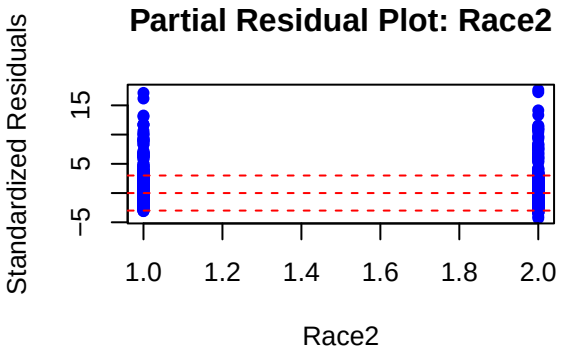
```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + educ3 + country + title4, data = STEM2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -172329  -21606   -1925   16777  704653
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             74517.36    1512.77  49.259  < 2e-16 ***
## yearsofexperience        4360.82      62.57  69.691  < 2e-16 ***
## yearsatcompany          -1540.55     103.57 -14.875  < 2e-16 ***
## genderFemale            -3287.76     767.66  -4.283 1.85e-05 ***
## Race2White              -2444.29     659.48  -3.706 0.000211 ***
## educ3Master's Degree     6242.27     627.82   9.943  < 2e-16 ***
## educ3PhD                31305.07    1422.55  22.006  < 2e-16 ***
## countryIndia           -59832.75    1791.45 -33.399  < 2e-16 ***
## countryUnited Kingdom    5622.27    2292.88   2.452 0.014213 *
## countryUS               45743.20    1485.38  30.796  < 2e-16 ***
## title4Management         4027.31     824.84   4.883 1.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39920 on 18914 degrees of freedom
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.5003
## F-statistic:  1895 on 10 and 18914 DF,  p-value: < 2.2e-16
```

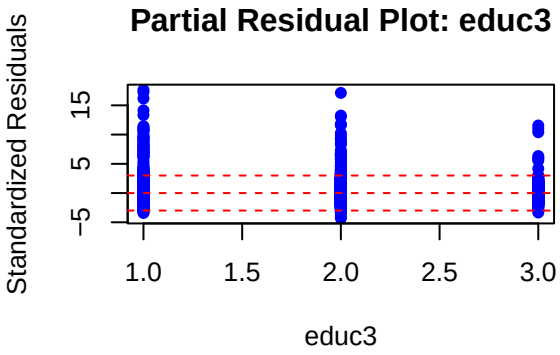
Residual Plots (Detect Lack of Fit) We inspected the residual plots of the main effects model to assess potential lack of fit. Standardized residuals were plotted against the predicted values (Y) and also against

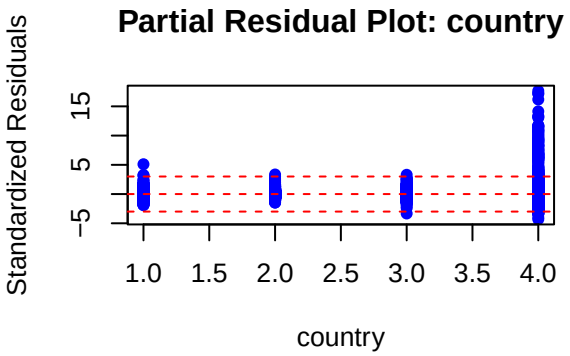
each of the independent variables. Several residuals lie beyond 2 standard deviations from the mean. We will be removing these outliers to evaluate whether this improves the model's predictive power. It is also important to note these residuals constitute only 2.5% of the dataset, which falls below the 5% threshold, making their removal statistically justifiable.

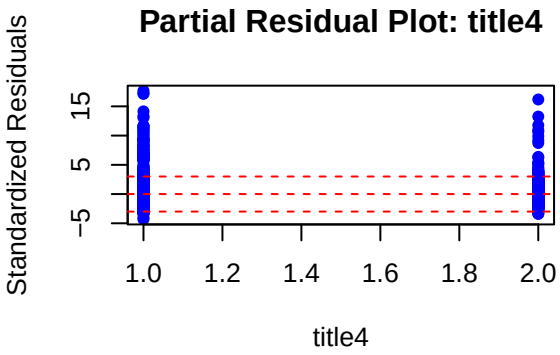
```
par(mfrow = c(2, 2))

# Plot residuals against predicted value
plot(rstandard(Main) ~ predict(Main),
     xlab = "Fitted values", ylab = "Standardized Residuals",
     main = "Standardized Residual Plot", pch=16, col="blue")
abline(h = 0, col = "red", lty = 2)
abline(h = c(-3, 0, 3), col = "red", lty = 2)
# same as above but not standardized
#plot(Main, 1, pch=16, col="blue")

# Plot residuals against each individual x
predictors <- c("yearsofexperience", "yearsatcompany", "gender", "Race2", "educ3", "country", "title4")
plot_residuals(Main, STEM2, predictors)
```



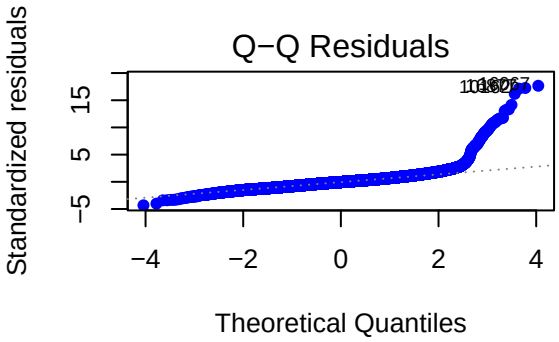









```
# Check normality
plot(Main, 2, pch=16, col="blue")
```




```
# From regression model
outliers <- which(abs(rstandard(Main)) > 2)
#STEM2[outliers, ] #display

# Delete outliers
STEM3 <- STEM2[-outliers, ]
```

Check outliers

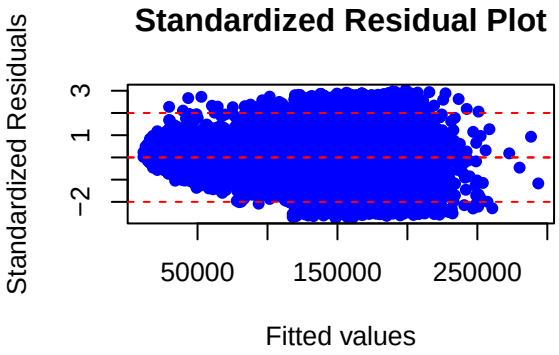
Refit the model After removing these outliers, we observed an immediate improvement in the adjusted R-squared, which increased to 64.27%, representing a 14% improvement from the previous model that included the outliers. The RSE also improved, decreasing to 28,720, compared to the previous value of 39,920 (close to 40k).

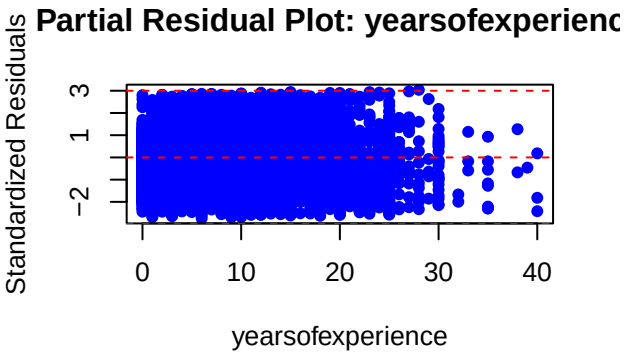
```
# Refit the model
Main <- lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
             gender + Race2 + educ3 + country + title4 + Fortune_500, data = STEM3)
summary(Main)
```

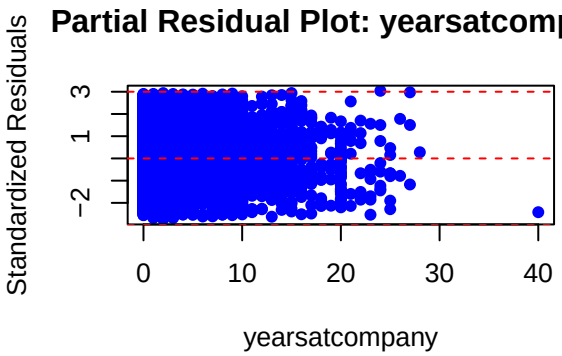
```
##
## Call:
```

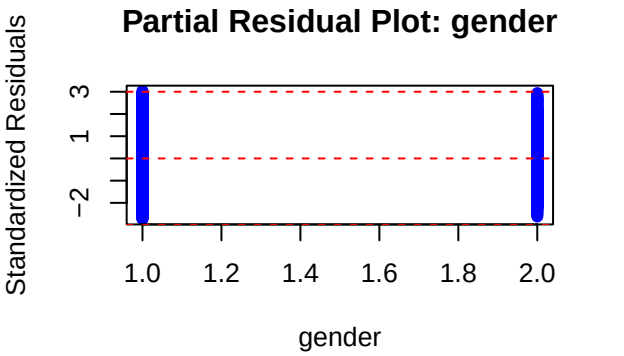
```
## lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500,
##     data = STEM3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78663 -19167   -255  17808  87257
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              75189.50    1111.11  67.670  < 2e-16 ***
## yearsofexperience         4214.92      46.77  90.118  < 2e-16 ***
## yearsatcompany           -1593.09      77.13 -20.655  < 2e-16 ***
## genderFemale             -3127.99     558.62  -5.599 2.18e-08 ***
## Race2White               -2708.69     482.22  -5.617 1.97e-08 ***
## educ3Master's Degree      6942.93     458.02  15.158  < 2e-16 ***
## educ3PhD                 30627.52    1046.01  29.280  < 2e-16 ***
## countryIndia            -59190.02    1293.76 -45.751  < 2e-16 ***
## countryUnited Kingdom     5941.16    1662.34   3.574 0.000352 ***
## countryUS                44426.91    1075.20  41.320  < 2e-16 ***
## title4Management          3570.24     608.43   5.868 4.49e-09 ***
## Fortune_500Yes           -1518.87     439.30  -3.457 0.000547 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28720 on 18442 degrees of freedom
## Multiple R-squared:  0.6429, Adjusted R-squared:  0.6427
## F-statistic:  3019 on 11 and 18442 DF,  p-value: < 2.2e-16
```

