PREDICTION OF GENDER PAY EQUALITY

Aims

- The aim is to accurately predict if there is any discrimination going on starting salary paid based on sex.

Background

- The analysis is performed on a dataset 'Bank.csv' containing 93 employees of a bank having same category of job and same skill. This dataset provides 6 recorded features for each employee that includes Bsal (base salary at the time of hire), Sex (Male or Female), Senior (Months since first hired), Age (Age in months), Educ (Years of Education) and Exper (work experience prior to job with the bank in months).

Statistical methods and results

Variables: -

The dataset has six variables.

1. Bsal (base salary at the time of hire)
2. Sex (Male or Female)
3. Senior (Months since first hired)
4. Age (Age in months)
5. Educ (Years of Education)
6. Exper (work experience prior to job with the bank in months)

Computational methods

The study has been done using statistical package R and R Studio of version 1.4.1106. Boxplot () function has been used to get the boxplot of Bsal variable with respect to gender. "MASS" package has been used for creating the BoxCox plots to find if transformation is needed of our data. Normal "plot ()" function has been used to compare the change in Bsal with respect to predictor variables. The lm () was used for the linear regression model fitting on the bank data and the summary () command was used to get the summary statistics of the fitted model with the regression equation, residuals, and the other important statistics. The plot(lm) command was used to generate the diagnostic plots of the regression model. Par (mfrow =c (2,2)) command has been used to compact all the four diagnostic plots of linear regression models in one single window.

Findings: -

- It was found out that on average, equally qualified and experienced males received $351-$841 more starting salary then the females (Table 4). The equations for Bsal predictor are: -

For male,

Bsal = -600.7 + 23.99 Senior + 6.17 Age + 528.9Educ+ 197.6Exper - 0.279Age: Exper – 0.358Age: Educ - 3.33Senior: Educ

For female,

Bsal = -1200 + 23.99 Senior + 6.17 Age + 528.9Educ+ 197.6Exper - 0.279Age: Exper – 0.358Age: Educ - 3.33Senior: Educ

We loaded up the bank.csv data in RStudio and plotted the scatterplot matrix for all the explanatory and predictor variables to visualize the relationship between them.

