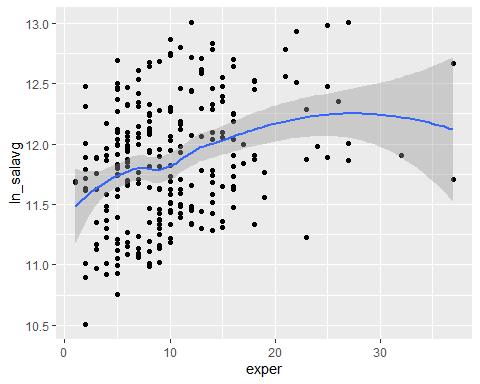
Model diagnostic

Xiaoyang Li

2019/12/8

# Model Diagnostic

## Functional form for continuous predictor



The only continuous variable 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

We added its power 2, 3, 4 to our model and its centering power 2, 3, 4 seperately to create polynomial models. But the summary result indicates that there is no association between higher order of exper and ln\_salavg. Alao, by using anova, we fail to reject our reg13. There is no need to use high order model to improve our reg13.

### Piecewise model

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

P-value of 2 splines is larger than 0.05 in result of piecewise model. Alao, Alao, by using anova, we fail to reject our reg13. So there is no need to use piecewise model to improve our reg13.

## Diagnostic

### Multicollinearity

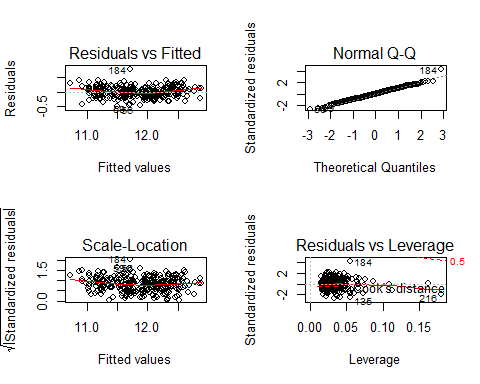
|  |  |
| --- | --- |
|  | x |
| gendermale | 1.207354 |
| factor(dept)2 | 1.894436 |
| factor(dept)3 | 1.775204 |
| factor(dept)4 | 2.748729 |
| factor(dept)5 | 3.930318 |
| factor(dept)6 | 3.038340 |
| clin1 | 1.983056 |
| cert1 | 1.396052 |
| rank2 | 1.103716 |
| rank3 | 1.148014 |

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.301580 |
| factor(dept)2 | 1.419295 |
| factor(dept)3 | 1.298414 |
| factor(dept)4 | 1.400235 |
| factor(dept)5 | 2.129589 |
| factor(dept)6 | 2.058084 |
| clin1 | 1.563786 |
| cert1 | 1.346859 |
| rank2 | 2.616493 |
| rank3 | 2.728842 |

|  |  |
| --- | --- |
|  | x |
| gendermale | 1.453647 |
| factor(dept)2 | 1.689401 |
| factor(dept)3 | 1.286818 |
| factor(dept)5 | 2.417561 |
| factor(dept)6 | 2.590909 |
| clin1 | 2.145632 |
| cert1 | 1.199235 |
| rank3 | 1.258038 |

After stratifying for experience, we tested VIF for 3 layers of experience. According to the vif results above, there is no predictors’ vif larger than 5. So none of predictor’s coefficient might be misleading due to collinearity.

### Assumptions



Based on the diagnostic plots above, 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. Excepr 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.

### Outlier and influential points

According to residuals vs fitted plot and scale-location plot, we can find that No.82, No.122, No. 184 observation are outliers. The No.184 obervation are far away from others.

After deleting no.184 observation, adjusted R-square 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. 82, No. 122), adjusted R-square improve didn’t improve very much. BBut, coefficient of gender changes a lot (0.098 to 0.084).

To summary, outliers containning No.82, No.122 and No. 184 observation. They are influential points in the same time, which means there might be other potential preditor we didn’t take into consideration in our model.

According to influence mesure result, we found that No. 19, No. 44, No. 56, No. 58, No. 59, No. 88, No. 122, No. 125, No. 135, No. 172, No. 184, No. 216, No. 239 are potential influential points. Except from outliers we deleted above, we delete other potential influential points in reg\_i.

After comparing summary results of reg13\_o\_ and reg13\_i, adjusted R-squared improve a little. Coefficients of all predictor don’t change largely. So none of our potential influential observation are really influential points after testing.