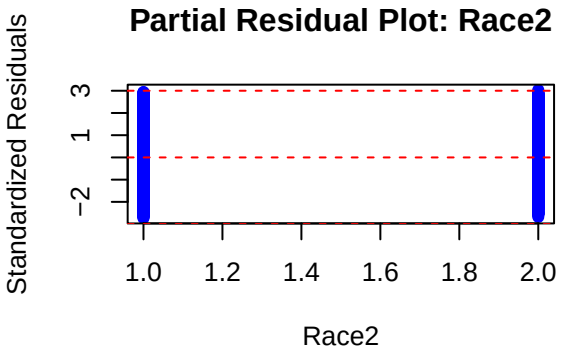
```r
par(mfrow = c(2, 2))

# Plot residuals against predicted value
plot(rstandard(Main) ~ predict(Main),
     xlab = "Fitted values", ylab = "Standardized Residuals",
     main = "Standardized Residual Plot", pch=16, col="blue")
abline(h = 0, col = "red", lty = 2)
abline(h = c(-2, 0, 2), col = "red", lty = 2)
# same as above but not standardized
#plot(Main, 1, pch=16, col="blue")

# Plot residuals against each individual x
predictors <- c("yearsofexperience", "yearsatcompany", "gender", "Race2", "educ3", "country", "title4")
plot_residuals(Main, STEM3, predictors)
```

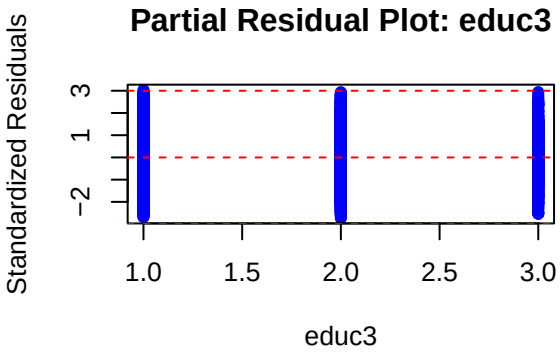

Standardized Residual Plot
Standardized Residuals
−2
1
3
50000
150000
250000
Fitted values

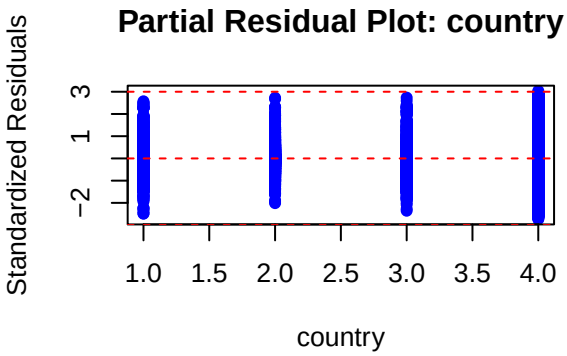

Partial Residual Plot: yearsofexperienc
Standardized Residuals
−2
1
3
0
10
20
30
40
yearsofexperience

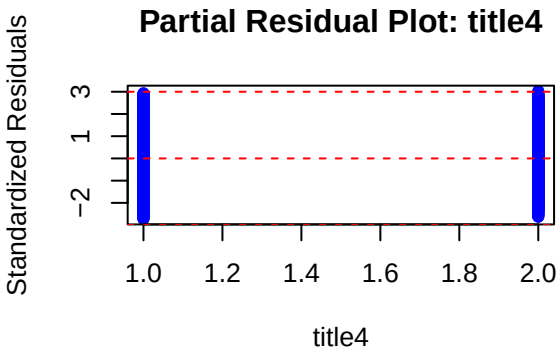

Partial Residual Plot: yearsatcompany
Standardized Residuals
−2
1
3
0
10
20
30
40
yearsatcompany


Partial Residual Plot: gender
Standardized Residuals
−2
1
3
1.0
1.2
1.4
1.6
1.8
2.0
gender

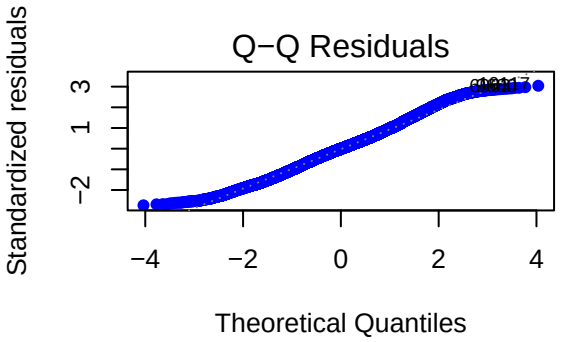









```
# Check normality
plot(Main, 2, pch=16, col="blue")
```



d) Stepwise Regression

Main effects model The second model we developed uses Stepwise Regression, one of the most commonly used variable selection methods. In the first run, we included all the predictors but only their main effects. This model was tested on the data after removing outliers, and as shown, both the RSE and the adjusted R-squared have improved. All individual predictors included in the model are statistically significant, as indicated by their respective t-tests.

```
# Perform stepwise regression on Main effects model
step(lm(formula = basesalary ~ yearsofexperience +
    yearsatcompany + gender + Race2 + educ3 + country + title4 + Fortune_500,
    data = STEM3),direction="both")
```

```
## Start:  AIC=378883.4
## basesalary ~ yearsofexperience + yearsatcompany + gender + Race2 +
##     educ3 + country + title4 + Fortune_500
##
##                     Df  Sum of Sq        RSS    AIC
## <none>                              1.5210e+13 378883
## - Fortune_500        1 9.8591e+09 1.5220e+13 378893
## - gender             1 2.5859e+10 1.5236e+13 378913
## - Race2              1 2.6022e+10 1.5236e+13 378913
## - title4             1 2.8398e+10 1.5238e+13 378916
## - yearsatcompany     1 3.5186e+11 1.5562e+13 379303
## - educ3              2 7.8212e+11 1.5992e+13 379805
```

```
## - yearsofexperience  1 6.6980e+12 2.1908e+13 385615
## - country            3 1.4799e+13 3.0009e+13 391418

##
## Call:
## lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500,
##     data = STEM3)
##
## Coefficients:
##           (Intercept)      yearsofexperience         yearsatcompany
##                 75189                   4215                  -1593
##          genderFemale             Race2White   educ3Master's Degree
##                 -3128                  -2709                   6943
##              educ3PhD           countryIndia  countryUnited Kingdom
##                 30628                 -59190                   5941
##             countryUS       title4Management         Fortune_500Yes
##                 44427                   3570                  -1519
```

```r
# Run the result of the stepwise regression
M2 <- lm(formula = basesalary ~ yearsofexperience +
    yearsatcompany + gender + Race2 + educ3 + country + title4 + Fortune_500,
    data = STEM3)
summary(M2)
```

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500,
##     data = STEM3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78663 -19167   -255  17808  87257
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              75189.50    1111.11  67.670  < 2e-16 ***
## yearsofexperience         4214.92      46.77  90.118  < 2e-16 ***
## yearsatcompany           -1593.09      77.13 -20.655  < 2e-16 ***
## genderFemale             -3127.99     558.62  -5.599 2.18e-08 ***
## Race2White               -2708.69     482.22  -5.617 1.97e-08 ***
## educ3Master's Degree      6942.93     458.02  15.158  < 2e-16 ***
## educ3PhD                 30627.52    1046.01  29.280  < 2e-16 ***
## countryIndia            -59190.02    1293.76 -45.751  < 2e-16 ***
## countryUnited Kingdom     5941.16    1662.34   3.574 0.000352 ***
## countryUS                44426.91    1075.20  41.320  < 2e-16 ***
## title4Management          3570.24     608.43   5.868 4.49e-09 ***
## Fortune_500Yes           -1518.87     439.30  -3.457 0.000547 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28720 on 18442 degrees of freedom
```

```
## Multiple R-squared:  0.6429, Adjusted R-squared:  0.6427
## F-statistic:  3019 on 11 and 18442 DF,  p-value: < 2.2e-16
```

Main effects model + Transformation + Interaction Next, we ran a second Stepwise Regression, this time incorporating both transformations and interactions identified as highly significant in the earlier analysis. The resulting model, shown on the screen, produced only a modest improvement: a slight increase of just over 1% in adjusted R-squared and a small reduction of RSE by only 100.