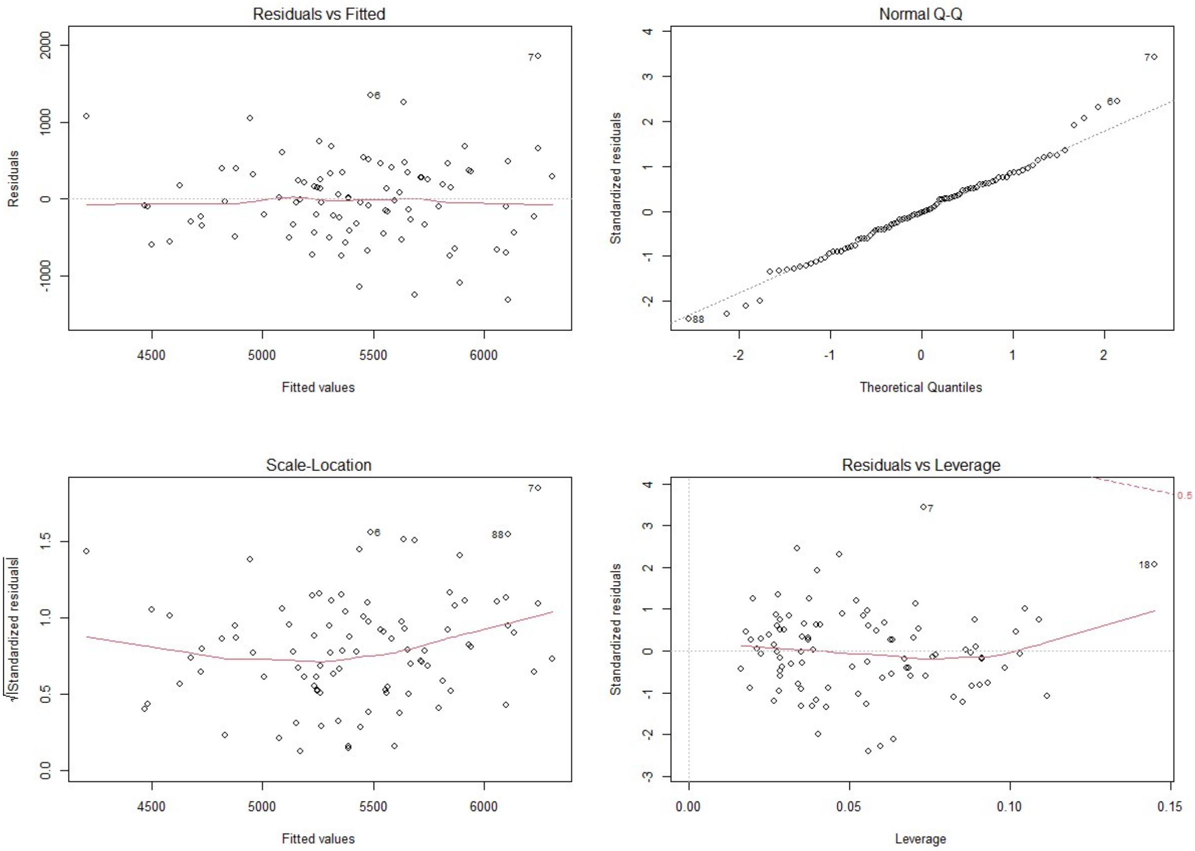


Figure 1: Scatterplot matrix of bank_data

From the above plot, the Exper and Educ variables are seen to have a strong linear relation whereas the relation among other variables doesn’t look to be that clear. The other predictor variables don’t have any strong relationships amongst themselves, showing that they are not strongly correlated.

Checking if experience and education are making any differences in pay,

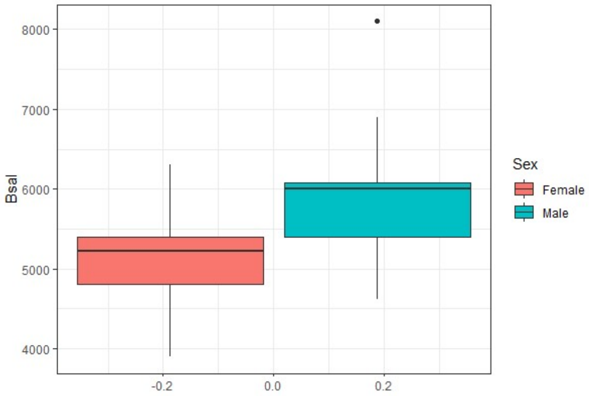
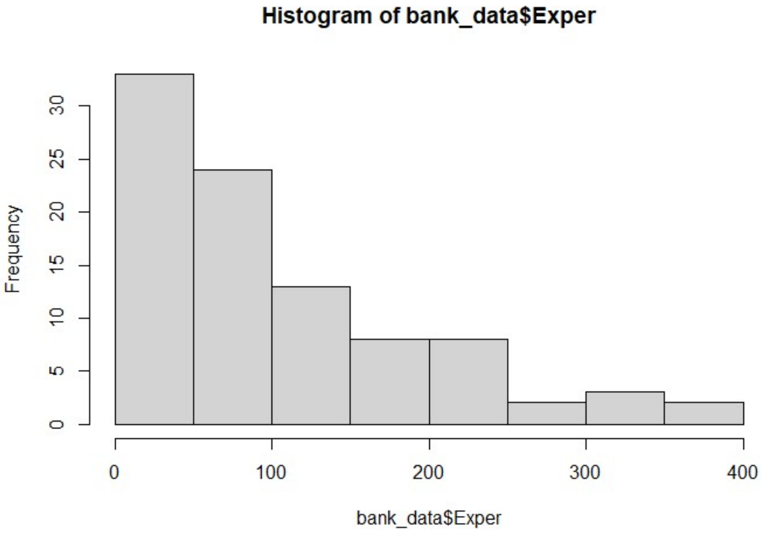


Figure 2: Boxplot for Bsal

Form the boxplot above as well, it is seen that the average starting salary is comparatively higher for male than for female without taking any other confounding variables. So, now we are going to check if there really exists a sex-based discrimination based on pay taking other non-discriminatory predictors in account as well.



1

Figure 3: Histogram for Exper variable data

Exper looks to be positively skewed, and so a square root transformation can be done to make it normal. After checking on further down, it did make significant change in our R^2 value of models. So, we will stick to transforming Exper predictor variable.

