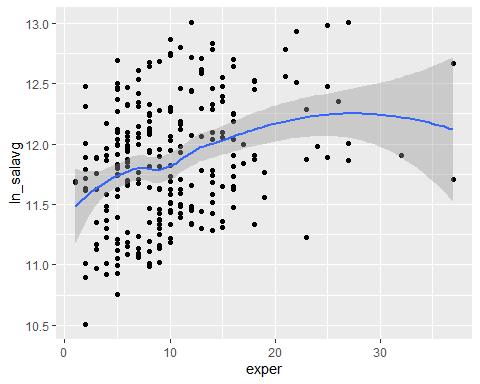
Model diagnostic

# Model Diagnostic

## Functional form for continuous predictor



The only continuous variable in our model is exper. So we made scatter plot for exper and ln\_salavg. According to the scatter plot, we hypothesized there is non-linear relationship between exper and ln\_salavg.

### Polynomial regression

For model 1 (interaction gender\*exper)

We added its centering power 2, 3, 4 to original model 1. But the summary result indicates that there is no association between higher order of exper and ln\_salavg. Alao, we fail to reject original model 1 by using anova to compare original model with polinomial model.

For model 2 (interaction gender\*rank)

We use the same process to test whether polynomial regression could improve model 2. The higher order model doesn’t perform better than original one.

### Piecewise model

To make piecewise model, I choose exper = 7.5 and exper = 9 as knots according to the scatter plot.

2 spline of piecewise model 1 are not significant, so there is not enough significant evidence to show piecewise form of model 1 is reasonable.

For model 2, P-value of spline at exper = 9 is 0.001 which indicate coefficient of exper change significantly here. However, the adjusted R-square is far less than original model 2. Therefore, our original model 2 fits the data better.

## Diagnostic

### Multicollinearity

For model 1 (interaction gender\*exper)

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.207354 |
| dept2 | 1.894436 |
| dept3 | 1.775204 |
| dept4 | 2.748729 |
| dept5 | 3.930318 |
| dept6 | 3.038340 |
| clin1 | 1.983056 |
| cert1 | 1.396052 |
| rank2 | 1.103716 |
| rank3 | 1.148014 |

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.301580 |
| dept2 | 1.419295 |
| dept3 | 1.298414 |
| dept4 | 1.400235 |
| dept5 | 2.129589 |
| dept6 | 2.058084 |
| clin1 | 1.563786 |
| cert1 | 1.346859 |
| rank2 | 2.616493 |
| rank3 | 2.728842 |

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.453647 |
| dept2 | 1.689401 |
| dept3 | 1.286818 |
| dept5 | 2.417561 |
| dept6 | 2.590909 |
| clin1 | 2.145632 |
| cert1 | 1.199235 |
| rank3 | 1.258038 |

For model 1, we tested VIF after stratification by experience and found there is no predictors’ vif larger than 5. So none of predictor’s coefficient might be misleading due to collinearity.

For model 2 (interaction gender\*rank)

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.388703 |
| dept2 | 1.788001 |
| dept3 | 1.646600 |
| dept4 | 2.506182 |
| dept5 | 3.759025 |
| dept6 | 3.218099 |
| clin1 | 1.822424 |
| cert1 | 1.387240 |

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.192175 |
| dept2 | 1.494380 |
| dept3 | 1.325727 |
| dept4 | 2.125445 |
| dept5 | 2.805415 |
| dept6 | 2.316121 |
| clin1 | 2.111921 |
| cert1 | 1.413204 |

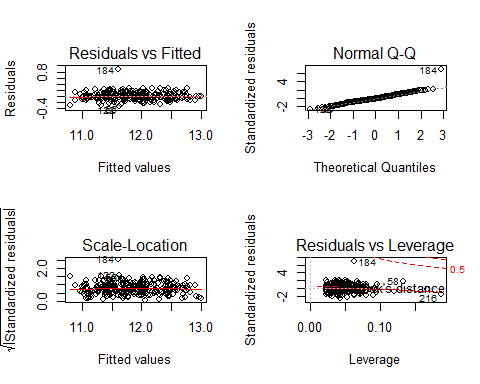
|  |  |
| --- | --- |
|  | x |
| gendermale | 1.211899 |
| dept2 | 1.502562 |
| dept3 | 1.436424 |
| dept4 | 1.235696 |
| dept5 | 1.900869 |
| dept6 | 1.897666 |
| clin1 | 1.661902 |
| cert1 | 1.284494 |

Also, we tested VIF after stratification by rank for model 2, there is no collinearity between predicor of model 2.

