Final Project

Yinbo Tang and Jonathan Hsu

2024-05-07

Read Data and Data Preprocessing

```r
job_placement_raw = read.csv("job_placement.csv")

# remove those who didn't find job.
salary_noRanking = job_placement_raw[job_placement_raw$salary > 0,]

#Change the university column name
names(salary_noRanking)[names(salary_noRanking) == "college_name"] = "University.Name"

# replace "--" by "-"
salary_noRanking$University.Name <- gsub("--", "-", salary_noRanking$University.Name)
# I found there is a wrong "-" between Urbana and Champiagn so I remove it here.
salary_noRanking$University.Name <- gsub("Urbana-Champaign", "Urbana Champaign",
                                         salary_noRanking$University.Name)


#import university ranking dataset
ranking = read.csv("US-News-Rankings-Universities-Through-2023.csv")

# Calculate the average ranking in past 10 years so we get a more reliable ranking
ranking$avg_10 = rowMeans(ranking[,c("X2023", "X2022", "X2021", "X2020",
                                     "X2019", "X2018", "X2017", "X2016", "X2015",
                                     "X2014")],
                          na.rm = TRUE)

# create a new dataset which only contains university name and rankign in past 10 years.
ranking_subset = ranking[, c("University.Name", "avg_10")]

#adding average ranking into salary dataset by merging function
salary_full = merge(salary_noRanking, ranking_subset, by = "University.Name")

salary_final = salary_full[, c("gpa", "age", "gender",
                               "years_of_experience", "avg_10", "salary")]

head(salary_final)
```

```
##   gpa age gender years_of_experience avg_10 salary
## 1 3.8  26   Male                   3   10.3  59000
## 2 3.7  25 Female                   1    5.1  63000
## 3 3.9  24   Male                   2    8.9  61000
## 4 3.7  23   Male                   2   21.6  67000
## 5 3.7  25   Male                   2    2.1  60000
## 6 3.6  24 Female                   1    4.7  65000
```

```
dim(salary_final)
```

```
## [1] 450   6
```

Part 1

Part 1.1 Fit First Model

```
first_model = lm(salary ~ gpa + age + years_of_experience + gender + avg_10,
                 data = salary_final)
model = lm(salary ~ gpa + years_of_experience + gender + avg_10, data = salary_final)
summary(first_model)
```

```
##
## Call:
## lm(formula = salary ~ gpa + age + years_of_experience + gender +
##     avg_10, data = salary_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6055.9  -757.6   589.4  1545.5  5240.5
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          26373.90    4760.49   5.540 5.19e-08 ***
## gpa                  10450.88    1294.22   8.075 6.43e-15 ***
## age                   -142.65      78.81  -1.810  0.07096 .
## years_of_experience    685.68     196.43   3.491  0.00053 ***
## genderMale            -467.30     188.99  -2.473  0.01379 *
## avg_10                  17.20       3.75   4.587 5.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1980 on 443 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3443, Adjusted R-squared:  0.3369
## F-statistic: 46.53 on 5 and 443 DF,  p-value: < 2.2e-16
```

```
summary(model) #Removing age cannot improve R square
```

```
##
## Call:
## lm(formula = salary ~ gpa + years_of_experience + gender + avg_10,
##     data = salary_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6296.9  -915.3   426.4  1419.5  5151.1
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          24270.171   4628.258   5.244 2.44e-07 ***
## gpa                  10054.343   1278.806   7.862 2.89e-14 ***
## years_of_experience    728.737    195.488   3.728 0.000218 ***
## genderMale            -453.487    189.319  -2.395 0.017018 *
## avg_10                  17.167      3.759   4.566 6.44e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1985 on 444 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3395, Adjusted R-squared:  0.3335
## F-statistic: 57.04 on 4 and 444 DF,  p-value: < 2.2e-16
```

a. Interpretation for coefficients of age: The coefficient of -142.65 for age suggests a slightly negative impact of age on salary. This finding implies that as individuals get older, their expected salary decreases by approximately \$142.65 annually, assuming all other variables such as GPA, years of experience, gender, and university ranking remain constant. This effect, while statistically marginal ($p < 0.1$), could indicate that younger individuals might be entering higher-paying industries or roles, or that older individuals might not be receiving proportional increases in salary with age. Alternatively, this trend could reflect the changing dynamics in job sectors where younger skill sets are more in demand.

b. Interpretation for Gender: The gender coefficient of -467.30 for males, which is statistically significant ($p < 0.05$), indicates that males are expected to earn \$467.30 less than their female counterparts, when other factors are held constant. This unexpected result could suggest a reverse gender pay gap in the specific dataset or sectors analyzed. It would be important to delve deeper into the contextual factors such as industry types, job roles, and geographical locations that might influence this trend. Additionally, examining company-specific policies on pay equality and promotions could provide further insights.

c. In regression models, each non-reference category of a categorical variable like gender adds one degree of freedom to the model. Since gender is coded with males as one category and females as the baseline, the presence of the male category adds one parameter to be estimated in the model, hence contributing 1 to the number of predictors (p) used in the analysis. This means that the model complexity increases slightly with the addition of the gender variable, which helps in understanding salary differentials across genders within the given dataset.