So, now constructing a preliminary model for the non-discriminatory variables to predict Bsal: -

Fit_pre<-lm (Bsal~Senior+Age+Educ+Exper, data=bank_data)

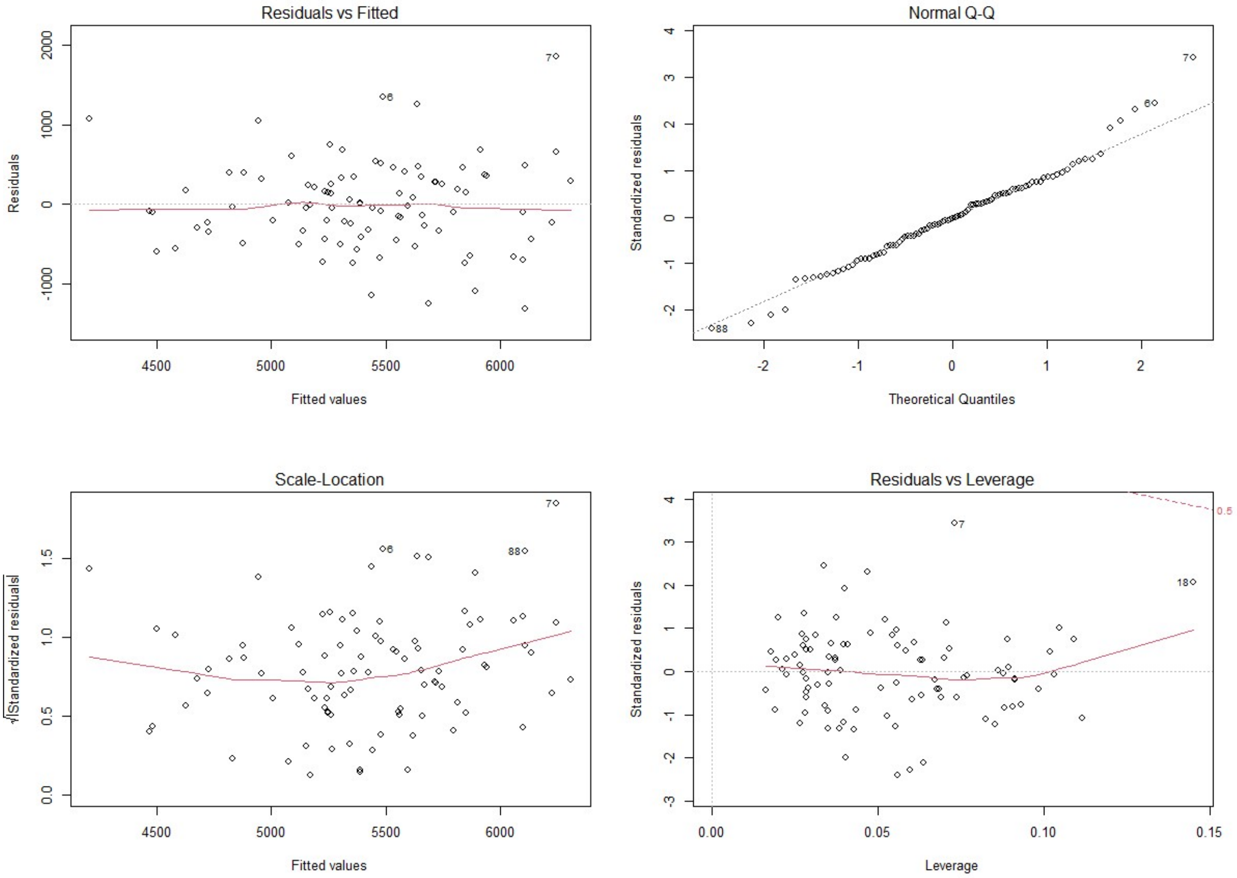
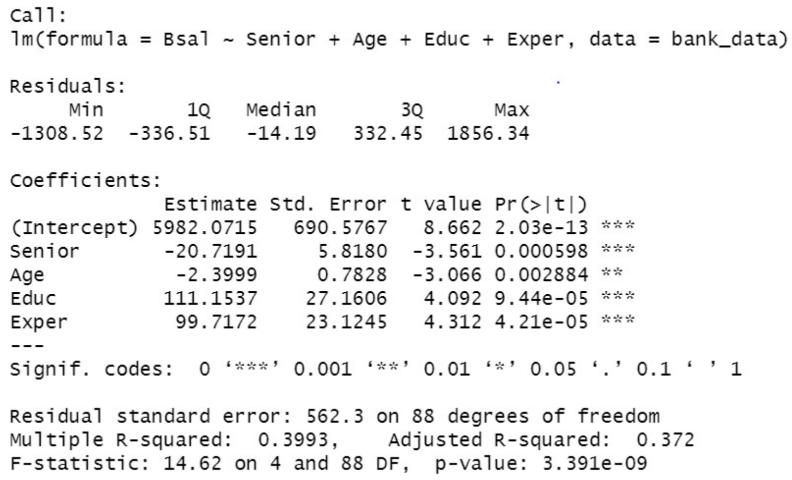


