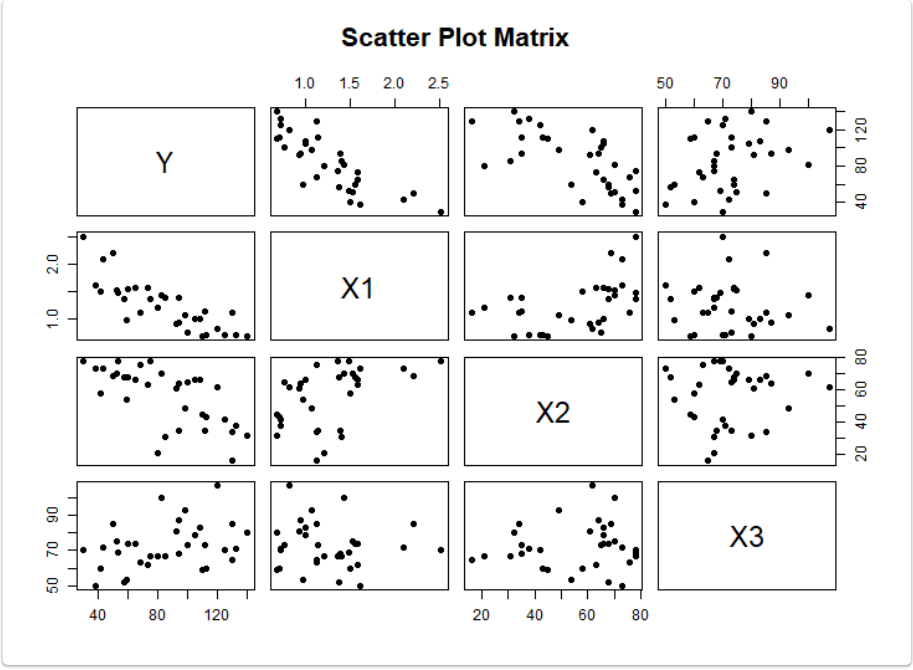
9.15




```r
            X1         X2          X3
X1  1.00000000 0.46773179 -0.08898262
X2  0.46773179 1.00000000  0.06848147
X3 -0.08898262 0.06848147  1.00000000
```

Y 和 X_1, X_2 有比较强的负相关性，不过和 X_3 相关性没有很明显。另一方面， X_i 之间的多重共线性并不严重。 X_1 和 X_2 之间可能有一些 pattern，不过相关系数 0.468 倒也说得过去。

c

```r
lm(formula = Y ~ X1 + X2 + X3, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-28.668  -7.002   1.518   9.905  16.006

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.0473    14.7737   8.126 5.84e-09 ***
X1          -39.9393     5.6000  -7.132 7.55e-08 ***
X2           -0.7368     0.1414  -5.211 1.41e-05 ***
X3            0.7764     0.1719   4.517 9.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.46 on 29 degrees of freedom
Multiple R-squared:  0.8548,    Adjusted R-squared:  0.8398
F-statistic: 56.92 on 3 and 29 DF,  p-value: 2.885e-12
```

从 p 值可以看到，都很显著，初步判断不能删除。

9.16

经过leaps包计算，得到以下最好的三个结果：

```r
Model 1 (Cp = 3.3 ): X1c, X2c, X3c, X1c:X2c
Model 2 (Cp = 3.38 ): X1c, X2c, X3c, I(X3c^2), X1c:X2c
Model 3 (Cp = 4.77 ): X1c, X2c, X3c, I(X2c^2), I(X3c^2), X1c:X2c
```

前两个差不多，都很接近 p。实际上从

```r
  n_vars        Cp     adjR2        R2       BIC                                                          which_vars
4      4  3.302215 0.8615103 0.8788215 -52.16365                                                    X1c, X2c, X3c, X1c:X2c
5      5  3.384990 0.8668497 0.8876545 -51.16476                                          X1c, X2c, X3c, I(X3c^2), X1c:X2c
6      6  4.766392 0.8652362 0.8905044 -48.51620                                X1c, X2c, X3c, I(X2c^2), I(X3c^2), X1c:X2c
7      7  6.344160 0.8623357 0.8924497 -45.61124                      X1c, X2c, X3c, I(X1c^2), I(X2c^2), I(X3c^2), X2c:X3c
```

```
3      3  6.512138 0.8397998 0.8548186 -49.69641                                                                  X1c, X2c, X3c
8      8  8.002543 0.8586982 0.8940236 -42.60122          X1c, X2c, X3c, I(X1c^2), I(X2c^2), I(X3c^2), X1c:X2c, X2c:X3c
9      9 10.000000 0.8525709 0.8940353 -39.10836 X1c, X2c, X3c, I(X1c^2), I(X2c^2), I(X3c^2), X1c:X2c, X1c:X3c, X2c:X3c
2      2 26.683707 0.7361820 0.7526706 -35.61261                                                                 X1c, X2c
1      1 48.509656 0.6313813 0.6429007 -26.98845                                                                      X1c
```