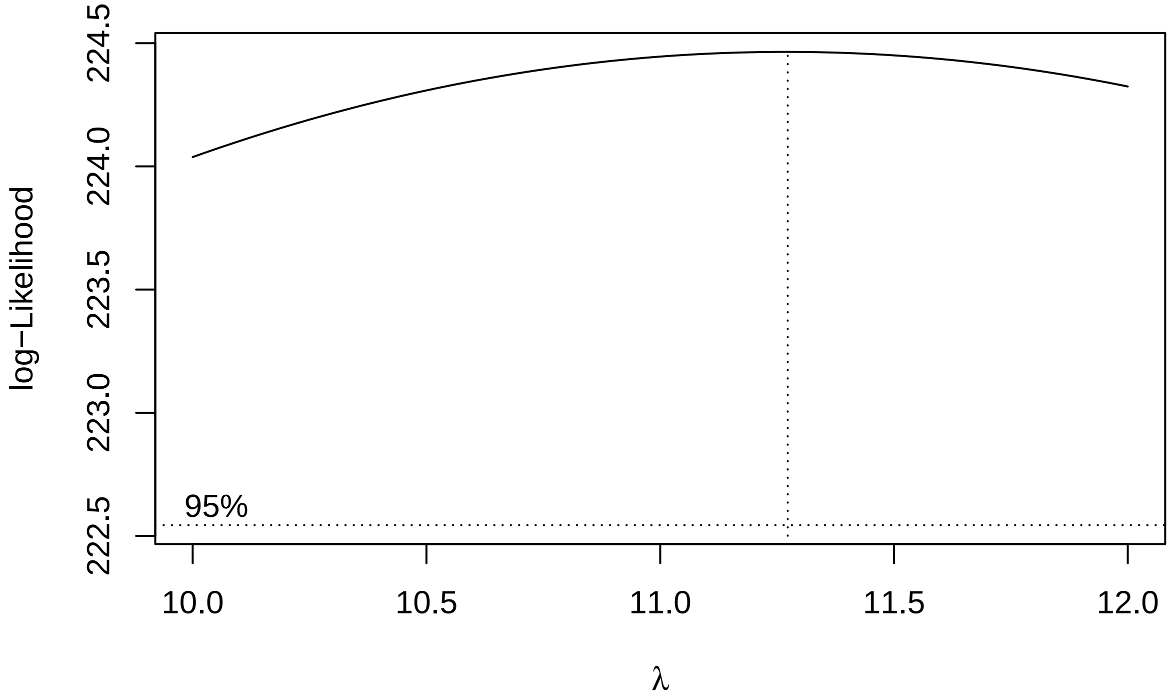
d. The choice of females as the baseline or reference category in the model means that all coefficients for gender are interpreted relative to female graduates. This methodological choice sets female graduates as the standard against which other categories (i.e., male graduates) are compared. This approach helps isolate the effect of being male on salary while controlling for other factors in the model.

Part 1.2 Box-Cox Transformation

```r
library("MASS")
```

```
## Warning: package 'MASS' was built under R version 4.2.3
```

```r
boxcox(first_model, plotit = TRUE, lambda = seq(10, 12, by = 0.1))
```



According to the Box-Cox plot transformation plot, the optimal λ suggested is around 11.3, which is the λ . So the recommended transformation is $y^{11.3}$ or $\frac{y^{11.3}-1}{11.3}$. Applying this transformation can enhance model fit by better meeting regression assumptions. It can also reduce the influence of outliers by adjusting scale disparities in the salary data. However, $y^{11.3}$ would be too large for my model. It may be time comsuming to fit a model with such scale of data. Therefore, I would keep my original modle for analysis.

Part 1.3 Fit Another Model

```
second_model = lm(salary ~ gpa + age + years_of_experience +
                    gender + avg_10 + gender * gpa, data = salary_final)
summary(second_model)
```

```
##
## Call:
## lm(formula = salary ~ gpa + age + years_of_experience + gender +
##     avg_10 + gender * gpa, data = salary_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6053.9  -742.4   585.5  1550.2  5270.9
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         25700.281   5580.516   4.605 5.39e-06 ***
## gpa                 10622.747   1492.467   7.118 4.47e-12 ***
## age                  -140.470     79.447  -1.768 0.077733 .
## years_of_experience   677.528    199.762   3.392 0.000757 ***
## genderMale            899.015   5892.624   0.153 0.878810
## avg_10                 17.086      3.785   4.514 8.18e-06 ***
## gpa:genderMale       -362.983   1564.660  -0.232 0.816654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1982 on 442 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3444, Adjusted R-squared:  0.3355
## F-statistic:  38.7 on 6 and 442 DF,  p-value: < 2.2e-16
```

This second model introduces an interaction between gender and GPA to examine if the effect of GPA on salary differs by gender. This adjustment is based on the hypothesis that gender might influence the relationship between academic performance and salary outcomes.

Main Effect of GPA: The coefficient for GPA is positive and significant (10622.747, p-value < 0.001), indicating a strong positive impact of GPA on salary for females (the baseline group).

Gender Effect: The coefficient for males is 899.015 but is not statistically significant (p-value = 0.878810), suggesting no significant salary difference for males compared to females when other factors are constant.

Interaction Term (gpa:genderMale): The interaction term coefficient is -362.983. We perform a t test here. The null hypothesis is coefficient of the interaction term is different from 0. However with a p-value of 0.816654, the test indicates that the effect of GPA on salary does not differ significantly between males and females. This result does not support the null hypothesis that females might gain more salary benefit from higher GPA levels due to negative job market biases.

- The adjusted R-squared is slightly lower at 0.3355, indicating a small decrease in explanatory power despite the added complexity.
- The overall model remains statistically significant (F-statistic p-value < 2.2e-16).

The analysis does not find evidence to support significant differences in how GPA affects salary by gender. Although the hypothesis was not confirmed, exploring such interaction effects is crucial for understanding potential gender dynamics in salary determinants. Future research may explore other variables or contexts where gender interactions could play a more pronounced role.

Part 1.4 Model Selection

```
n = length(resid(second_model))
salary_model_back_bic = step(second_model, direction = 'backward', k = log(n))
```

```
## Start:  AIC=6853.19
## salary ~ gpa + age + years_of_experience + gender + avg_10 +
##     gender * gpa
##
##                         Df Sum of Sq        RSS    AIC
## - gpa:gender             1    211419 1736551580 6847.1
## - age                    1  12280833 1748620994 6850.2
## <none>                               1736340161 6853.2
## - years_of_experience    1  45189984 1781530144 6858.6
## - avg_10                 1  80031462 1816371623 6867.3
##
## Step:  AIC=6847.14
## salary ~ gpa + age + years_of_experience + gender + avg_10
##
##                         Df Sum of Sq        RSS    AIC
## - age                    1  12843396 1749394976 6844.3
## <none>                               1736551580 6847.1
## - gender                 1  23966115 1760517695 6847.2
## - years_of_experience    1  47763874 1784315454 6853.2
## - avg_10                 1  82464716 1819016296 6861.9
## - gpa                  1 255608938 1992160518 6902.7
##
## Step:  AIC=6844.34
## salary ~ gpa + years_of_experience + gender + avg_10
##
##                        Df Sum of Sq        RSS    AIC
## - gender               1  22607105 1772002082 6844.0
## <none>                             1749394976 6844.3
## - years_of_experience  1  54752933 1804147909 6852.1
## - avg_10               1  82157451 1831552427 6858.8
## - gpa                  1 243558359 1992953335 6896.8
##
## Step:  AIC=6844
## salary ~ gpa + years_of_experience + avg_10
##
##                        Df Sum of Sq        RSS    AIC
## <none>                             1772002082 6844.0
## - years_of_experience  1  54918060 1826920142 6851.6
## - avg_10               1  85827089 1857829171 6859.1
## - gpa                  1 237909434 2009911515 6894.5
```

To achieve a model that can achieve the best balance between simplicity and explanatory power, backward method with bic as the metric is chosen as the process to iteratively remove the least significant predictors.