Figure 3: Diagnostic plots for the model Fit

(Residual vs Fitted) plots above shows regression doing a good job as there can be seen almost equal spread of the residuals above and below the horizontal line. Normal Q-Q plots shows almost majority of the residuals follows the straight line in the models without deviating severely, which tells us that the residuals are normally distributed. Scale-Location plot above as well shows that residuals are spread almost equally over the ranges of the predictor values for the model. This shows assumption of equal variance is met by the model. Residual vs Leverage plot shows that almost all the data lies on the horizontal line meaning they don't influence the fit heavily if they were to be removed.

Table 1: Summary of model fit



```
Call:
lm(formula = Bsal ~ Senior + Age + Educ + Exper, data = bank_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1308.52  -336.51   -14.19   332.45  1856.34 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5982.0715   690.5767   8.662 2.03e-13 ***
Senior       -20.7191     5.8180  -3.561 0.000598 ***
Age           -2.3999     0.7828  -3.066 0.002884 ** 
Educ         111.1537    27.1606   4.092 9.44e-05 ***
Exper         99.7172    23.1245   4.312 4.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 562.3 on 88 degrees of freedom
Multiple R-squared:  0.3993,	Adjusted R-squared:  0.372 
F-statistic: 14.62 on 4 and 88 DF,  p-value: 3.391e-09
```

This model explains 39% of variation in the basic salary while taking non-discriminatory variables as predictor variables. Here, hesterics shows almost all the predictors are good predictors with significant p-values. We further want to see if the R^2 values can be further improved to make the model more robust. So, we performed BoxCox analysis to figure out if any sort of transformation is needed.

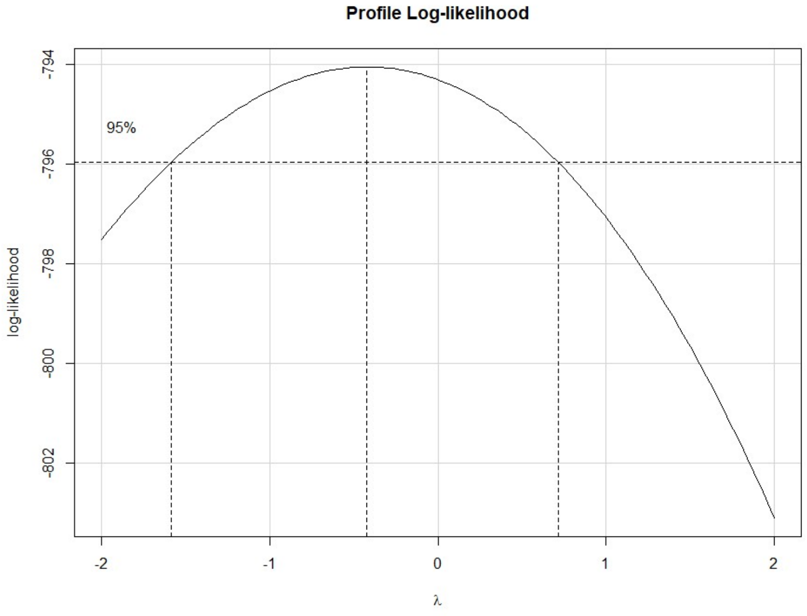


Figure 4: Boxcox Analysis for model Fit_pre

Form the above boxcox plot, we got our value of lambda near to 0 meaning it is suggesting log transformation.