```r
# Perform stepwise regression on Main effects model + Transformation + Interaction
step(lm(formula = basesalary ~ yearsofexperience
                  + sqrt(yearsofexperience)
                  + yearsatcompany + gender + Race2 + educ3 + country
                  + title4 + Fortune_500 + yearsofexperience:country +
          educ3:country + yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
          yearsofexperience:title4 + yearsofexperience:Fortune_500 + educ3:title4 +
          yearsatcompany:educ3 + Race2:educ3 + yearsatcompany:title4 + educ3:Fortune_500 +
          gender:educ3 + gender:title4 + yearsatcompany:Fortune_500 + gender:Race2,
                  data = STEM3),direction="both")
```

```
## Start:  AIC=378214.8
## basesalary ~ yearsofexperience + sqrt(yearsofexperience) + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500 +
##     yearsofexperience:country + educ3:country + yearsofexperience:educ3 +
##     yearsofexperience:yearsatcompany + yearsofexperience:title4 +
##     yearsofexperience:Fortune_500 + educ3:title4 + yearsatcompany:educ3 +
##     Race2:educ3 + yearsatcompany:title4 + educ3:Fortune_500 +
##     gender:educ3 + gender:title4 + yearsatcompany:Fortune_500 +
##     gender:Race2
##
##                                    Df  Sum of Sq        RSS    AIC
## - gender:educ3                      2 1.7476e+08 1.4623e+13 378211
## - educ3:Fortune_500                 2 1.3092e+09 1.4624e+13 378212
## - Race2:educ3                       2 1.7430e+09 1.4624e+13 378213
## - gender:Race2                      1 3.4520e+08 1.4623e+13 378213
## - educ3:title4                      2 2.7710e+09 1.4625e+13 378214
## <none>                                           1.4623e+13 378215
## - yearsofexperience:title4          1 1.9257e+09 1.4625e+13 378215
## - gender:title4                     1 2.8210e+09 1.4626e+13 378216
## - educ3:country                     6 1.5563e+10 1.4638e+13 378222
## - yearsatcompany:title4             1 9.7517e+09 1.4632e+13 378225
## - yearsofexperience:educ3           2 1.3204e+10 1.4636e+13 378227
## - yearsatcompany:educ3              2 1.5754e+10 1.4638e+13 378231
## - yearsofexperience:yearsatcompany  1 1.6254e+10 1.4639e+13 378233
## - yearsofexperience:country         3 3.4543e+10 1.4657e+13 378252
## - yearsofexperience:Fortune_500     1 4.0170e+10 1.4663e+13 378263
## - yearsatcompany:Fortune_500        1 7.4293e+10 1.4697e+13 378306
## - sqrt(yearsofexperience)           1 3.1926e+11 1.4942e+13 378611
##
## Step:  AIC=378211
## basesalary ~ yearsofexperience + sqrt(yearsofexperience) + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500 +
##     yearsofexperience:country + educ3:country + yearsofexperience:educ3 +
##     yearsofexperience:yearsatcompany + yearsofexperience:title4 +
```

```
##     yearsofexperience:Fortune_500 + educ3:title4 + yearsatcompany:educ3 +
##     Race2:educ3 + yearsatcompany:title4 + educ3:Fortune_500 +
##     gender:title4 + yearsatcompany:Fortune_500 + gender:Race2
##
##                                    Df  Sum of Sq        RSS    AIC
## - educ3:Fortune_500                 2 1.2911e+09 1.4624e+13 378209
## - Race2:educ3                       2 1.7253e+09 1.4625e+13 378209
## - gender:Race2                      1 4.6148e+08 1.4623e+13 378210
## - educ3:title4                      2 2.7218e+09 1.4626e+13 378210
## <none>                                           1.4623e+13 378211
## - yearsofexperience:title4          1 1.9064e+09 1.4625e+13 378211
## - gender:title4                     1 2.7282e+09 1.4626e+13 378212
## + gender:educ3                      2 1.7476e+08 1.4623e+13 378215
## - educ3:country                     6 1.5610e+10 1.4639e+13 378219
## - yearsatcompany:title4             1 9.7626e+09 1.4633e+13 378221
## - yearsofexperience:educ3           2 1.3155e+10 1.4636e+13 378224
## - yearsatcompany:educ3              2 1.6019e+10 1.4639e+13 378227
## - yearsofexperience:yearsatcompany  1 1.6211e+10 1.4639e+13 378229
## - yearsofexperience:country         3 3.4519e+10 1.4657e+13 378249
## - yearsofexperience:Fortune_500     1 4.0146e+10 1.4663e+13 378260
## - yearsatcompany:Fortune_500        1 7.4366e+10 1.4697e+13 378303
## - sqrt(yearsofexperience)           1 3.1942e+11 1.4942e+13 378608
##
## Step:  AIC=378208.7
## basesalary ~ yearsofexperience + sqrt(yearsofexperience) + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500 +
##     yearsofexperience:country + educ3:country + yearsofexperience:educ3 +
##     yearsofexperience:yearsatcompany + yearsofexperience:title4 +
##     yearsofexperience:Fortune_500 + educ3:title4 + yearsatcompany:educ3 +
##     Race2:educ3 + yearsatcompany:title4 + gender:title4 + yearsatcompany:Fortune_500 +
##     gender:Race2
##
##                                    Df  Sum of Sq        RSS    AIC
## - Race2:educ3                       2 1.8302e+09 1.4626e+13 378207
## - gender:Race2                      1 4.5457e+08 1.4625e+13 378207
## - educ3:title4                      2 2.8680e+09 1.4627e+13 378208
## <none>                                           1.4624e+13 378209
## - yearsofexperience:title4          1 1.9092e+09 1.4626e+13 378209
## - gender:title4                     1 2.7186e+09 1.4627e+13 378210
## + educ3:Fortune_500                 2 1.2911e+09 1.4623e+13 378211
## + gender:educ3                      2 1.5664e+08 1.4624e+13 378212
## - educ3:country                     6 1.5074e+10 1.4639e+13 378216
## - yearsatcompany:title4             1 9.7738e+09 1.4634e+13 378219
## - yearsofexperience:educ3           2 1.3054e+10 1.4637e+13 378221
## - yearsatcompany:educ3              2 1.6008e+10 1.4640e+13 378225
## - yearsofexperience:yearsatcompany  1 1.6254e+10 1.4640e+13 378227
## - yearsofexperience:country         3 3.4587e+10 1.4659e+13 378246
## - yearsofexperience:Fortune_500     1 4.1294e+10 1.4665e+13 378259
## - yearsatcompany:Fortune_500        1 7.4684e+10 1.4699e+13 378301
## - sqrt(yearsofexperience)           1 3.1916e+11 1.4943e+13 378605
##
## Step:  AIC=378207
## basesalary ~ yearsofexperience + sqrt(yearsofexperience) + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500 +
```

```
##     yearsofexperience:country + educ3:country + yearsofexperience:educ3 +
##     yearsofexperience:yearsatcompany + yearsofexperience:title4 +
##     yearsofexperience:Fortune_500 + educ3:title4 + yearsatcompany:educ3 +
##     yearsatcompany:title4 + gender:title4 + yearsatcompany:Fortune_500 +
##     gender:Race2
##
##                                    Df  Sum of Sq        RSS    AIC
## - gender:Race2                      1 4.6786e+08 1.4626e+13 378206
## - educ3:title4                      2 2.8131e+09 1.4629e+13 378207
## <none>                                           1.4626e+13 378207
## - yearsofexperience:title4          1 1.9421e+09 1.4628e+13 378207
## - gender:title4                     1 2.7422e+09 1.4629e+13 378208
## + Race2:educ3                       2 1.8302e+09 1.4624e+13 378209
## + educ3:Fortune_500                 2 1.3960e+09 1.4625e+13 378209
## + gender:educ3                      2 1.3851e+08 1.4626e+13 378211
## - educ3:country                     6 1.5635e+10 1.4642e+13 378215
## - yearsatcompany:title4             1 9.6651e+09 1.4636e+13 378217
## - yearsofexperience:educ3           2 1.2484e+10 1.4639e+13 378219
## - yearsatcompany:educ3              2 1.5722e+10 1.4642e+13 378223
## - yearsofexperience:yearsatcompany  1 1.6285e+10 1.4642e+13 378226
## - yearsofexperience:country         3 3.4598e+10 1.4661e+13 378245
## - yearsofexperience:Fortune_500     1 4.1373e+10 1.4667e+13 378257
## - yearsatcompany:Fortune_500        1 7.4622e+10 1.4701e+13 378299
## - sqrt(yearsofexperience)           1 3.1870e+11 1.4945e+13 378603
##
## Step:  AIC=378205.6
## basesalary ~ yearsofexperience + sqrt(yearsofexperience) + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500 +
##     yearsofexperience:country + educ3:country + yearsofexperience:educ3 +
##     yearsofexperience:yearsatcompany + yearsofexperience:title4 +
##     yearsofexperience:Fortune_500 + educ3:title4 + yearsatcompany:educ3 +
##     yearsatcompany:title4 + gender:title4 + yearsatcompany:Fortune_500
##
##                                    Df  Sum of Sq        RSS    AIC
## - educ3:title4                      2 2.7604e+09 1.4629e+13 378205
## <none>                                           1.4626e+13 378206
## - yearsofexperience:title4          1 1.9590e+09 1.4628e+13 378206
## + gender:Race2                      1 4.6786e+08 1.4626e+13 378207
## + Race2:educ3                       2 1.8435e+09 1.4625e+13 378207
## - gender:title4                     1 2.9959e+09 1.4629e+13 378207
## + educ3:Fortune_500                 2 1.3885e+09 1.4625e+13 378208
## + gender:educ3                      2 2.4615e+08 1.4626e+13 378209
## - educ3:country                     6 1.5652e+10 1.4642e+13 378213
## - yearsatcompany:title4             1 9.5828e+09 1.4636e+13 378216
## - yearsofexperience:educ3           2 1.2439e+10 1.4639e+13 378217
## - yearsatcompany:educ3              2 1.5699e+10 1.4642e+13 378221
## - yearsofexperience:yearsatcompany  1 1.6317e+10 1.4643e+13 378224
## - Race2                             1 2.7932e+10 1.4654e+13 378239
## - yearsofexperience:country         3 3.4693e+10 1.4661e+13 378243
## - yearsofexperience:Fortune_500     1 4.1370e+10 1.4668e+13 378256
## - yearsatcompany:Fortune_500        1 7.4659e+10 1.4701e+13 378298
## - sqrt(yearsofexperience)           1 3.1832e+11 1.4945e+13 378601
##
## Step:  AIC=378205
```