First Iteration: Removed the gender-GPA interaction due to its minimal impact on model performance as indicated by the largest decrease in BIC.

Second Iteration: Age was removed next, further lowering the BIC, suggesting that age did not provide sufficient explanatory power relative to its complexity cost.

Third Iteration: Gender was then eliminated, slightly reducing the BIC again, which indicated that gender did not significantly influence the model's explanatory ability after considering other factors.

With the steps that were taken to eliminate the variables, we were able to come up with a model that has the best balance between simplicity and the need to capture essential variability in the salary data.

The chosen model, emphasizing gpa, years of experience, and university ranking, achieves a robust balance between simplicity and predictive accuracy, making it suitable for real-world applications and further validation.

Part 1.5 Model Comparison

```r
summary(first_model)
```

```
##
## Call:
## lm(formula = salary ~ gpa + age + years_of_experience + gender +
##     avg_10, data = salary_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6055.9  -757.6   589.4  1545.5  5240.5
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         26373.90    4760.49   5.540 5.19e-08 ***
## gpa                 10450.88    1294.22   8.075 6.43e-15 ***
## age                  -142.65      78.81  -1.810  0.07096 .  
## years_of_experience   685.68     196.43   3.491  0.00053 ***
## genderMale           -467.30     188.99  -2.473  0.01379 *  
## avg_10                 17.20       3.75   4.587 5.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1980 on 443 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3443, Adjusted R-squared:  0.3369 
## F-statistic: 46.53 on 5 and 443 DF,  p-value: < 2.2e-16
```

```r
summary(salary_model_back_bic)
```

```
##
## Call:
## lm(formula = salary ~ gpa + years_of_experience + avg_10, data = salary_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6522.0  -897.2   208.1  1599.2  5304.4 
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         24514.076   4651.704   5.270 2.13e-07 ***
## gpa                  9928.699   1284.513   7.730 7.25e-14 ***
## years_of_experience   729.833    196.525   3.714  0.00023 ***
## avg_10                 17.531      3.776   4.643 4.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1996 on 445 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3309, Adjusted R-squared:  0.3264 
## F-statistic: 73.37 on 3 and 445 DF,  p-value: < 2.2e-16
```

The models that have been selected to do the comparison are the model that was first fitted (full model) and the model that fitted in the previous part, 1.4 (gpa, years of experience, and university ranking).

First Model (Full Model) Variables: GPA, age, years of experience, gender, average university ranking (avg_10). Adjusted R-squared: 0.3369 This model, though more complex, shows a slightly higher adjusted R-squared, suggesting better explanatory power but at the risk of overfitting.

Second Model (Reduced Model) Variables: GPA, years of experience, average university ranking (avg_10). Adjusted R-squared: 0.3264 Simpler and more generalizable with a modest trade-off in explanatory power.

As adjusted R^2 considering the explanatory power and model complexity, I would keep the first model for further analysis as it has a higher Adjusted R-squared.

Part 2

Part a Reason for Selected Model

```r
summary(first_model)
```

```
##
```

```
## Call:
## lm(formula = salary ~ gpa + age + years_of_experience + gender + 
##     avg_10, data = salary_final)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6055.9  -757.6   589.4  1545.5  5240.5 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          26373.90    4760.49   5.540 5.19e-08 ***
## gpa                  10450.88    1294.22   8.075 6.43e-15 ***
## age                   -142.65      78.81  -1.810  0.07096 .  
## years_of_experience    685.68     196.43   3.491  0.00053 ***
## genderMale            -467.30     188.99  -2.473  0.01379 *  
## avg_10                  17.20       3.75   4.587 5.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1980 on 443 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3443, Adjusted R-squared:  0.3369 
## F-statistic: 46.53 on 5 and 443 DF,  p-value: < 2.2e-16
```

```
model = lm(salary ~ gpa + years_of_experience + gender + avg_10, data = salary_final)
summary(model)
```

```
## 
## Call:
## lm(formula = salary ~ gpa + years_of_experience + gender + avg_10, 
##     data = salary_final)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6296.9  -915.3   426.4  1419.5  5151.1 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          24270.171   4628.258   5.244 2.44e-07 ***
## gpa                  10054.343   1278.806   7.862 2.89e-14 ***
## years_of_experience    728.737    195.488   3.728 0.000218 ***
## genderMale            -453.487    189.319  -2.395 0.017018 *  
## avg_10                  17.167      3.759   4.566 6.44e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1985 on 444 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3395, Adjusted R-squared:  0.3335 
## F-statistic: 57.04 on 4 and 444 DF,  p-value: < 2.2e-16
```

The model chosen is:

$$salary = \beta_0 + \beta_1 \times gpa + \beta_2 \times age + \beta_3 \times years\_of\_eperience + \beta_4 \times gender + \beta_5 \times avg\_10 + \epsilon$$

The model was chosen because of these following reasons:

Relevance and Interpretability: The chosen predictors (GPA, years of experience, and university ranking) are highly relevant and intuitively important for predicting salaries. Each factor is directly related to job performance and potential, which are critical considerations in salary decisions. This relevance enhances the interpretability of the model, making it easier to communicate and justify in professional and academic settings.

Simplicity: The model strikes a balance between simplicity and the ability to capture significant variations in salary. By focusing on these key predictors, the model avoids unnecessary complexity while maintaining robust predictive power.