So, after log transformation, our model becomes: -

Fit_trans<-lm (log (Bsal)~Senior+Age+Educ+Exper, data=bank_data)

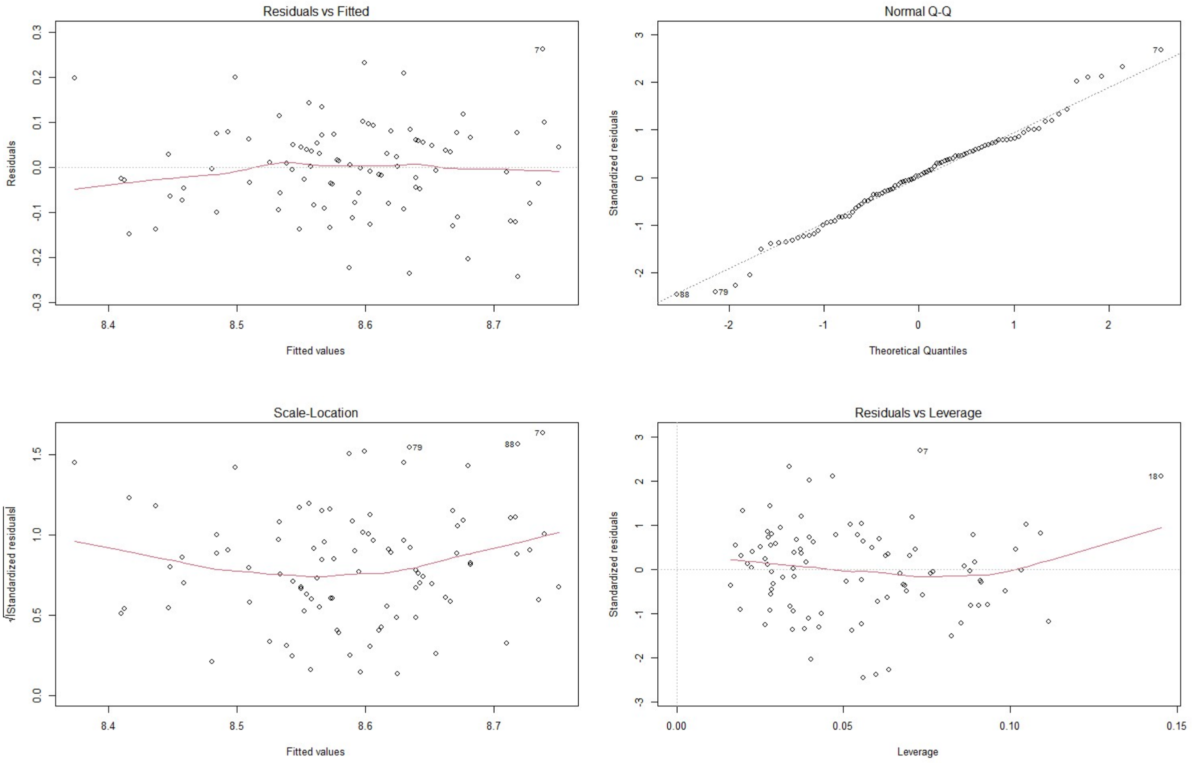
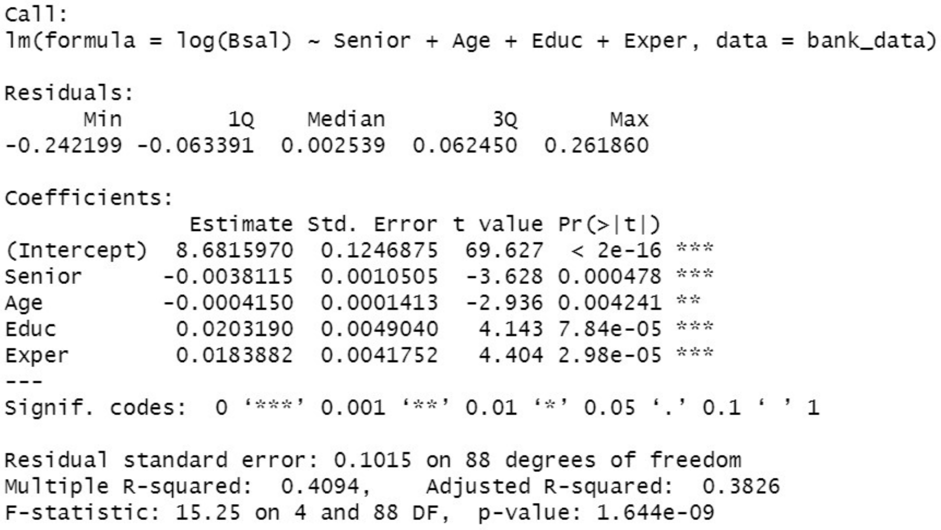