### Assumptions

For model 1 (interaction gender\*exper)

par(mfrow = c(2,2))  
plot(reg13)

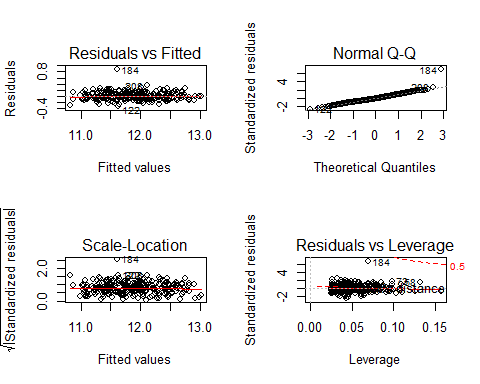


Based on the diagnostic plots for model 1, we can see the constant variance assumptions hold well. The curves in residuals vs fitted plot are horizontal and bounce around 0 which indicates that constant variance is satisfied in our model. Scale-location plot shows the same thing. No.184 observations deviates from the normal line on the qq plot means it might be a outlier. Except from this observation, other observations satisfy normality pretty well.

In addition, the same observation is close to the 0.5 cook’s distance in residuals vs leverage plot, which means No.184 observation might be influential.

For model 2 (interaction gender\*rank)

par(mfrow = c(2,2))  
plot(reg14)



Diagnostic plots for model 2 show similar things, except from No.184 observation, other observations satisfy all of our assumptions very well. It might be influential point in the same time.

### Outlier and influential points

For model 1 (interaction gender\*exper)

According to residuals vs fitted plot and scale-location plot, we can find that No.56, No.122, No. 184 observation are outliers. The No.184 obervation are far away from others.

After deleting no.184 observation, adjusted R-square of model 1 improve from 0.9337 to 0.9458. In addition, several coefficient change. Specifically, coefficient of gender changes a lot (0.129 to 0.098). Then we deleted other outliers(No. 56, No. 122) seperately, adjusted R-square didn’t improve. coefficient of gender changes from 0.098 to 0.123.

To summary, outliers containning No.56, No.122 and No. 184 observation, only No.184 is influential points in the same time, which means there might be other potential preditor we didn’t take into consideration in our model 1.

influential1 %>% knitr::kable()

|  |  |
| --- | --- |
| case | cook.d |
| 56 | 0.0220416 |
| 58 | 0.0403659 |
| 59 | 0.0155546 |
| 73 | 0.0170638 |
| 82 | 0.0217901 |
| 122 | 0.0214374 |
| 135 | 0.0243694 |
| 182 | 0.0160004 |
| 184 | 0.2419279 |
| 216 | 0.0484364 |
| 220 | 0.0233413 |

According to cook’s distance of every observation, we found that No. 56, No. 58, No. 59, No. 73, No. 82, No. 122, No. 135, No. 182, No. 184, No. 216, No. 220 are potential influential points (criterion Di > 4/n).

Among those potential influential points, only deleting No. 184 or No. 216 cause huge change to coefficient in model1 (criterion>10%). So No. 184 and No. 216 are influential points.

For model 2 (interaction gender\*rank)

According to residuals vs fitted plot and scale-location plot, we can find that No. 122, No. 184, No.208 observation are outliers. The No.184 obervation are far away from others.

After deleting no.184 observation, adjusted R-square improve from 0.9322 to 0.9445. In addition, several coefficient change significantly. Specifically, coefficient of gender changes a lot (0.074 to 0.046). Then we deleted other outliers(No. 122, No. 208) seperately, coefficients of predictor don’t changes a lot.

To summary, outliers containning No.122, No.184 and No. 208 observation. Only No. 184 observation is influential points in the same time, which means there might be other potential preditor we didn’t take into consideration in our model 2.

influential2 %>% knitr::kable()

|  |  |
| --- | --- |
| case | cook.d |
| 56 | 0.0221121 |
| 58 | 0.0246957 |
| 59 | 0.0157807 |
| 73 | 0.0257712 |
| 82 | 0.0200849 |
| 122 | 0.0201920 |
| 184 | 0.2459881 |
| 208 | 0.0158925 |

According to influence mesure result, we found that No. 56, No. 58, No. 59, No. 73, No. 82, No. 122, No. 182, No. 184, No. 208 are potential influential points. Except from outliers we deleted above, we delete other potential influential points.

Among those potential influential points, only deleting No. 184 causes huge change to coefficient in model1 (criterion>10%). So No. 184 are influential points.