Part b Fitted Model

Fitted model for male is:

$$\widehat{salary} = 26373.9 + 10450.88 \times gpa - 142.65 \times age + 685.68 \times years\_of\_eperience - 467.30 \times genderMale + 17.20 \times avg\_10$$

The fitted model for female is:

$$\widehat{salary} = 26373.9 \ + \ 10450.88 \times gpa \ - 142.65 \times age \ + 685.68 \times years\_of\_eperience \ + 17.20 \times avg\_10$$

Part c N and P for Model

```r
n = length(resid(first_model))

p =length(coef(first_model))

n
```

```
## [1] 449
```

```r
p
```

```
## [1] 6
```

n for this model is 449, and p for this model is 6.

Part d Standard Deviation for Y

```r
sd_y = sd(salary_final$salary)

sd_y_hat = sd(predict(first_model))

sd_y
```

```
## [1] 2437.678
```

```r
sd_y_hat
```

```
## [1] 1426.698
```

Actual Salary Standard Deviation (sd_y): The actual salaries have a standard deviation of 2437.678, which represents the typical variation in salary within your dataset.

Predicted Salary Standard Deviation (sd_y_hat): The predicted salaries from your model have a standard deviation of 1426.698, which is substantially lower than the actual salaries.

The lower standard deviation of the predicted salaries suggests that the model tends to average out salary outcomes. This results in predictions that are less varied than the actual data. Also, the fact that sd_y_hat is nearly half of sd_y indicates that the model does not capture a significant portion of the variability in actual salaries. This could be due to several factors, such as the absence of critical predictors, nonlinear relationships that the model cannot capture, or inherent variability in the salary data that is not explained by the variables included in the model. Lastly, the substantial difference in standard deviations raises a question about whether the model's simplicity—while beneficial for generalization—might be overly restrictive in terms of capturing the full complexity and dynamics of salary determinations.

The comparison of these two standard deviations highlights a limitation in the model's ability to capture all the variations in salary data. It suggests that while the model is useful for predicting general trends and making broadly accurate salary predictions, it may not be as effective for applications requiring precise prediction of individual salaries. This observation should guide further refinement of the model, potentially by exploring additional predictors, incorporating non-linear transformations, or using different modeling techniques that might capture more of the salary data's inherent variability.

Part e Collinearity

```r
library(car)
```

```
## Loading required package: carData
```

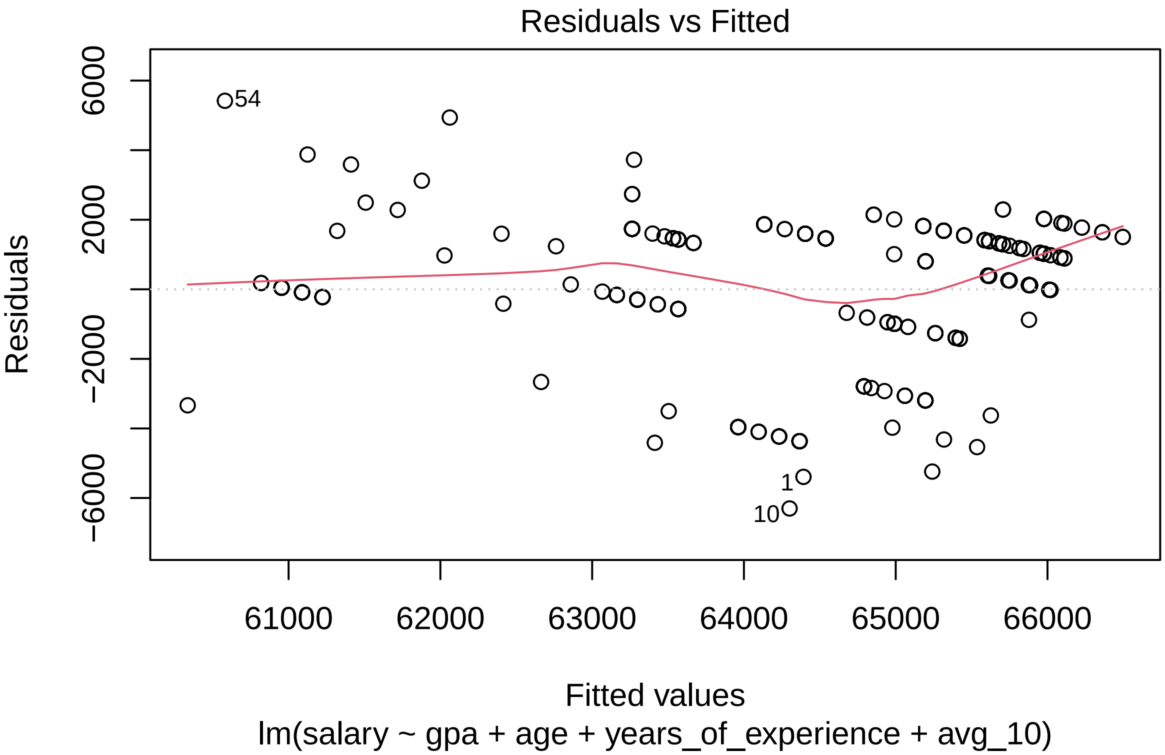
```r
model_no_category = first_model = lm(salary ~ gpa + age + years_of_experience + avg_10,
                                      data = salary_final)
vif(model_no_category)
```