Figure 5: Diagnostic plot for model Fit_trans



```
Call:
lm(formula = log(Bsal) ~ Senior + Age + Educ + Exper, data = bank_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.242199 -0.063391  0.002539  0.062450  0.261860 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.6815970  0.1246875  69.627  < 2e-16 ***
Senior      -0.0038115  0.0010505  -3.628 0.000478 ***
Age         -0.0004150  0.0001413  -2.936 0.004241 ** 
Educ         0.0203190  0.0049040   4.143 7.84e-05 ***
Exper        0.0183882  0.0041752   4.404 2.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1015 on 88 degrees of freedom
Multiple R-squared:  0.4094,	Adjusted R-squared:  0.3826 
F-statistic: 15.25 on 4 and 88 DF,  p-value: 1.644e-09
```

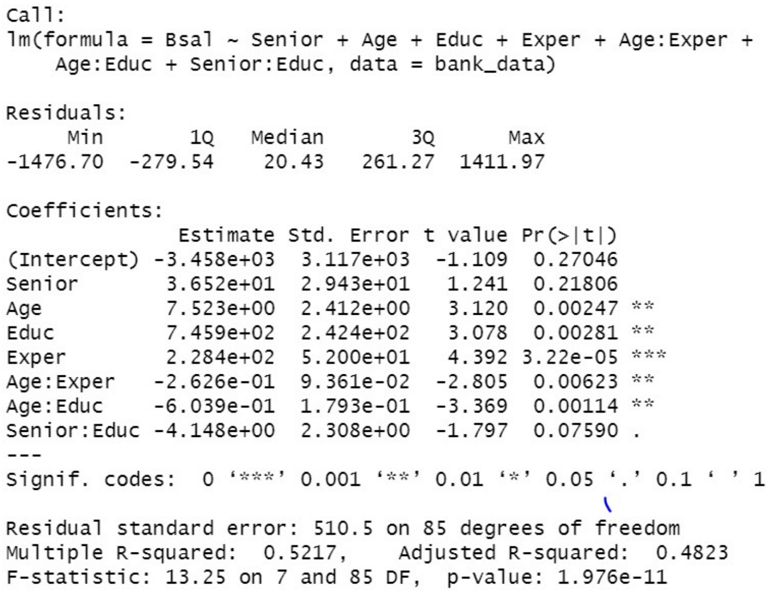
So, the R^2 value shows a bit of improvement but is not of significant manner. This transformed model explains the 40% variation in Bsal to 39% of untransformed model. So, as the diagnostic plots doesn’t show any significant changes as well, we stick with the Fit_pre model. But looks like it’s good either way going with log transformation or no transformation at all.

As of now to find the optimal set of variables and their interaction terms, we use Step () function to find out the possible potential interactions of variables.

We get a couple of models and choosing on the base of AIC value, we chose the model with least AIC value here i.e., AIC =1167.43. so, our model will be: -

Fit_final<-lm (Bsal ~ Senior+Age+Educ+Exper+Age: Exper + Age: Educ + Senior: Educ, data=bank_data)

