HW 6

Fangling Zhang

6.3. It is true that adding predictor variables to a regression model can never reduce R^2. It means that model fitting is better. However, it does not imply better model. More variables leads the variance of the model to be larger. Also too many variables will make the model hard to be explained. There should be a balance. We should not always include all available predictor variables in the model.

6.18

(a)

```
  The decimal point is at the |

   0 | 0000000000000000
   2 | 00000000000000000000000
   4 | 00000
   6 | 0
   8 | 0
  10 | 00
  12 | 00000
  14 | 0000000000000
  16 | 0000000000
  18 | 000
  20 | 00
```

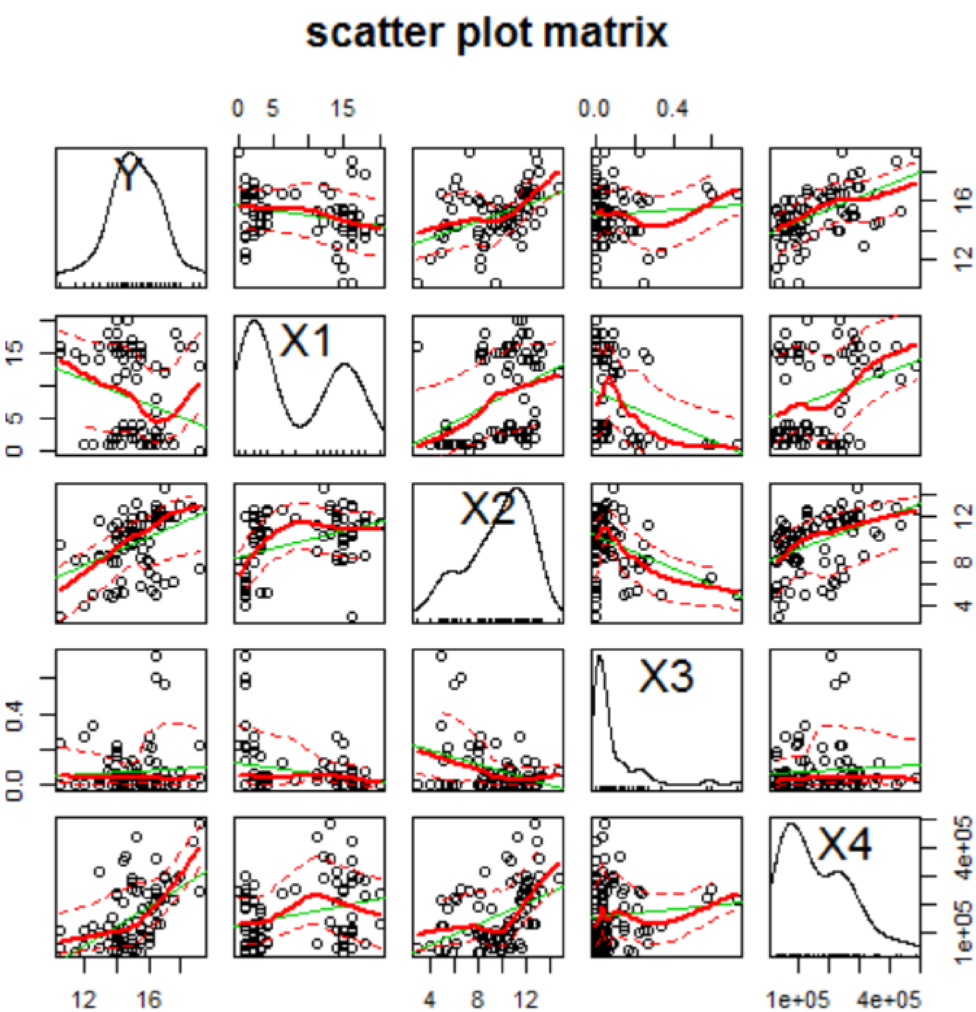
```
  The decimal point is at the |

   2 | 0
   4 | 080003358
   6 | 012613
   8 | 0001223456001555689
  10 | 01334456667777812334466668
  12 | 0001111577788 9002
  14 | 6
```

```
  The decimal point is 1 digit(s) to the left of the |

  0 | 00000000000000000000000000000233333333333444444555555667 8889
  1 | 023444469
  2 | 1223477
  3 | 3
  4 |
  5 | 7
  6 | 0
  7 | 3
```

```
  The decimal point is 5 digit(s) to the right of the |

  0 | 333333444444
  0 | 555666667778899
  1 | 000001111222333334
  1 | 578889
  2 | 011122334444
  2 | 555788899
  3 | 002
  3 | 567
  4 | 23
  4 | 8
```

The stem and leaf plot of X1 indicates that ages of property are frequently located in intervals [0,6] or [12,18]. X2's stem and leaf plot indicates that operating expenses and taxes are mostly located in interval [8,14]. Vacancy rates (X3) are mostly below 0.1. Rental rates (X4) are frequently located in interval [30000, 300000].