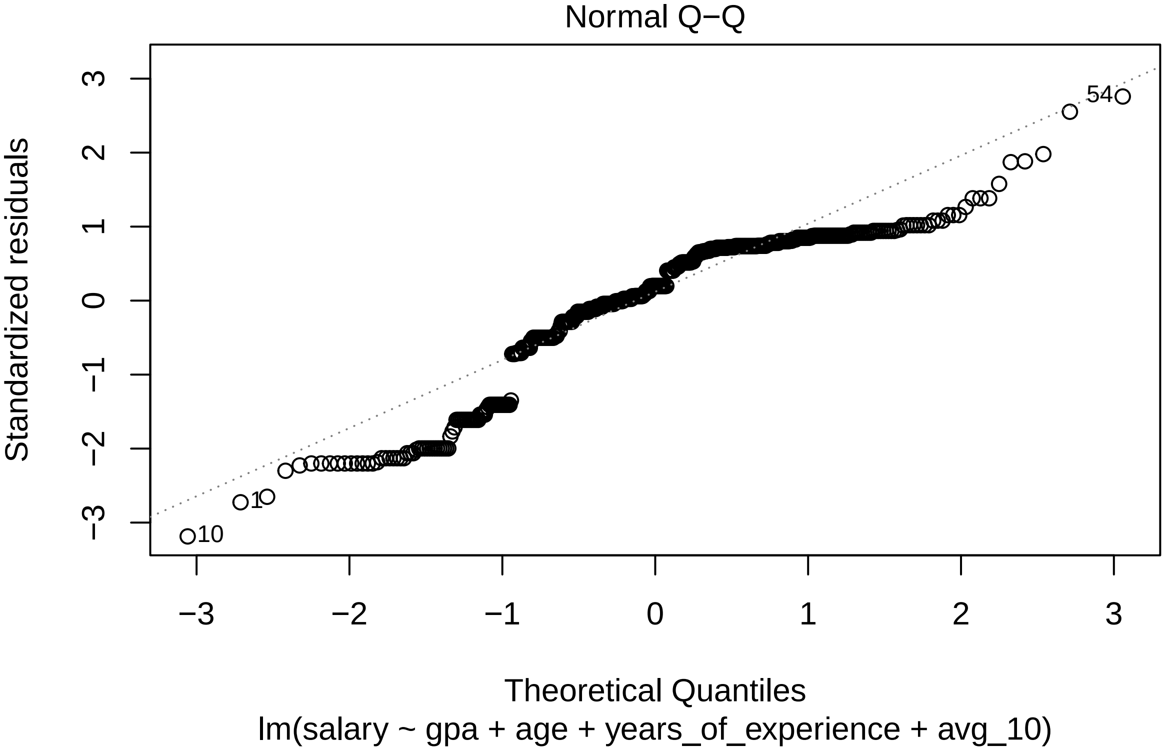
```
##                 gpa                 age years_of_experience              avg_10
##            2.997242            1.031671            2.902687            1.933464
```

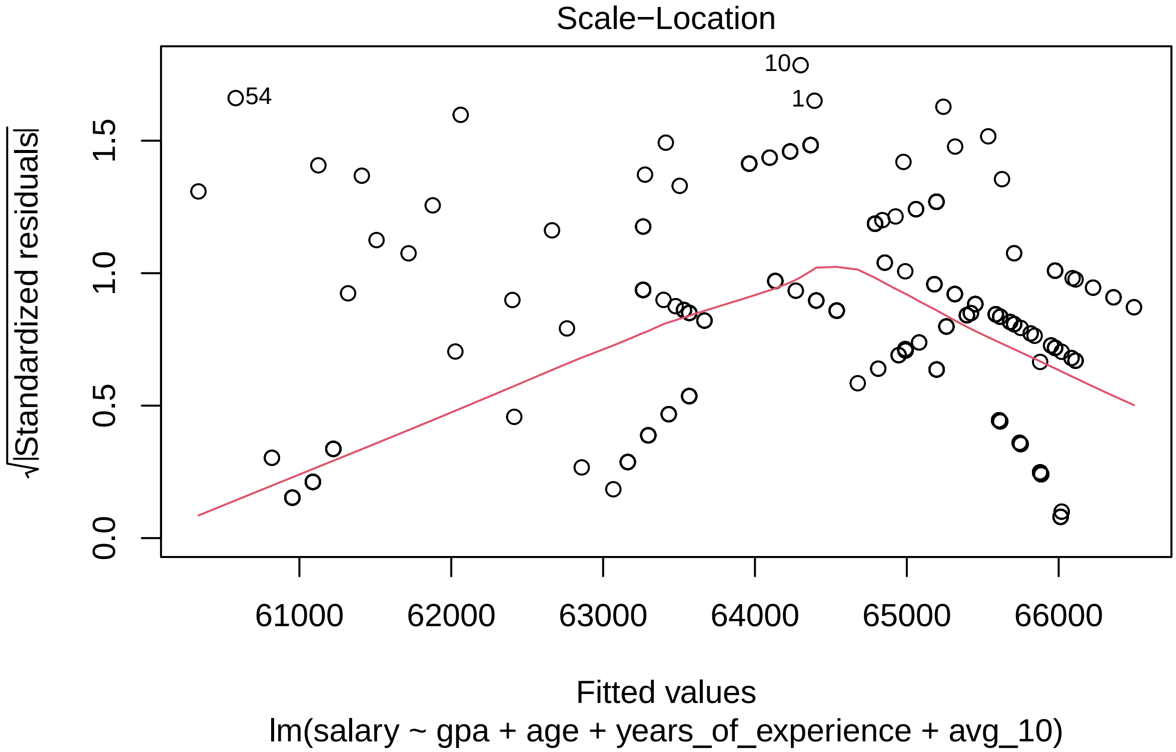
Based on the VIF results, there is no substantial collinearity among the predictors in the regression model. The values indicate that while there is some mild correlation between GPA and years of experience, it is not at a level that would undermine the validity of the regression analysis. Therefore, there is no need to be overly concerned about collinearity affecting the outcomes of this specific model. This suggests that the model's estimates are stable and that multicollinearity does not significantly bias the interpretation of the coefficients.