```
## basesalary ~ yearsofexperience + sqrt(yearsofexperience) + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500 +
##     yearsofexperience:country + educ3:country + yearsofexperience:educ3 +
##     yearsofexperience:yearsatcompany + yearsofexperience:title4 +
##     yearsofexperience:Fortune_500 + yearsatcompany:educ3 + yearsatcompany:title4 +
##     gender:title4 + yearsatcompany:Fortune_500
##
##                                    Df  Sum of Sq        RSS    AIC
## <none>                                              1.4629e+13 378205
## + educ3:title4                      2 2.7604e+09 1.4626e+13 378206
## - yearsofexperience:title4          1 2.3352e+09 1.4632e+13 378206
## + gender:Race2                      1 4.1512e+08 1.4629e+13 378207
## - gender:title4                     1 2.8908e+09 1.4632e+13 378207
## + Race2:educ3                       2 1.7881e+09 1.4627e+13 378207
## + educ3:Fortune_500                 2 1.5377e+09 1.4628e+13 378207
## + gender:educ3                      2 1.6046e+08 1.4629e+13 378209
## - educ3:country                     6 1.5547e+10 1.4645e+13 378213
## - yearsatcompany:title4             1 9.9890e+09 1.4639e+13 378216
## - yearsofexperience:educ3           2 1.4375e+10 1.4644e+13 378219
## - yearsatcompany:educ3              2 1.5549e+10 1.4645e+13 378221
## - yearsofexperience:yearsatcompany  1 1.6242e+10 1.4645e+13 378224
## - Race2                             1 2.8145e+10 1.4657e+13 378239
## - yearsofexperience:country         3 3.4441e+10 1.4664e+13 378242
## - yearsofexperience:Fortune_500     1 4.1562e+10 1.4671e+13 378255
## - yearsatcompany:Fortune_500        1 7.5012e+10 1.4704e+13 378297
## - sqrt(yearsofexperience)           1 3.1700e+11 1.4946e+13 378599

##
## Call:
## lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience) +
##     yearsatcompany + gender + Race2 + educ3 + country + title4 +
##     Fortune_500 + yearsofexperience:country + educ3:country +
##     yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
##     yearsofexperience:title4 + yearsofexperience:Fortune_500 +
##     yearsatcompany:educ3 + yearsatcompany:title4 + gender:title4 +
##     yearsatcompany:Fortune_500, data = STEM3)
##
## Coefficients:
##                                (Intercept)
##                                   67906.59
##                           yearsofexperience
##                                    1024.80
##                      sqrt(yearsofexperience)
##                                   14035.89
##                                yearsatcompany
##                                    -3567.82
##                                genderFemale
##                                    -2642.68
##                                 Race2White
##                                    -2836.14
##                         educ3Master's Degree
##                                    5731.92
##                                    educ3PhD
##                                   13571.29
```

```
##                                countryIndia
##                                   -62082.07
##                         countryUnited Kingdom
##                                    -4520.74
##                                   countryUS
##                                    37273.15
##                              title4Management
##                                     4498.61
##                               Fortune_500Yes
##                                     -677.69
##              yearsofexperience:countryIndia
##                                      397.55
##    yearsofexperience:countryUnited Kingdom
##                                     1500.80
##                 yearsofexperience:countryUS
##                                     1039.19
##           educ3Master's Degree:countryIndia
##                                    -1578.04
##                         educ3PhD:countryIndia
##                                    18374.78
## educ3Master's Degree:countryUnited Kingdom
##                                     2810.50
##              educ3PhD:countryUnited Kingdom
##                                     3292.56
##              educ3Master's Degree:countryUS
##                                     1956.40
##                            educ3PhD:countryUS
##                                    20546.02
##     yearsofexperience:educ3Master's Degree
##                                      -79.58
##                  yearsofexperience:educ3PhD
##                                     -910.12
##            yearsofexperience:yearsatcompany
##                                       45.68
##          yearsofexperience:title4Management
##                                     -183.75
##            yearsofexperience:Fortune_500Yes
##                                     -661.27
##        yearsatcompany:educ3Master's Degree
##                                     -434.52
##                      yearsatcompany:educ3PhD
##                                     1091.94
##             yearsatcompany:title4Management
##                                      589.61
##               genderFemale:title4Management
##                                    -2608.13
##               yearsatcompany:Fortune_500Yes
##                                     1652.89
```

```
# Run the result of the stepwise regression
summary(lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience) +
    yearsatcompany + gender + Race2 + educ3 + country + title4 +
    Fortune_500 + yearsofexperience:country + educ3:country +
    yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
```

```
    yearsofexperience:title4 + yearsofexperience:Fortune_500 +
    yearsatcompany:educ3 + yearsatcompany:title4 + gender:title4 +
    yearsatcompany:Fortune_500, data = STEM3))
```

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience) +
##     yearsatcompany + gender + Race2 + educ3 + country + title4 +
##     Fortune_500 + yearsofexperience:country + educ3:country +
##     yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
##     yearsofexperience:title4 + yearsofexperience:Fortune_500 +
##     yearsatcompany:educ3 + yearsatcompany:title4 + gender:title4 +
##     yearsatcompany:Fortune_500, data = STEM3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83235 -19169   -191  17656  96967
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              67906.59    1821.83  37.274
## yearsofexperience                         1024.80     262.80   3.900
## sqrt(yearsofexperience)                  14035.89     702.51  19.980
## yearsatcompany                           -3567.82     199.64 -17.871
## genderFemale                             -2642.68     613.44  -4.308
## Race2White                               -2836.14     476.40  -5.953
## educ3Master's Degree                      5731.92    2358.34   2.430
## educ3PhD                                 13571.29    7698.59   1.763
## countryIndia                            -62082.07    2084.09 -29.789
## countryUnited Kingdom                    -4520.74    2909.37  -1.554
## countryUS                                37273.15    1727.03  21.582
## title4Management                          4498.61    1207.81   3.725
## Fortune_500Yes                            -677.69     685.48  -0.989
## yearsofexperience:countryIndia             397.55     255.24   1.558
## yearsofexperience:countryUnited Kingdom   1500.80     328.50   4.569
## yearsofexperience:countryUS               1039.19     205.51   5.057
## educ3Master's Degree:countryIndia        -1578.04    2873.33  -0.549
## educ3PhD:countryIndia                    18374.78   13841.90   1.327
## educ3Master's Degree:countryUnited Kingdom  2810.50    3487.07   0.806
## educ3PhD:countryUnited Kingdom            3292.56    9721.92   0.339
## educ3Master's Degree:countryUS            1956.40    2339.69   0.836
## educ3PhD:countryUS                       20546.02    7716.37   2.663
## yearsofexperience:educ3Master's Degree     -79.58      90.47  -0.880
## yearsofexperience:educ3PhD                -910.12     213.94  -4.254
## yearsofexperience:yearsatcompany            45.68      10.10   4.523
## yearsofexperience:title4Management        -183.75     107.15  -1.715
## yearsofexperience:Fortune_500Yes          -661.27      91.41  -7.234
## yearsatcompany:educ3Master's Degree       -434.52     155.40  -2.796
## yearsatcompany:educ3PhD                   1091.94     389.23   2.805
## yearsatcompany:title4Management            589.61     166.24   3.547
## genderFemale:title4Management            -2608.13    1366.98  -1.908
## yearsatcompany:Fortune_500Yes             1652.89     170.07   9.719
##                                         Pr(>|t|)
```

```
## (Intercept)                                    < 2e-16 ***
## yearsofexperience                             9.67e-05 ***
## sqrt(yearsofexperience)                        < 2e-16 ***
## yearsatcompany                                 < 2e-16 ***
## genderFemale                                  1.66e-05 ***
## Race2White                                    2.68e-09 ***
## educ3Master's Degree                          0.015088 *
## educ3PhD                                      0.077946 .
## countryIndia                                   < 2e-16 ***
## countryUnited Kingdom                         0.120237
## countryUS                                      < 2e-16 ***
## title4Management                              0.000196 ***
## Fortune_500Yes                                0.322854
## yearsofexperience:countryIndia                0.119347
## yearsofexperience:countryUnited Kingdom       4.94e-06 ***
## yearsofexperience:countryUS                   4.31e-07 ***
## educ3Master's Degree:countryIndia             0.582874
## educ3PhD:countryIndia                         0.184368
## educ3Master's Degree:countryUnited Kingdom    0.420266
## educ3PhD:countryUnited Kingdom                0.734860
## educ3Master's Degree:countryUS                0.403064
## educ3PhD:countryUS                            0.007759 **
## yearsofexperience:educ3Master's Degree        0.379044
## yearsofexperience:educ3PhD                    2.11e-05 ***
## yearsofexperience:yearsatcompany              6.15e-06 ***
## yearsofexperience:title4Management            0.086395 .
## yearsofexperience:Fortune_500Yes              4.86e-13 ***
## yearsatcompany:educ3Master's Degree           0.005177 **
## yearsatcompany:educ3PhD                       0.005031 **
## yearsatcompany:title4Management               0.000391 ***
## genderFemale:title4Management                 0.056414 .
## yearsatcompany:Fortune_500Yes                  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28180 on 18422 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.656
## F-statistic:  1136 on 31 and 18422 DF,  p-value: < 2.2e-16
```