(b)




```
> cor(data)
            V1          V2         V3          V4         V5
V1  1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
V2 -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
V3  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
V4  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
V5  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

In the above correlation matrix, V1 represents Y, V2~V5 represents X1~X4 seperately.

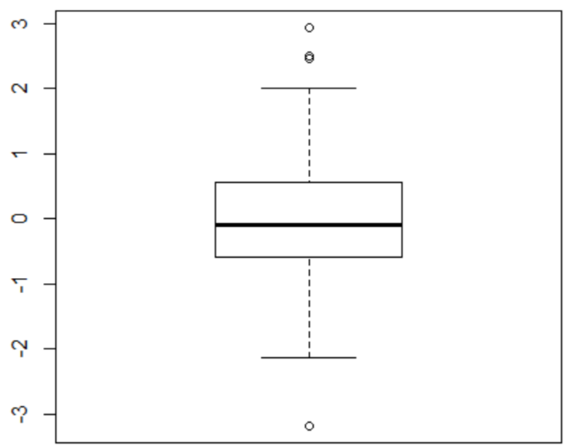
From this two matrix, we can see that Y, X2 and X4 have positive correlation with each other.

(c) Estimated regression function: Y = 12.20 -0.14X1+0.28X2+0.62X3+0.00008X4

The X3 here is not obviously correlated to Y.

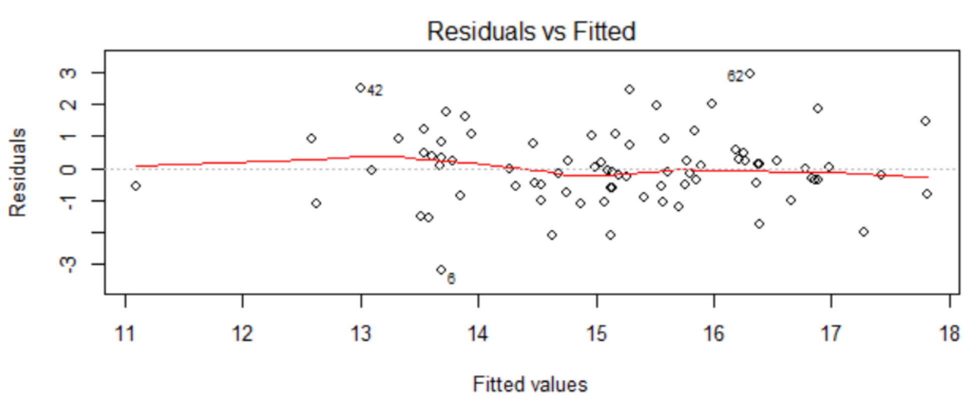
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1872 -0.5911 -0.0910  0.5579  2.9441 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
X3           6.193e-01  1.087e+00   0.570     0.57    
X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

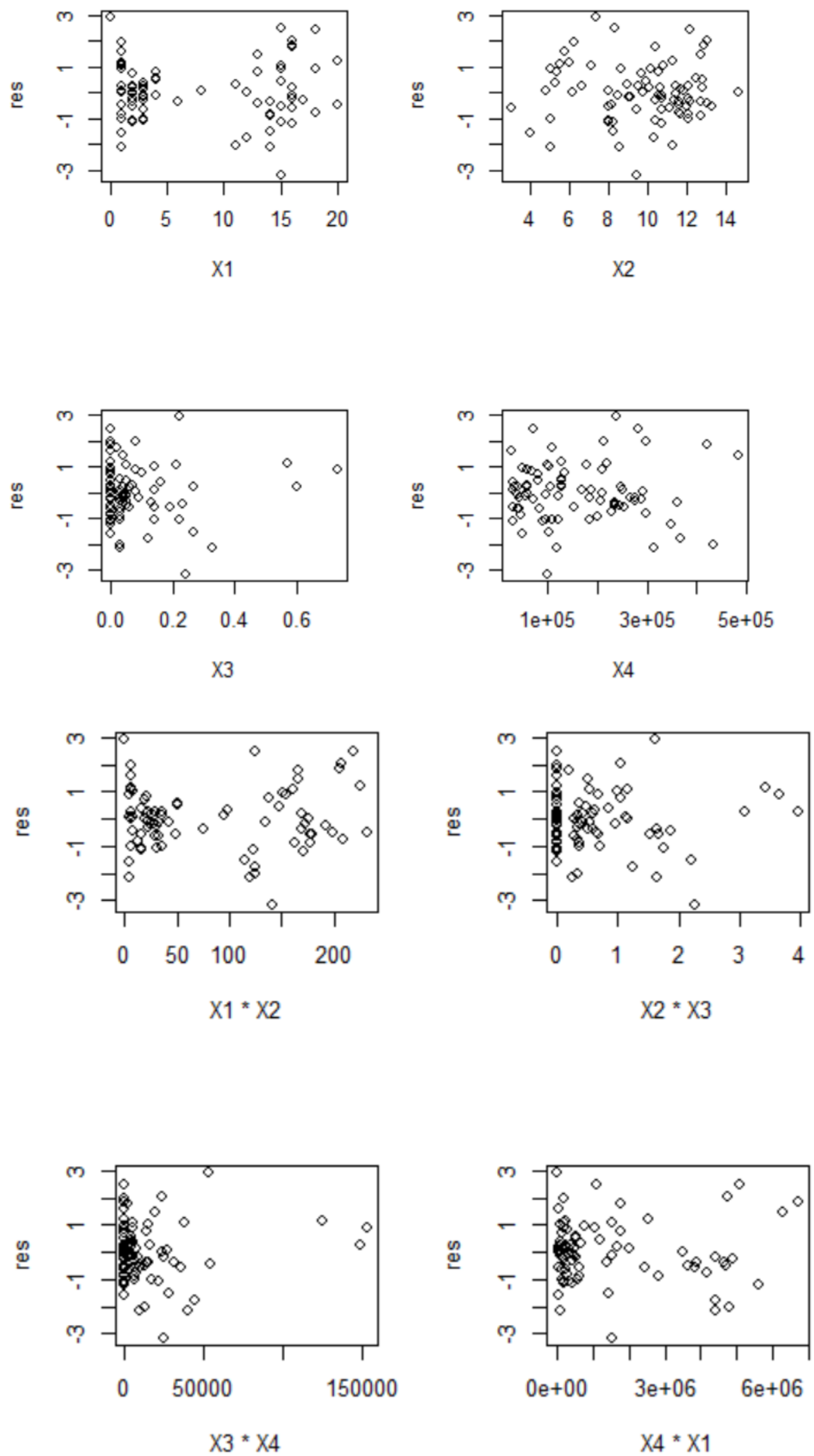
(d)



The box plot of residuals above indicates that the distribution appears to be fairly symmetrical.

(e)





The plots of residuals against all the fitted values, each predictor, each two factors interaction term indicates that there are no obviously positive or negative trends of residuals with these variables.

(f) As we do not have the same values for each of the X variables in this dataset, we do not have a replicate group here. Thus we cannot conduct a formal test for lack of fit here.

19

(a)

The alternatives:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$
$$H_a: \text{not all } \beta_k \ (k = 1, \ldots, p-1) \text{ equal zero}$$

The decision rule:

$$\text{If } F^* \leq F(1-\alpha; p-1, n-p), \text{ conclude } H_0$$
$$\text{If } F^* > F(1-\alpha; p-1, n-p), \text{ conclude } H_a$$

$F^* = \frac{MSR}{MSE}$ =34.582/1.293=26.755

As F (0.95, 4, 76) = 2.49, and here we get F*=26.755>2.49, we conclude Ha, which means that there is a regression relation between the response variable Y and the set of X. That imply that not all β1, β2, β3, β4 equal zero. P-value of the test is 7.272e-14

(b)

The Bonferroni joint confidence intervals can be used to estimate several regression coefficients simultaneously. The confidence limits with family confidence coefficient 1-α are:

$b_k \pm Bs\{b_k\}$, where $B = t(1-\alpha/2g; n-p)$

As α=0.05, g=4, B=t(1-0.05/8;76)=2.339

b_k = -1.420336e-01   2.820165e-01   6.193435e-01   7.924302e-06

$\underset{p\times p}{\mathbf{s}^2\{\mathbf{b}\}} = MSE(\mathbf{X}'\mathbf{X})^{-1}$ =

	X0	X1	X2	X3	X4
X0	3.340347e-01	-3.939727e-04	-3.244950e-02	-3.242039e-01	1.596968e-07
X1	-3.939727e-04	4.555089e-04	-2.706849e-04	4.064096e-03	-5.571354e-09
X2	-3.244950e-02	-2.706849e-04	3.990762e-03	2.831991e-02	-3.970806e-08
X3	-3.242039e-01	4.064096e-03	2.831991e-02	1.181167e+00	-4.842359e-07
X4	1.596968e-07	-5.571354e-09	-3.970806e-08	-4.842359e-07	1.917611e-12

$s\{b_k\}$ =

2.134265e-02 6.317248e-02 1.086815e+00 1.384778e-06

The confidence intervals of $\beta_1, \beta_2, \beta_3$, and β_4 jointly are [-0.197, -0.087], [0.120, 0.444], [-2.161, 3.400], [4.381e-6, 1.127e-5].

The probability that $\beta_1, \beta_2, \beta_3$, and β_4 are all in these four intervals is 95%.

(c)

$R^2 = \frac{SSR}{SSTO}$ =138.237/236.558=0.5847

That means when the four predictor variables, age(X1), operating expenses and taxes (X2), vacancy rates(X3), total square footage(X4), are considered, the variation in properties rental rates(Y) is reduced by 58.74 percent.

20.

$\mathbf{X}'_h$ =