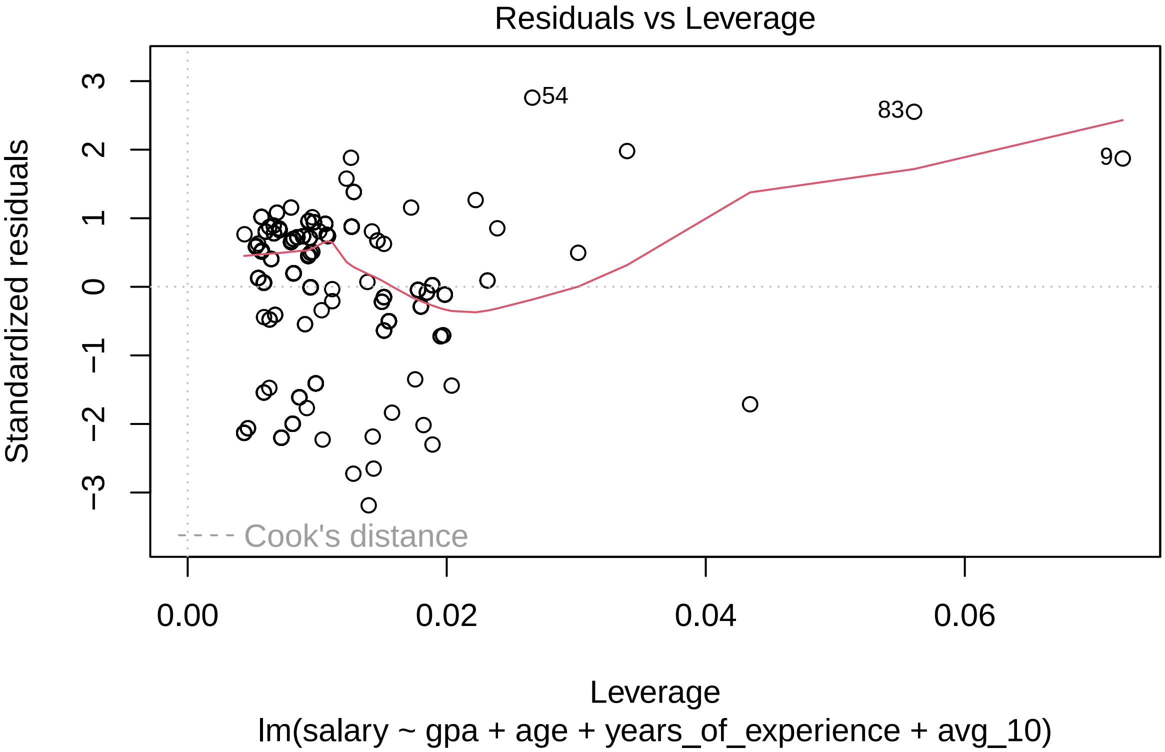
Part f Model Assumptions

```r
plot(first_model)
```


Residuals vs Fitted
Residuals
6000
2000
−2000
−6000
54
1
10
61000
62000
63000
64000
65000
66000
Fitted values
lm(salary ~ gpa + age + years_of_experience + avg_10)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
54
1
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(salary ~ gpa + age + years_of_experience + avg_10)





- Linearity: Based on residuals vs. fitted plot, there are some outliers and the red line is not relatively flat, indicating potential non-linearity in the relationship between the predictors and the response variable. Therefore indicating that the model might not adequately capture the true relationship, suggesting the need for exploring non-linear transformations or adding interaction terms.
- Equal variance: Based on residuals vs. fitted plot, the residuals seem to fan out or have varying spreads across the range of fitted values as seen in the scale-location plot. This suggests that the variance

of the residuals is not constant, which violates the homoscedasticity assumption. Considerations for transformations of the response variable or using heteroscedasticity-consistent standard errors should be contemplated.

- Leverage and Influence: The Q-Q plot shows that residuals deviate from the straight line, especially at the tails. This indicates that residuals are not normally distributed. The violation of this assumption can affect the reliability of hypothesis tests on the coefficients. Transformations of the response variable or robust regression methods may be needed. There are a few points with higher leverage, but they don't seem to excessively influence the model, as indicated by Cook's distance being within acceptable limits. This suggests that while these points are outliers in the X space, they are not excessively influencing the regression outcomes.
- Multicollinearity: As previously analyzed (part e), the VIF values are all well below 10, suggesting that multicollinearity is not a concern for this model. The predictors are sufficiently independent in terms of explaining the variability in the response variable, supporting the reliability of the estimated coefficients.
- Independent: This assumption is generally checked by considering the data collection process, which is not visible from plots but needs context understanding. Would need to ensure that data points are independent of each other.

Part g Unusual Observations

```r
hatvalues(first_model)[which(hatvalues(first_model) > 2 * p / n)]
```

```
##          2          6          7          9         83 
## 0.03014947 0.03392551 0.04342242 0.07219530 0.05607998 
```

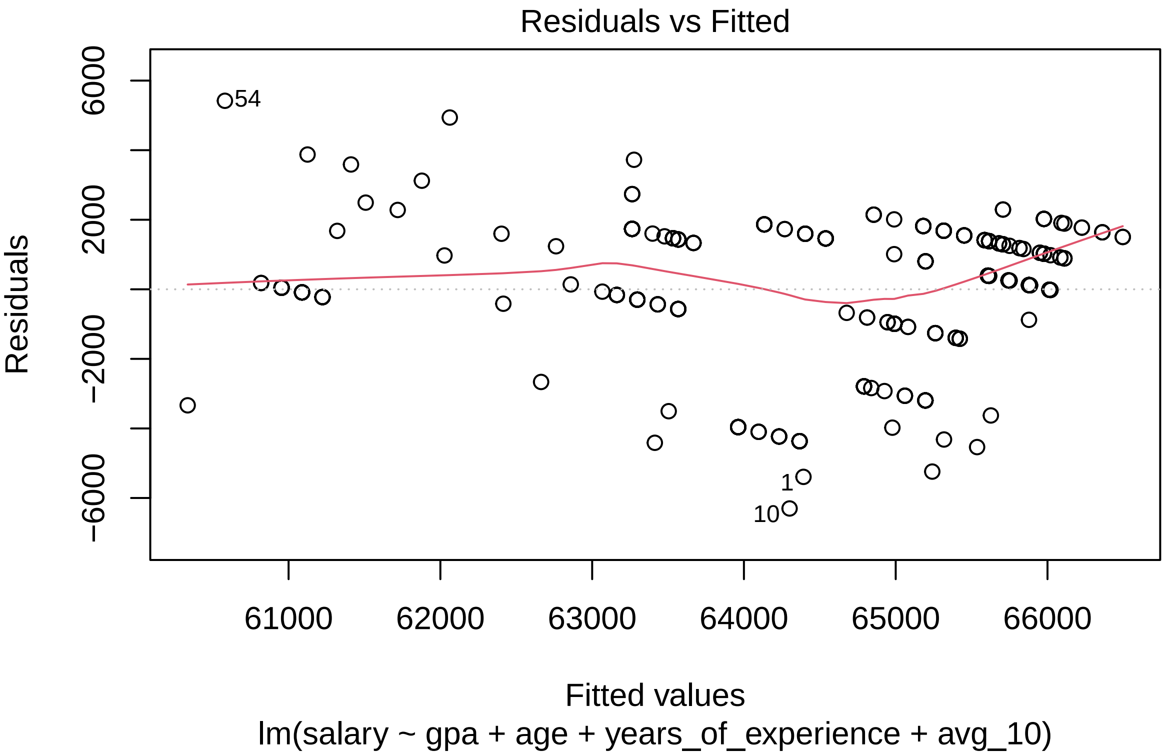
```r
rstandard(first_model)[which(abs(rstandard(first_model)) > 2)]
```