Table 2: summary table for model Fit_final

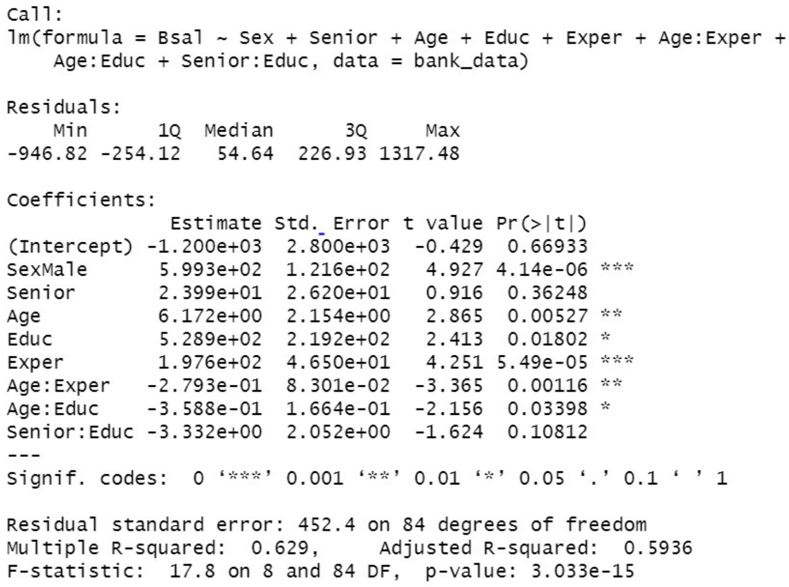


```
Call:
lm(formula = Bsal ~ Senior + Age + Educ + Exper + Age:Exper +
    Age:Educ + Senior:Educ, data = bank_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1476.70  -279.54    20.43   261.27  1411.97

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.458e+03  3.117e+03  -1.109  0.27046
Senior       3.652e+01  2.943e+01   1.241  0.21806
Age          7.523e+00  2.412e+00   3.120  0.00247 **
Educ         7.459e+02  2.424e+02   3.078  0.00281 **
Exper        2.284e+02  5.200e+01   4.392 3.22e-05 ***
Age:Exper   -2.626e-01  9.361e-02  -2.805  0.00623 **
Age:Educ    -6.039e-01  1.793e-01  -3.369  0.00114 **
Senior:Educ -4.148e+00  2.308e+00  -1.797  0.07590 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 510.5 on 85 degrees of freedom
Multiple R-squared:  0.5217,    Adjusted R-squared:  0.4823
F-statistic: 13.25 on 7 and 85 DF,  p-value: 1.976e-11
```

The R^2 value improves to explaining 52% of variation in Bsal. The Senior variable and its interaction term senior: Educ don’t look to be that significant as their p-value looks quite higher than 0.005, proving their presence doesn’t make significant change in prediction of Bsal. Now, we are all good to add Sex in our model. So, our final model turns out to be: -

Fit_final<-lm (Bsal ~ Sex+Senior+Age+Educ+Exper+Age: Exper + Age: Educ + Senior: Educ, data=bank_data)

Table 3: Summary table for model Fit_final



```
Call:
lm(formula = Bsal ~ Sex + Senior + Age + Educ + Exper + Age:Exper +
    Age:Educ + Senior:Educ, data = bank_data)

Residuals:
    Min      1Q  Median      3Q     Max
-946.82 -254.12   54.64  226.93 1317.48

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.200e+03  2.800e+03  -0.429  0.66933
SexMale      5.993e+02  1.216e+02   4.927 4.14e-06 ***
Senior       2.399e+01  2.620e+01   0.916  0.36248
Age          6.172e+00  2.154e+00   2.865  0.00527 **
Educ         5.289e+02  2.192e+02   2.413  0.01802 *
Exper        1.976e+02  4.650e+01   4.251 5.49e-05 ***
Age:Exper   -2.793e-01  8.301e-02  -3.365  0.00116 **
Age:Educ    -3.588e-01  1.664e-01  -2.156  0.03398 *
Senior:Educ -3.332e+00  2.052e+00  -1.624  0.10812
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 452.4 on 84 degrees of freedom
Multiple R-squared:  0.629,     Adjusted R-squared:  0.5936
F-statistic:  17.8 on 8 and 84 DF,  p-value: 3.033e-15
```