e) All-Possible-Regressions Selection

We then used the All-Possible-Regressions Selection procedure to identify the "best" subset of variables. With 11 variables, there are 2,047 possible subsets of first-order models, representing different combinations of predictors.

```r
# Create new binary columns for each level of educ3
STEM3$Bachelor <- ifelse(STEM3$educ3 == "College or below", 1, 0)
STEM3$Masters <- ifelse(STEM3$educ3 == "Master's Degree", 1, 0)
STEM3$PhD <- ifelse(STEM3$educ3 == "PhD", 1, 0)
```

```r
# Create new binary columns for each level of country
STEM3$Canada <- ifelse(STEM3$country == "Canada", 1, 0)
STEM3$India <- ifelse(STEM3$country == "India", 1, 0)
```

```r
STEM3$UK <- ifelse(STEM3$country == "United Kingdom", 1, 0)
STEM3$US <- ifelse(STEM3$country == "US", 1, 0)
```

Show the possible models only for the "best" model for each value of p The table on screen shows the "best" model for each number of predictors (p). For example, the best one-variable model includes the country India, while the best two-variable model includes India and years of experience, and so on.

```r
Model_all_2=regsubsets(basesalary~yearsofexperience+yearsatcompany+gender+Race2+
                         Fortune_500+title4+Masters+PhD+India+UK+US,data=STEM3, nvmax = 11)
summary(Model_all_2)
```

```
## Subset selection object
## Call: regsubsets.formula(basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + Fortune_500 + title4 + Masters + PhD + India +
##     UK + US, data = STEM3, nvmax = 11)
## 11 Variables  (and intercept)
##                    Forced in Forced out
## yearsofexperience      FALSE      FALSE
## yearsatcompany         FALSE      FALSE
## genderFemale           FALSE      FALSE
## Race2White             FALSE      FALSE
## Fortune_500Yes         FALSE      FALSE
## title4Management       FALSE      FALSE
## Masters                FALSE      FALSE
## PhD                    FALSE      FALSE
## India                  FALSE      FALSE
## UK                     FALSE      FALSE
## US                     FALSE      FALSE
## 1 subsets of each size up to 11
## Selection Algorithm: exhaustive
##           yearsofexperience yearsatcompany genderFemale Race2White
## 1  ( 1 )  " "               " "            " "          " "
## 2  ( 1 )  "*"               " "            " "          " "
## 3  ( 1 )  "*"               " "            " "          " "
## 4  ( 1 )  "*"               " "            " "          " "
## 5  ( 1 )  "*"               "*"            " "          " "
## 6  ( 1 )  "*"               "*"            " "          " "
## 7  ( 1 )  "*"               "*"            " "          " "
## 8  ( 1 )  "*"               "*"            "*"          " "
## 9  ( 1 )  "*"               "*"            "*"          "*"
## 10  ( 1 ) "*"               "*"            "*"          "*"
## 11  ( 1 ) "*"               "*"            "*"          "*"
##           Fortune_500Yes title4Management Masters PhD India UK  US
## 1  ( 1 )  " "            " "              " "     " " "*"   " " " "
## 2  ( 1 )  " "            " "              " "     " " "*"   " " " "
## 3  ( 1 )  " "            " "              " "     " " "*"   " " "*"
## 4  ( 1 )  " "            " "              " "     "*" "*"   " " "*"
## 5  ( 1 )  " "            " "              " "     "*" "*"   " " "*"
## 6  ( 1 )  " "            " "              "*"     "*" "*"   " " "*"
## 7  ( 1 )  " "            "*"              "*"     "*" "*"   " " "*"
## 8  ( 1 )  " "            "*"              "*"     "*" "*"   " " "*"
## 9  ( 1 )  " "            "*"              "*"     "*" "*"   " " "*"
```

```
## 10  ( 1 ) " "            "*"              "*"     "*" "*"   "*" "*"
## 11  ( 1 ) "*"            "*"              "*"     "*" "*"   "*" "*"
```

Summarize the "best subset" models along with their criteria Next, we summarized the best subset models along with their criteria such as the R-squared, MSE, CP, PRESS.

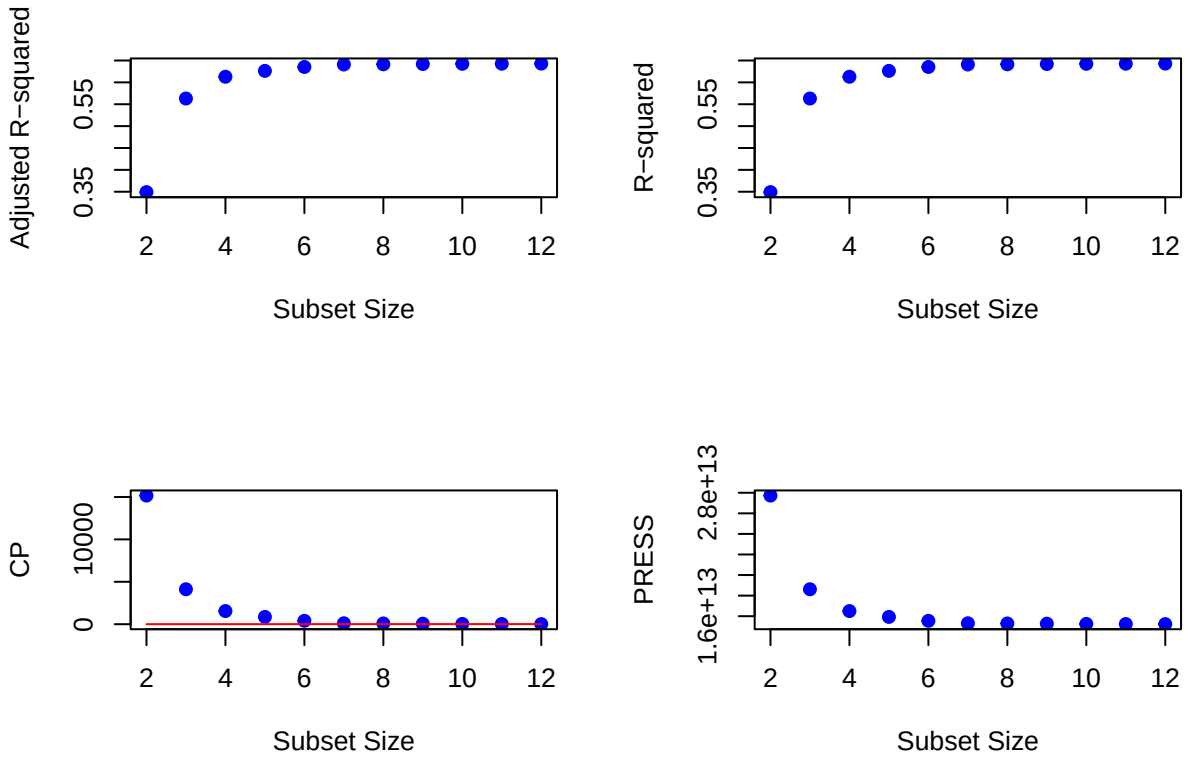
In the table, we can see that the 11th subset maximizes the R-squared and adjusted R-squared while also minimizing the rest of the criteria. Plotting these quantities against the number of predictors, shows that all variables should be included in the group of the most important predictors.

```
# Result
cbind(SUM$which,round(cbind(Rsq,AdRsq,CP,BIC,RSS,AIC,PRESS,MSE,MSE1),4))
```

```
##    (Intercept) yearsofexperience yearsatcompany genderFemale Race2White
## 1            1                 0              0            0          0
## 2            1                 1              0            0          0
## 3            1                 1              0            0          0
## 4            1                 1              0            0          0
## 5            1                 1              1            0          0
## 6            1                 1              1            0          0
## 7            1                 1              1            0          0
## 8            1                 1              1            1          0
## 9            1                 1              1            1          1
## 10           1                 1              1            1          1
## 11           1                 1              1            1          1
##    Fortune_500Yes title4Management Masters PhD India UK US    Rsq  AdRsq
## 1               0                0       0   0     1  0  0 0.3492 0.3492
## 2               0                0       0   0     1  0  0 0.5631 0.5630
## 3               0                0       0   0     1  0  1 0.6128 0.6128
## 4               0                0       0   1     1  0  1 0.6263 0.6262
## 5               0                0       0   1     1  0  1 0.6352 0.6351
## 6               0                0       1   1     1  0  1 0.6408 0.6407
## 7               0                1       1   1     1  0  1 0.6414 0.6412
## 8               0                1       1   1     1  0  1 0.6419 0.6418
## 9               0                1       1   1     1  0  1 0.6425 0.6423
## 10              0                1       1   1     1  1  1 0.6427 0.6425
## 11              1                1       1   1     1  1  1 0.6429 0.6427
##            CP        BIC          RSS        AIC        PRESS        MSE
## 1  15162.2415  -7907.103 2.772151e+13  -7922.749 2.772586e+13 1502358218
## 2   4117.9777 -15250.220 1.861117e+13 -15273.689 1.861644e+13 1008680705
## 3   1550.2279 -17471.481 1.649178e+13 -17502.773 1.649804e+13  893863440
## 4    857.3962 -18114.306 1.591872e+13 -18153.421 1.592656e+13  862850118
## 5    399.1027 -18549.900 1.553910e+13 -18596.838 1.554915e+13  842318802
## 6    109.7603 -18827.663 1.529881e+13 -18882.424 1.531046e+13  829338874
## 7     85.1767 -18844.305 1.527689e+13 -18906.890 1.529038e+13  828195246
## 8     57.2388 -18864.332 1.525220e+13 -18934.740 1.526723e+13  826901512
## 9     32.2815 -18881.429 1.522997e+13 -18959.660 1.524680e+13  825740916
## 10    21.9541 -18883.929 1.521980e+13 -18969.983 1.523802e+13  825234427
## 11    12.0000 -18886.064 1.520994e+13 -18979.941 1.522991e+13  824744575
##          MSE1
## 1  1502358218
## 2  1008680705
## 3   893863440
```

```
## 4   862850118
## 5   842318802
## 6   829338874
## 7   828195246
## 8   826901512
## 9   825740916
## 10  825234427
## 11  824744575
```

```
par(mfrow=c(2,2))
plot(p,Rsq,xlab="Subset Size",ylab="Adjusted R-squared", pch=19, col="blue")
plot(p,Rsq,xlab="Subset Size",ylab="R-squared", pch=19, col="blue")
plot(p,CP,xlab="Subset Size",ylab="CP", pch=19, col="blue")
lines(y=p+1,x=p, col="red")
plot(p,PRESS,xlab="Subset Size",ylab="PRESS", pch=19, col="blue")
```