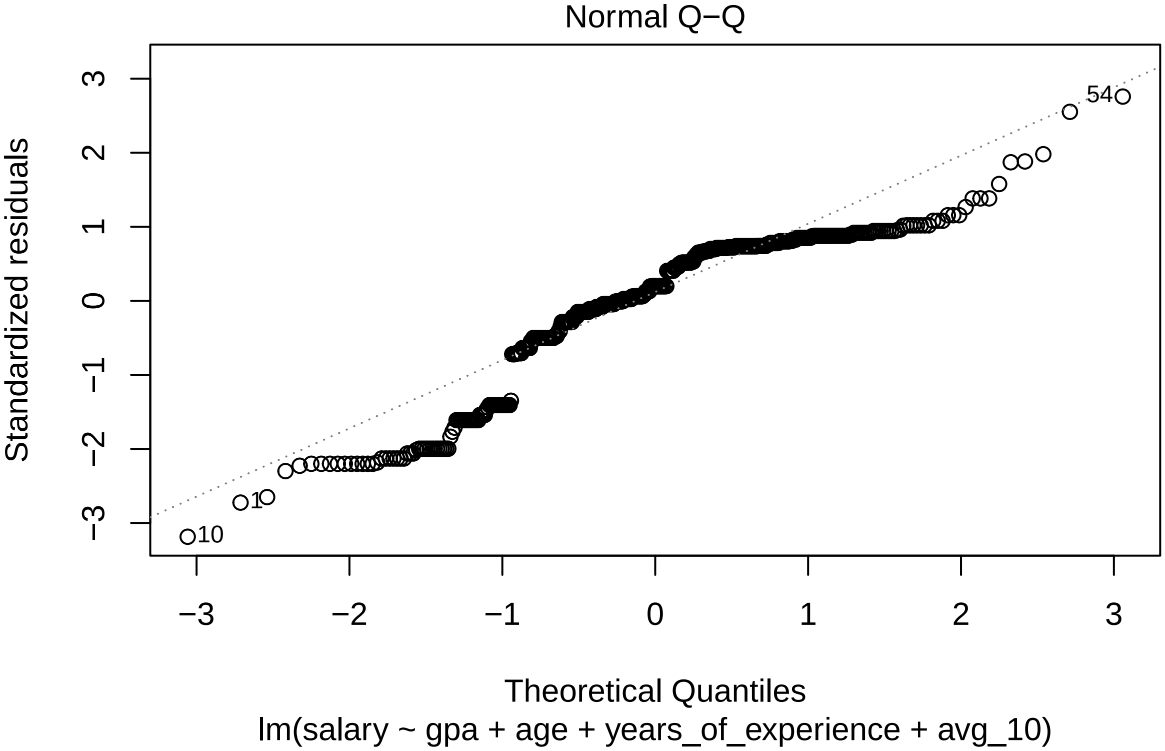
```
##         1         3        10        54        83        93        94        95 
## -2.725389 -2.016109 -3.186553  2.758571  2.552413 -2.129816 -2.129816 -2.129816 
##        96        97        98        99       102       103       104       106 
## -2.062283 -2.200827 -2.200827 -2.200827 -2.200827 -2.062283 -2.200827 -2.200827 
##       107       110       111       112       115       116       118       122 
## -2.129816 -2.200827 -2.129816 -2.129816 -2.200827 -2.129816 -2.200827 -2.062283 
##       123       234       279       301       450 
## -2.200827 -2.650982 -2.227790 -2.299870 -2.184176 
```

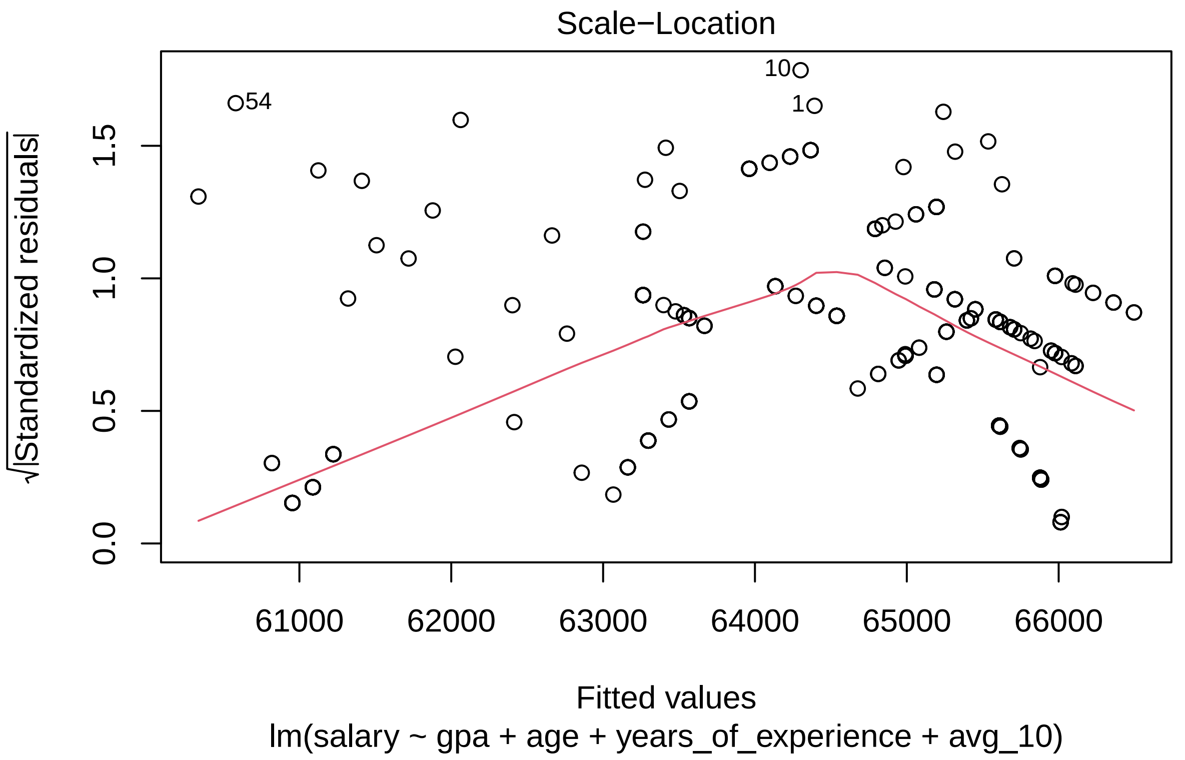
```r
cooks.distance(first_model)[which(cooks.distance(first_model) > 4 / n)]
```