```
     x0 V1    V2   V3     V4
[1,]  1  5  8.25 0.00 250000
[2,]  1  6  8.50 0.23 270000
[3,]  1 14 11.50 0.11 300000
[4,]  1 12 10.25 0.00 310000
```

b =
```
  (Intercept)            X1            X2            X3            X4
 1.220059e+01 -1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06
```

$\hat{Y}_h = \mathbf{X}'_h \mathbf{b} =$
```
[1,] 15.79813
[2,] 16.02754
[3,] 15.90072
[4,] 15.84339
```

$s^2\{\hat{Y}_h\} = \mathbf{X}'_h \mathbf{s}^2\{\mathbf{b}\}\mathbf{X}_h$ = 0.07733061 0.05566105 0.04935496 0.06714763

$s\{\hat{Y}_h\}$ = 0.2780838 0.2359259 0.2221598 0.2591286

If we use the Working-Hotelling method, then

$W^2 = pF(1-\alpha; p, n-p)$ =5*F(1-0.05;5,81-5)=11.675, so W=3.417.

If we use Bonfessoni simultaneous confidence intervals, then

$B = t(1-\alpha/2g; n-p)$ =t(1-0.05/10;81-5)=2.642.

We can see that B<W here, so at last, we use Bonferroni simultaneous confidence intervals. The Bonferroni confidence limits are: $\hat{Y}_h \pm Bs\{\hat{Y}_h\}$

The simultaneous interval estimates of the mean rates for four typical properties are as follows:

```
[1,] 15.06341 16.53285
[2,] 15.40420 16.65087
[3,] 15.31376 16.48769
[4,] 15.15875 16.52802
```

21.

To predict intervals for the rental rates of these 3 properties separately, the 1-α prediction limits for a new observation $Y_{h(\text{new})}$ corresponding to $\mathbf{X}_h$ are:

$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{\text{pred}\}$, where $s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} = MSE(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$

$\hat{Y}_h$ =
```
15.14850
15.54249
16.91384
```

$s^2\{\text{pred}\}$ = `1.328955 1.330628 1.426951`

t(1-0.05/2;81-5)=1.992

The separate prediction interval for the rates of these 3 properties are as follows:

```
[1,] 12.85249 17.44450
[2,] 13.24504 17.83994
[3,] 14.53469 19.29299
```

As these three properties are predicted separately, the predictions can hardly be fairly precisely.

The family confidence level for the set of three predictions is 0.95*0.95*0.95=0.857

6.1 (a)

$$X = \begin{bmatrix} 1 & X_{11} & X_{11}X_{12} \\ 1 & X_{21} & X_{21}X_{22} \\ 1 & X_{31} & X_{31}X_{32} \\ 1 & X_{41} & X_{41}X_{42} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

6.26.

The correlation coefficient between the observed values y_i and the fitted values $\hat{y}_i$

$$r^2_{y.\hat{y}} = \left(\frac{\text{cov}(y.\hat{y})}{\sqrt{\text{var}(y)\text{var}(\hat{y})}}\right)^2 = \frac{[\text{cov}(y.\hat{y})]^2}{\text{var}(y)\,\text{var}(\hat{y})}$$

$$\because \text{cov}(y,\hat{y}) = \text{cov}(\hat{y}+e,\hat{y}) = \text{cov}(\hat{y}.\hat{y}) + \text{cov}(e,\hat{y})$$
$$= \text{cov}(\hat{y}.\hat{y}) = \text{var}(\hat{y})$$

$$\therefore r^2_{y.\hat{y}} = \frac{(\text{var}(\hat{y}))^2}{\text{var}(y)\,\text{var}(\hat{y})} = \frac{\text{var}(\hat{y})}{\text{var}\,y} = \frac{SSR}{SSTO} = R^2$$