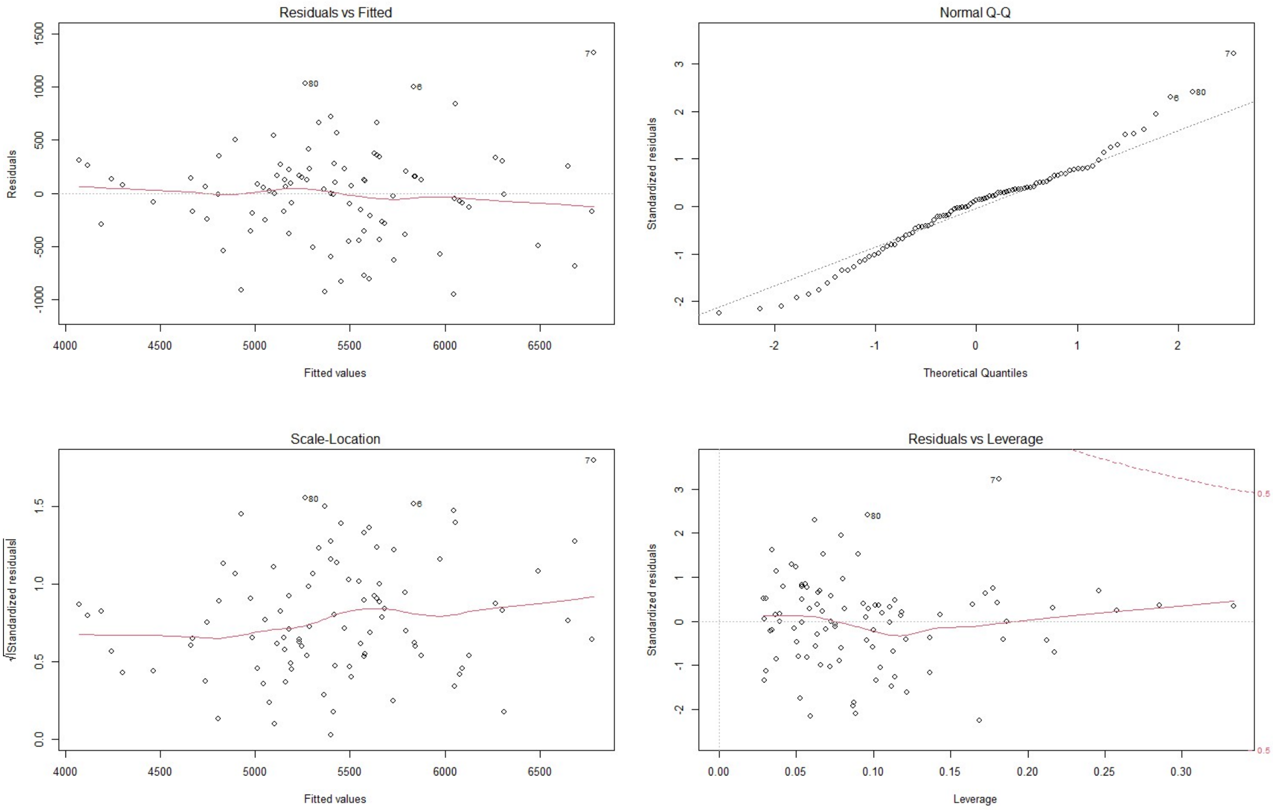
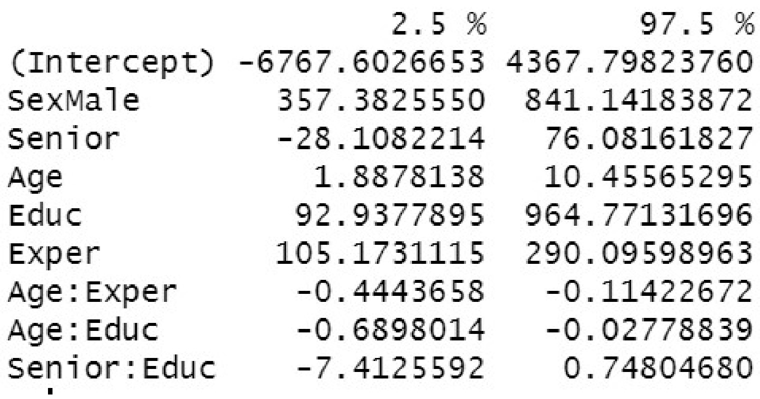


Figure 6: Diagnostic plot for model Fit_final

So, after adding Sex in the model, our final model explains 62.9% of the variation in Bsal and predictor terms like Age, Education, Experience, and interaction term Age: Exper are looking significant except Senior, interaction term of Senior and interaction term Age: Educ as their p-values are quite higher than 0.05. The diagnostic plots look good even though the Q-Q plot doesn't look that ideal. The Residual vs fitted plot has shown that the residuals are equally spread, and the Residuals and Leverage plot does show only few outliers like 7 which have a high leverage otherwise most of the data points doesn't look to have high influence in the model if is removed.

Table 4: confidence interval for predictors of model Fit_final



```
                   2.5 %       97.5 %
(Intercept) -6767.6026653 4367.79823760
SexMale       357.3825550  841.14183872
Senior        -28.1082214   76.08161827
Age             1.8878138   10.45565295
Educ           92.9377895  964.77131696
Exper         105.1731115  290.09598963
Age:Exper      -0.4443658   -0.11422672
Age:Educ       -0.6898014   -0.02778839
Senior:Educ    -7.4125592    0.74804680
```

From the 95% confidence interval generated above as well, the model predicts that there is certainly a sex-based discrimination going on paid starting salary, with the prediction of male getting $357 - $841 higher than females.