可以看出模型3 的 Cp 也不大。所以没什么区别。

9.19

利用了olsrr 包：

0.15-> Y ~ X1c + X2c + X3c + I(X2c^2) + I(X1c^2)

```
                         Model Summary                                 language-r
----------------------------------------------------------------
R                       0.938       RMSE                10.613
R-Squared               0.880       MSE                112.643
Adj. R-Squared          0.858       Coef. Var           13.770
Pred R-Squared          0.812       AIC                263.549
MAE                     8.092       SBC                274.025
----------------------------------------------------------------
```

0.1 -> 做出来一样

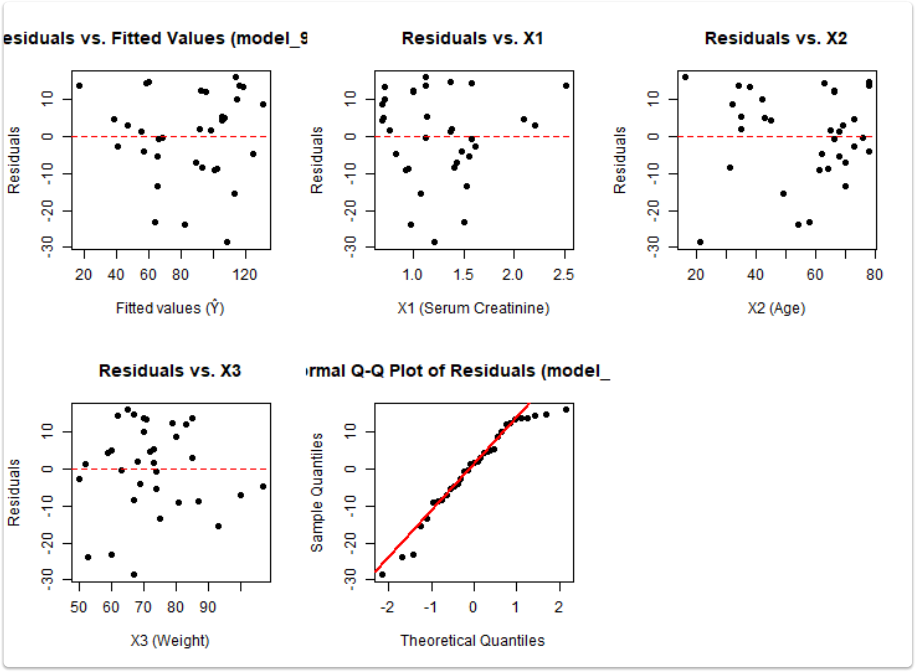
比对发现没有交叉项，此外调整 R^2 也更小。不如之前好。

10.21

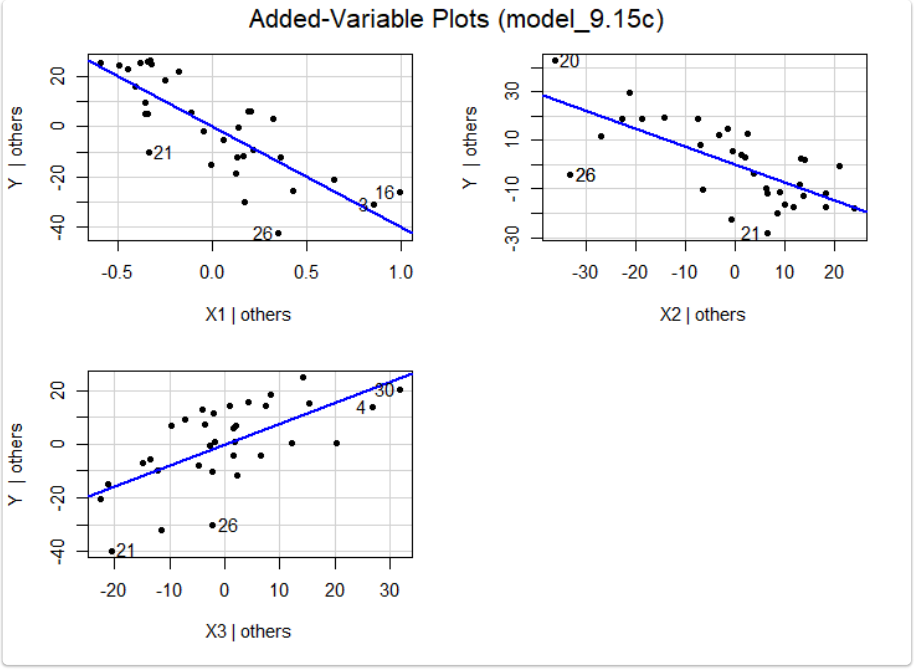
vif一下

```
      X1       X2       X3                                                 language-r
1.304608 1.300377 1.023997
```