5. Model Fitting

Our last step is Model Fitting, where we used stratified random sampling per country to split the data into 80% training and 20% test.

Prepare 80% Training and 20% Test Data

```
# Function to split data into training and test sets for each country
split_data_by_country <- function(data, train_ratio = 0.8) {
  # Group by country and split within each group
  data_split <- data %>%
    group_by(country) %>%
    group_split() %>%
    lapply(function(group) {
      set.seed(123)  # Ensures reproducibility
      sample_index <- sample(1:nrow(group), size = floor(train_ratio * nrow(group)))
      list(
        train = group[sample_index, ],
        test = group[-sample_index, ]
      )
    })

  # Combine training and test sets from all countries
  train_data <- bind_rows(lapply(data_split, `[[`, "train"))
  test_data <- bind_rows(lapply(data_split, `[[`, "test"))

  return(list(train = train_data, test = test_data))
}

# Example usage
set.seed(123)  # For reproducibility
data_split <- split_data_by_country(STEM3, train_ratio = 0.8)

train_set <- data_split$train
test_data <- data_split$test

# Verify proportions
table(train_set$country) / table(STEM3$country)
```

```
##
##         Canada          India United Kingdom             US
##      0.7989418      0.8000000      0.7995992      0.7999873
```

```
table(test_data$country) / table(STEM3$country)
```

```
##
##         Canada          India United Kingdom             US
##      0.2010582      0.2000000      0.2004008      0.2000127
```

```
#set the base level
train_set$gender <- relevel(train_set$gender, ref = "Male")
train_set$Race2 <- relevel(train_set$Race2, ref = "Non-White")
train_set$educ3 <- relevel(train_set$educ3, ref = "College or below")
train_set$Fortune_500 <- relevel(train_set$Fortune_500, ref = "No")
train_set$title4 <- relevel(train_set$title4, ref = "Non-Management")
train_set$country <- relevel(train_set$country, ref = "Canada")
```

```r
test_data$gender <- relevel(test_data$gender, ref = "Male")
test_data$Race2 <- relevel(test_data$Race2, ref = "Non-White")
test_data$educ3 <- relevel(test_data$educ3, ref = "College or below")
test_data$Fortune_500 <- relevel(test_data$Fortune_500, ref = "No")
test_data$title4 <- relevel(test_data$title4, ref = "Non-Management")
test_data$country <- relevel(test_data$country, ref = "Canada")
```

```r
# Function to Calculate Performance Metrics
calculate_metrics <- function(model, test_data) {
  # Predicted values
  predicted <- predict(model, newdata = test_data)

  # Actual values
  actual <- test_data[[as.character(formula(model)[[2]])]]

  # Calculate metrics
  residuals <- actual - predicted
  mse <- mean(residuals^2)
  rmse <- sqrt(mse)
  mae <- mean(abs(residuals))
  r_squared <- summary(model)$r.squared

  return(data.frame(
    MSE = mse,
    RMSE = rmse,
    MAE = mae,
    R_Squared = r_squared
  ))
}
```

```r
# Function to Plot Predicted vs. Actual
plot_predictions <- function(model, test_data) {
  # Predicted values
  predicted <- predict(model, newdata = test_data)

  # Actual values
  actual <- test_data[[as.character(formula(model)[[2]])]]

  # Create scatter plot
  library(ggplot2)
  ggplot(data = data.frame(Actual = actual, Predicted = predicted), aes(x = Actual, y = Predicted)) +
    geom_point(color = "blue", alpha = 0.6) +
    geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Predicted vs. Actual Values",
         x = "Actual Values",
         y = "Predicted Values") +
    scale_x_continuous(labels = function(x) format(x, big.mark = ",", scientific = FALSE)) +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",", scientific = FALSE)) +
    theme_minimal()
}
```

Comparing the Models

We have two models for comparison: the **Main Effects model** and the **Main Effects model with Transformations and Interactions**.

When comparing the predicted vs. actual value graphs side by side, there is no noticeable difference between the two. Based on their adjusted R-squared values, incorporating transformations and interactions only led to a slight improvement in model fit, increasing the adjusted R-squared by just over 1%. Therefore, the simpler Main Effects model is the best, as it provides comparable performance with fewer complexities.

```
# Fit the model
model <- lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
    gender + Race2 + educ3 + country + title4 + Fortune_500,
    data = train_set)
summary(model)
```

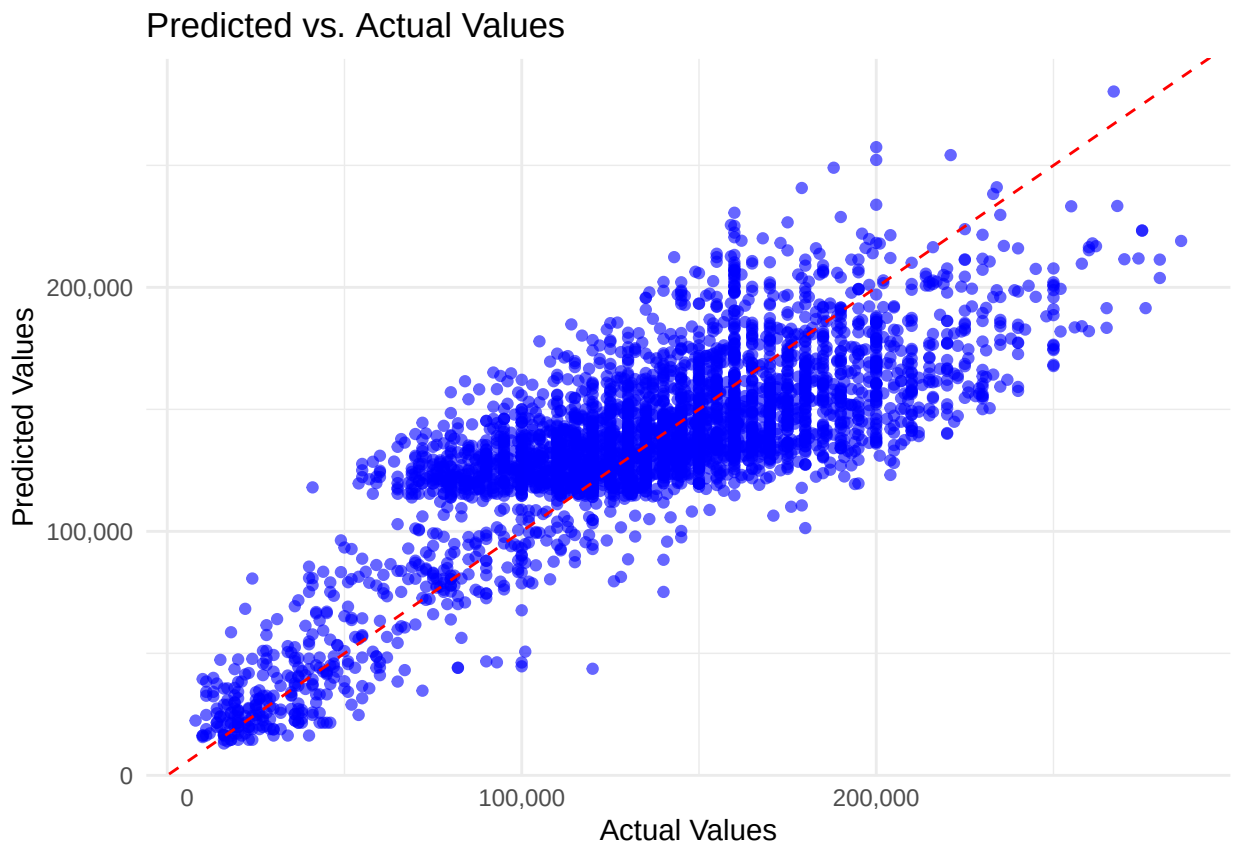
1) Main Effects (Simple) / All-Possible Regressions Model