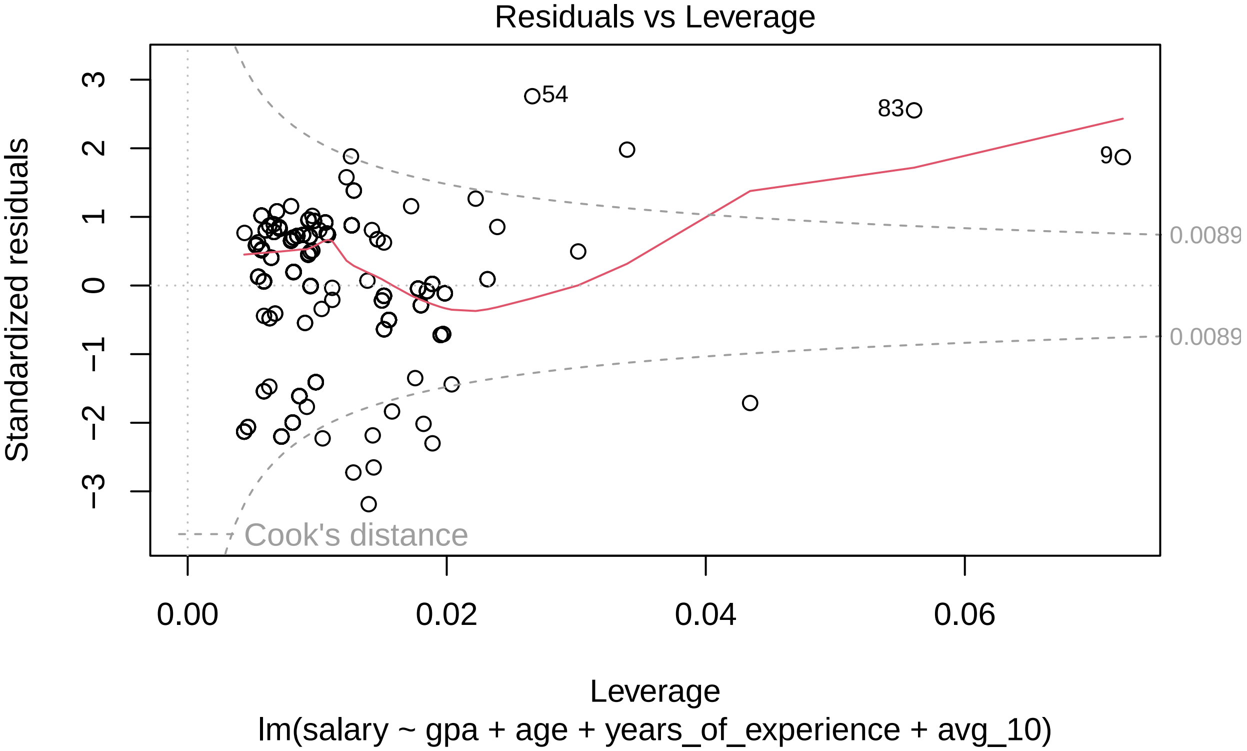
```
##           1           3           4           6           7           9 
## 0.019250098 0.015076585 0.009047526 0.027531323 0.026625193 0.054484693 
##          10          54          83         234         235         279 
## 0.028786636 0.041599603 0.077411314 0.020473561 0.010804907 0.010450110 
##         301         450 
## 0.020381835 0.013828374 
```

```r
plot(first_model, cook.levels = c(4 / n))
```


Residuals vs Fitted
Residuals
6000
2000
−2000
−6000
54
1
10
61000
62000
63000
64000
65000
66000
Fitted values
lm(salary ~ gpa + age + years_of_experience + avg_10)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
54
1
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(salary ~ gpa + age + years_of_experience + avg_10)





According to the Residuals vs Leverage plot above, more than 10 influential points with large Cook's distance can be found (points that outside of the curves according to the graph using threshold of $4/n$). It would be essential to remove these influential observations from the dataset and then refit the regression model. This step can help improve the model's robustness and the accuracy of the model's estimates.

Part h LOO RMSE

```r
residuals = resid(first_model)

leverage = hatvalues(first_model)

n = length(residuals)

RMSE = sqrt(1/n * sum((residuals/(1-leverage))^2))

RMSE
```

```
## [1] 2002.655
```

Using leave-one-out cross-validation method (LOOCV), The estimated Root Mean Square Error (RMSE) is 2002.655. With an RMSE of 2002.655, it is an moderate level of prediction error relative to the scale of salary values being predicted. This level of accuracy can be acceptable depending on the tolerance of errors for this data exploration. Exploring model enhancements such as including more predictive features, using polynomial or interaction terms to capture non-linear relationships, or employing more sophisticated modelling techniques can help lower the RMSE if needed.

Part i Model Complexity

With a p of 6 and n of over 400, the size of the model would not be a concern, as it has a ratio of over 66 observations per parameter, significantly above the recommended minimum of 10 observations per parameter. This high ratio can ensure that each parameter is well-supported by the data, enhancing the reliability, generalizability, and stability of the model's estimates.

Part 3

(Statistical Tests were performed in part 1 and 2, for example in part 1.2, we performed a test for interaction term for gpa and gender)