没有严重的多重共线性。



残差来看，没有很明显的pattern，倒是 X 的分布有一些左右偏。而qq图看出存在右尾比较厚的情况。



都是在控制其他变量后， X_i 与 Y 有比较明显的相关关系。从第一问的p值也能看出。并无修改必要。

10.22

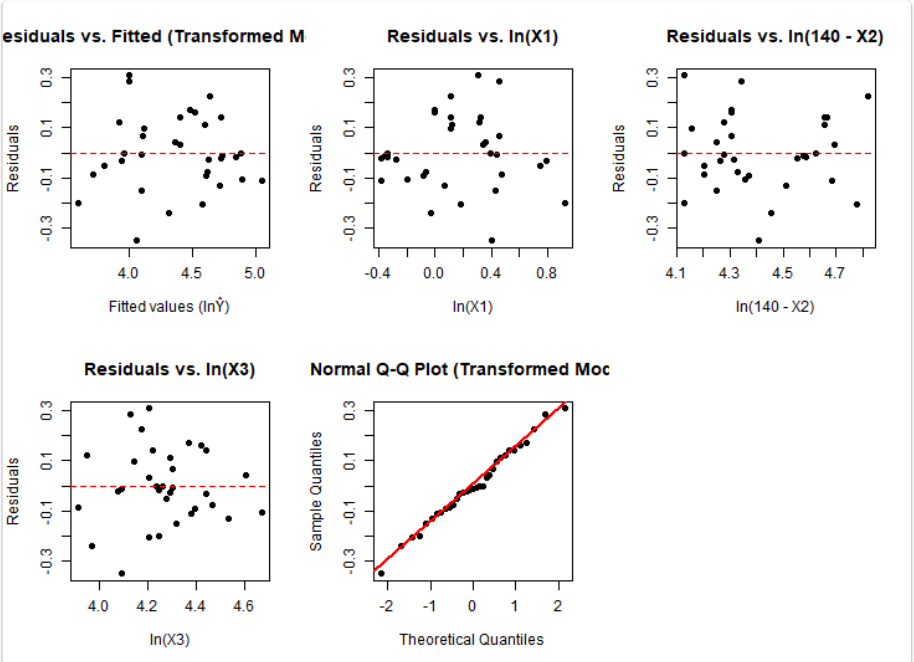
```r
Call:
lm(formula = lnY ~ lnX1 + ln_140_minus_X2 + lnX3, data = data_transformed)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34973 -0.08901 -0.01296  0.11124  0.30979

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.04269    1.01919  -2.004   0.0545 .
lnX1            -0.71195    0.09203  -7.736 1.57e-08 ***
ln_140_minus_X2  0.74736    0.15696   4.761 4.92e-05 ***
lnX3             0.75745    0.15923   4.757 4.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1571 on 29 degrees of freedom
Multiple R-squared:  0.8661,    Adjusted R-squared:  0.8523
F-statistic: 62.54 on 3 and 29 DF,  p-value: 8.963e-13
```

误差小了很多。



看上去很舒服。问题基本得到解决。

```
          lnX1 ln_140_minus_X2            lnX3
      1.339318        1.330109        1.016032
```

没有严重的多重共线性。

```
      rstudent unadjusted p-value Bonferroni p
29 -2.519951273           0.017719      0.58473
```

不能拒绝原假设。没有显著利群值

```
         1          2          3          4          5          6          7          8          9         10         11
0.10121516 0.09260872 0.17595758 0.16677520 0.15649272 0.12733416 0.12471745 0.06960083 0.05446932 0.09517693 0.06701848
        12         13         14         15         16         17         18         19         20         21         22
0.11035144 0.13974330 0.09850386 0.07759886 0.18279037 0.08545999 0.12520107 0.11977313 0.18973903 0.13798308 0.19688693
        23         24         25         26         27         28         29         30         31         32         33
0.05509560 0.14696625 0.13033477 0.16990975 0.11849288 0.10008368 0.07528335 0.23206967 0.05814422 0.06937774 0.14884448
```

这些事 hat 对角元。
利用ai所写

```r
# 经验法则 1: hii > 2p/n
cutoff_leverage1 <- 2 * p_params / n_obs
high_leverage_points1 <- which(hat_values > cutoff_leverage1)
cat(paste("\nHigh leverage points (hii > 2p/n =", round(cutoff_leverage1, 3), "):\n"))
if (length(high_leverage_points1) > 0) {
  print(data.frame(Case = high_leverage_points1, Leverage = hat_values[high_leverage_points1]))
} else {
  cat("None\n")
}
```

跑出来没有outlyiers。

```
   Case     DFFITS
28   28  0.7385345
29   29 -0.7190139
```

```
> print(dfbetas_output_df)
        (Intercept)       lnX1 ln_140_minus_X2       lnX3
Case 28   0.5298824 -0.1505673      -0.5768082 -0.1873520
Case 29  -0.1973071 -0.3099959      -0.1334312  0.4202176
```

```
  Case   Cooks_D
28   28 0.1201784
29   29 0.1091149
```

可以认为这两个点一定程度上比较有高影响，但是还没有达到非常显著的程度。