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
##     gender + Race2 + educ3 + country + title4 + Fortune_500,
##     data = train_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78883 -19131   -233  17935  87113
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             75182.37    1239.89  60.636  < 2e-16 ***
## yearsofexperience        4216.43      52.22  80.749  < 2e-16 ***
## yearsatcompany          -1590.86      85.34 -18.640  < 2e-16 ***
## genderFemale            -3261.57     624.89  -5.219 1.82e-07 ***
## Race2White              -2479.52     537.52  -4.613 4.01e-06 ***
## educ3Master's Degree     7141.82     510.64  13.986  < 2e-16 ***
## educ3PhD                30619.74    1160.41  26.387  < 2e-16 ***
## countryIndia           -58863.63    1443.55 -40.777  < 2e-16 ***
## countryUnited Kingdom    5487.59    1853.50   2.961 0.003075 **
## countryUS               44442.25    1199.36  37.055  < 2e-16 ***
## title4Management         3611.23     679.90   5.311 1.10e-07 ***
## Fortune_500Yes          -1748.43     490.60  -3.564 0.000367 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28640 on 14750 degrees of freedom
## Multiple R-squared:  0.6446, Adjusted R-squared:  0.6444
## F-statistic:  2432 on 11 and 14750 DF,  p-value: < 2.2e-16
```

```
# Calculate metrics
metrics <- calculate_metrics(model, test_data)
print(metrics)
```

```
##         MSE     RMSE      MAE R_Squared
## 1 843204523 29037.98 22846.76 0.6446266
```

```r
# Plot predictions
plot_predictions(model, test_data)
```




```r
# Fit the model
model1 <- lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience) +
    yearsatcompany + gender + Race2 + educ3 + country + title4 +
    Fortune_500 + yearsofexperience:country + educ3:country +
    yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
    yearsofexperience:title4 + yearsofexperience:Fortune_500 +
    yearsatcompany:educ3 + yearsatcompany:title4 + gender:title4 +
    yearsatcompany:Fortune_500, data = train_set)
summary(model1)
```

2) Stepwise Model (with Interaction + Transformation)

```
##
## Call:
## lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience) +
##     yearsatcompany + gender + Race2 + educ3 + country + title4 +
```

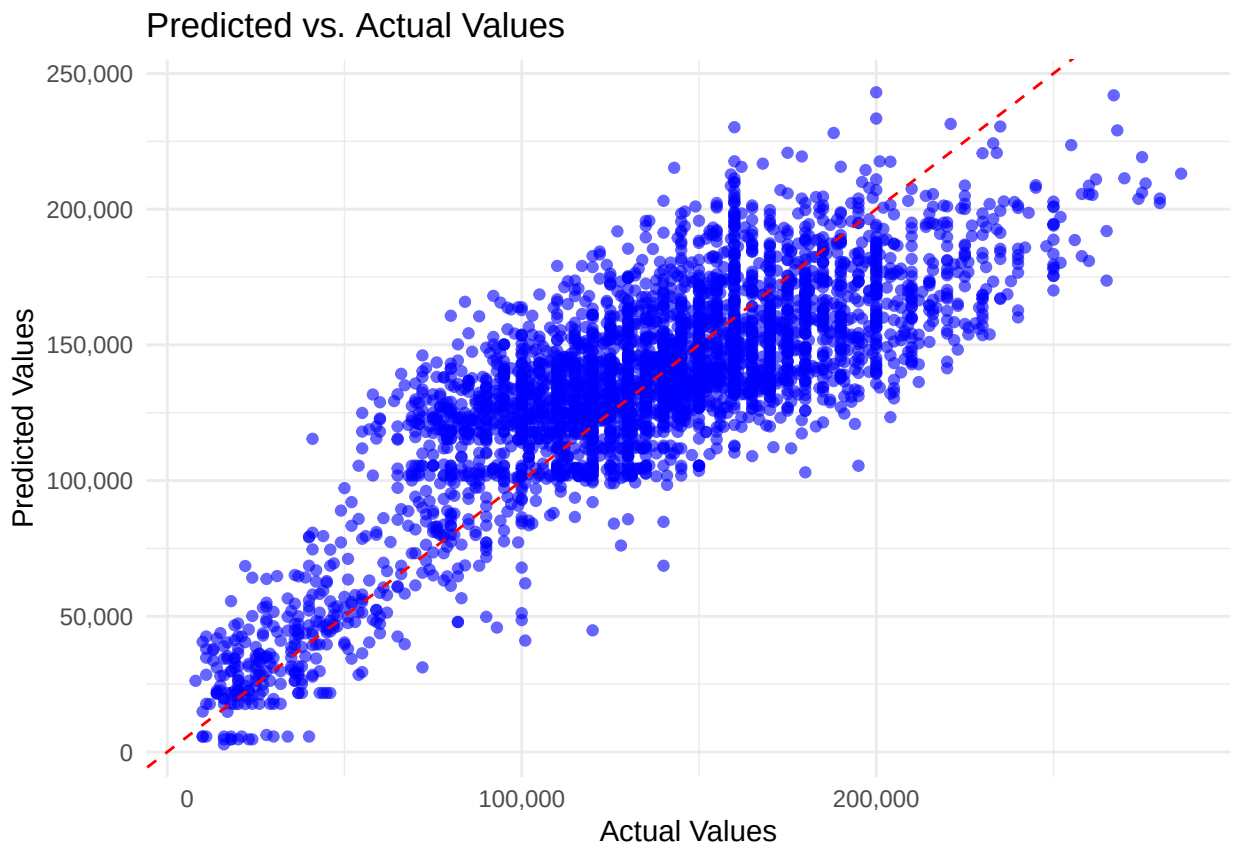
```
##     Fortune_500 + yearsofexperience:country + educ3:country +
##     yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
##     yearsofexperience:title4 + yearsofexperience:Fortune_500 +
##     yearsatcompany:educ3 + yearsatcompany:title4 + gender:title4 +
##     yearsatcompany:Fortune_500, data = train_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82114 -19214   -227  17592  95637
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             68624.642   2051.711  33.448
## yearsofexperience                         846.580    298.205   2.839
## sqrt(yearsofexperience)                 13961.358    780.857  17.880
## yearsatcompany                          -3439.553    221.042 -15.561
## genderFemale                            -2859.999    686.318  -4.167
## Race2White                              -2569.952    531.136  -4.839
## educ3Master's Degree                     5640.456   2607.301   2.163
## educ3PhD                                12597.151   8662.497   1.454
## countryIndia                           -62928.627   2345.916 -26.825
## countryUnited Kingdom                   -4404.649   3249.079  -1.356
## countryUS                               36843.763   1948.599  18.908
## title4Management                         4368.340   1354.980   3.224
## Fortune_500Yes                          -1029.803    765.488  -1.345
## yearsofexperience:countryIndia            655.519    290.682   2.255
## yearsofexperience:countryUnited Kingdom  1582.211    369.664   4.280
## yearsofexperience:countryUS              1166.895    237.925   4.904
## educ3Master's Degree:countryIndia       -2215.693   3176.548  -0.698
## educ3PhD:countryIndia                   19732.923  14376.915   1.373
## educ3Master's Degree:countryUnited Kingdom  1321.971   3865.937   0.342
## educ3PhD:countryUnited Kingdom           5241.508  10928.154   0.480
## educ3Master's Degree:countryUS           1784.586   2590.372   0.689
## educ3PhD:countryUS                      22163.118   8679.694   2.553
## yearsofexperience:educ3Master's Degree     -0.167    101.608  -0.002
## yearsofexperience:educ3PhD              -1066.213    235.177  -4.534
## yearsofexperience:yearsatcompany           43.113     11.205   3.847
## yearsofexperience:title4Management       -185.549    120.420  -1.541
## yearsofexperience:Fortune_500Yes         -603.206    102.245  -5.900
## yearsatcompany:educ3Master's Degree      -457.032    172.280  -2.653
## yearsatcompany:educ3PhD                  1284.804    439.962   2.920
## yearsatcompany:title4Management           601.130    184.211   3.263
## genderFemale:title4Management            -2350.592   1531.281  -1.535
## yearsatcompany:Fortune_500Yes             1538.696    189.154   8.135
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## yearsofexperience                        0.00453 **
## sqrt(yearsofexperience)                  < 2e-16 ***
## yearsatcompany                           < 2e-16 ***
## genderFemale                            3.10e-05 ***
## Race2White                              1.32e-06 ***
## educ3Master's Degree                     0.03053 *
## educ3PhD                                 0.14591
## countryIndia                             < 2e-16 ***
```

```
## countryUnited Kingdom                        0.17523
## countryUS                                    < 2e-16 ***
## title4Management                             0.00127 **
## Fortune_500Yes                               0.17855
## yearsofexperience:countryIndia               0.02414 *
## yearsofexperience:countryUnited Kingdom    1.88e-05 ***
## yearsofexperience:countryUS                9.47e-07 ***
## educ3Master's Degree:countryIndia            0.48549
## educ3PhD:countryIndia                        0.16992
## educ3Master's Degree:countryUnited Kingdom  0.73239
## educ3PhD:countryUnited Kingdom               0.63150
## educ3Master's Degree:countryUS               0.49088
## educ3PhD:countryUS                           0.01068 *
## yearsofexperience:educ3Master's Degree       0.99869
## yearsofexperience:educ3PhD                 5.84e-06 ***
## yearsofexperience:yearsatcompany             0.00012 ***
## yearsofexperience:title4Management           0.12337
## yearsofexperience:Fortune_500Yes           3.72e-09 ***
## yearsatcompany:educ3Master's Degree          0.00799 **
## yearsatcompany:educ3PhD                      0.00350 **
## yearsatcompany:title4Management              0.00110 **
## genderFemale:title4Management                0.12479
## yearsatcompany:Fortune_500Yes              4.46e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28110 on 14730 degrees of freedom
## Multiple R-squared:  0.658,  Adjusted R-squared:  0.6573
## F-statistic: 914.2 on 31 and 14730 DF,  p-value: < 2.2e-16
```

```r
# Calculate metrics
metrics <- calculate_metrics(model1, test_data)
print(metrics)
```

```
##         MSE     RMSE      MAE R_Squared
## 1 810020417 28460.86 22530.21 0.6579944
```

```r
# Plot predictions
plot_predictions(model1, test_data)
```



Predicting the salary of a new-graduate in Canada

We also attempted to use the prediction function to estimate the salary of a male, non-White, college graduate in Canada who landed a non-managerial role at a start-up company (non-Fortune 500). The results show that we are 95% confident that a person with these characteristics will earn between $72,752 and $77,612, with a prediction interval ranging from $18,994 to $131,370.

```
New <- data.frame(yearsofexperience = 0, yearsatcompany = 0, gender = "Male",
                  Race2 = "Non-White", educ3 = "College or below",
                  country = "Canada", title4 = "Non-Management",
                  Fortune_500 = "No")
predict(model, New, interval = "confidence", level = 0.95)
```

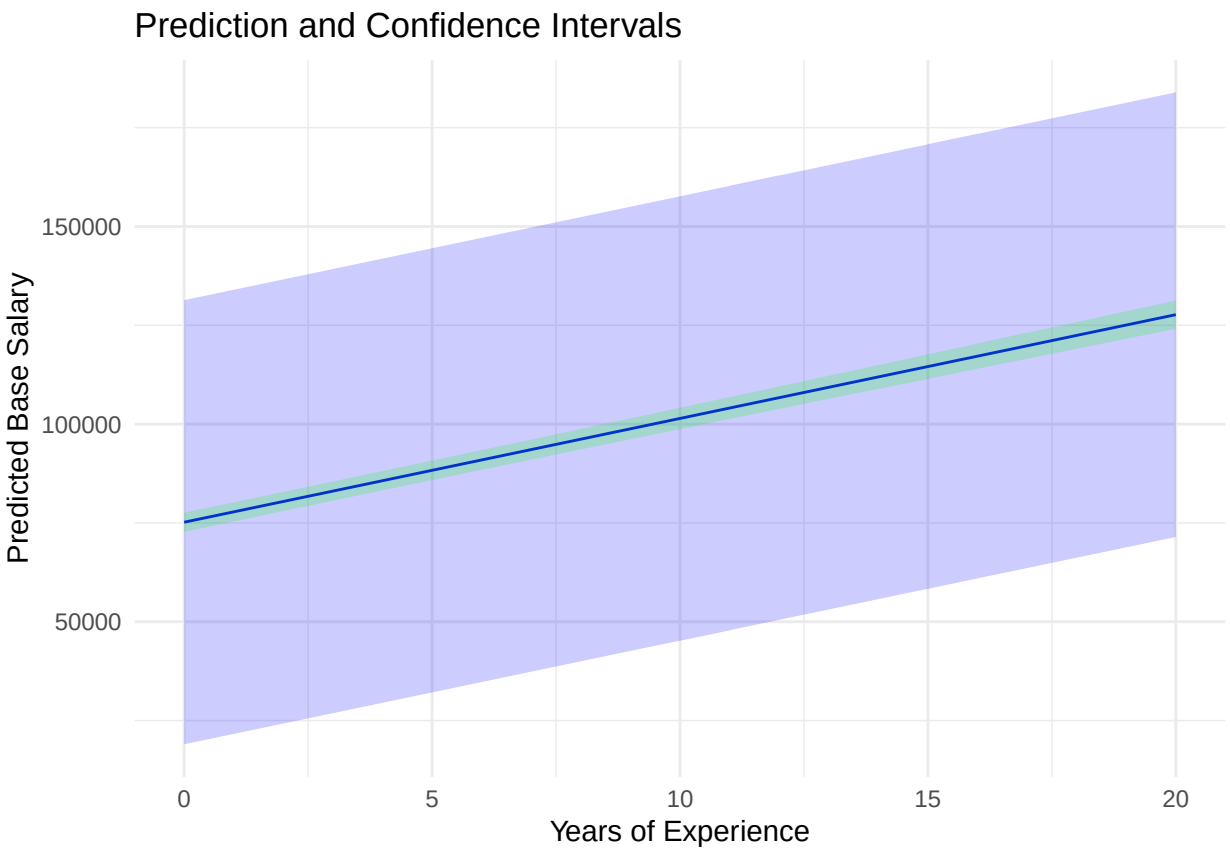
```
##        fit      lwr      upr
## 1 75182.37 72752.03 77612.72
```

```
predict(model, New, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 75182.37 18994.32 131370.4
```

```
# Load necessary library
library(ggplot2)
```

```r
# Generate new data for a range of `yearsofexperience`
new_data <- data.frame(
  yearsofexperience = seq(0, 20, by = 1),  # Example range of years
  yearsatcompany = seq(0, 20, by = 1),
  gender = "Male",
  Race2 = "Non-White",
  educ3 = "College or below",
  country = "Canada",
  title4 = "Non-Management",
  Fortune_500 = "No"
)

# Get predictions with intervals
predictions <- predict(model, new_data, interval = "prediction", level = 0.95)
confidence <- predict(model, new_data, interval = "confidence", level = 0.95)

# Combine predictions with new data
results <- cbind(new_data, predictions, confidence = confidence[, c("lwr", "upr")])

# Plot using ggplot2
ggplot(results, aes(x = yearsofexperience, y = fit)) +
  geom_line(color = "blue") +  # Predicted values
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) +  # Prediction interval
  geom_ribbon(aes(ymin = confidence.lwr, ymax = confidence.upr), fill = "green", alpha = 0.2) +  # Conf
  labs(
    title = "Prediction and Confidence Intervals",
    x = "Years of Experience",
    y = "Predicted Base Salary"
  ) +
  theme_minimal()
```



Best Model

In summary, we believe that the simpler model, which includes only the main effects term and all the predictors—such as years of experience, years at company, gender, race, country, education, Fortune 500 status, and title—provides a strong prediction of the base salary for a STEM employee residing in the US, UK, Canada, and India.

$$\begin{aligned}\text{BaseSalary} = \beta_0 &+ \beta_1 \cdot \text{YearsOfExperience} + \beta_2 \cdot \text{YearsAtCompany} + \beta_3 \cdot \text{GenderFemale} \\ &+ \beta_4 \cdot \text{RaceWhite} + \beta_5 \cdot \text{EducMaster'sDegree} + \beta_6 \cdot \text{EducPhD} \\ &+ \beta_7 \cdot \text{CountryIndia} + \beta_8 \cdot \text{CountryUK} + \beta_9 \cdot \text{CountryUS} \\ &+ \beta_{10} \cdot \text{TitleManagement} + \beta_{11} \cdot \text{Fortune500Yes} + \epsilon\end{aligned}$$

Limitations and Future Improvements

Our model comes with several limitations and opportunities for improvement.

- **Other contributing factors unaccounted for:**
  - The current model excludes significant predictors of base salary, such as:
    - **Tech Stack:** Top-paying programming languages and frameworks
    - **Industry**: Sector-specific variations (e.g., FinTech, AI/ML)
    - **Company Profitability:** Impact of employer financial health

    - **Prestige of Previous Employers:** Influence of working for high-profile companies (e.g., FAANG)

- **Inestimable parameters:** Some parameter combinations cannot be estimated due to missing data or sparse representation, limiting the model's ability to generalize for these cases.
- **Insufficient data for certain countries:** The dataset is heavily skewed, with the US accounting for 90% of responses. This imbalance reduces the model's ability to accurately represent salary trends in less-represented countries like the UK, Canada, and India.
- **Income disparity and outliers:** Income disparity within the US is pronounced, with several extreme outliers significantly inflating salaries compared to other countries. This likely affects the model's performance and skews predictions.