STA206 hw4

Zhen Zhang

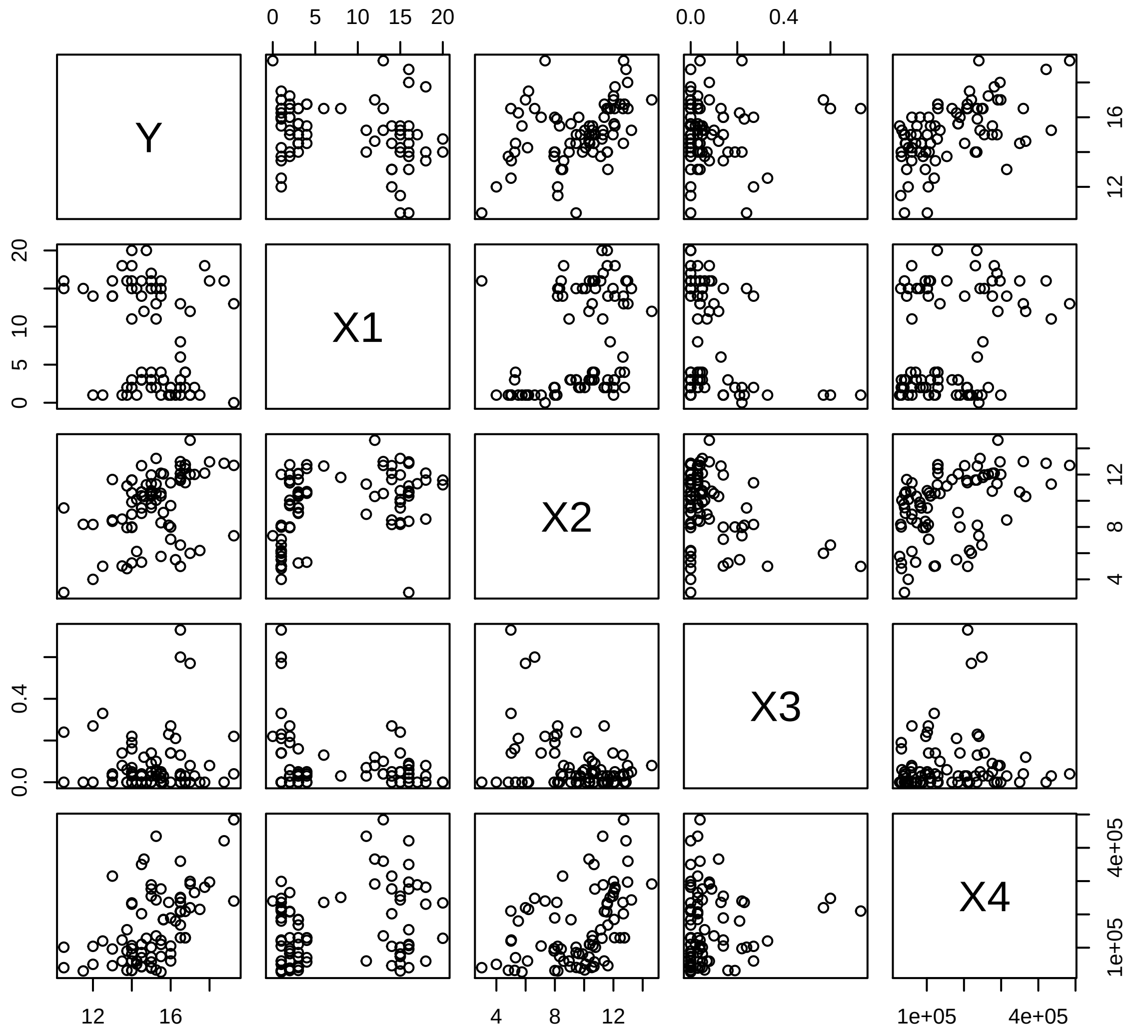
October 23, 2015

(a) First read the data:

```
property <- read.table("property.txt")
property[, 5] <- as.double(property[, 5])
colnames(property) <- c("Y", paste0("X", 1:4))
```

Draw the scatterplot matrix:

```
pairs(property)
```



and the correlation matrix:

```r
cor(property)
```

```
##              Y         X1         X2          X3         X4
## Y   1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## X1 -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## X2  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## X3  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## X4  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

I can see there is a modest realtionship between Y and X_2 or X_4, Y and X_1 has a negative relationship.

(b) The regression:

```r
fit1 <- lm(Y ~ ., data = property)
```

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = Y ~ ., data = property)
##
## 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
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

So the least square estimators are: β_0 = 12.201, β_1 = -0.142, β_2 = 0.282, β_3 = 0.619, β_4 = 0.000.

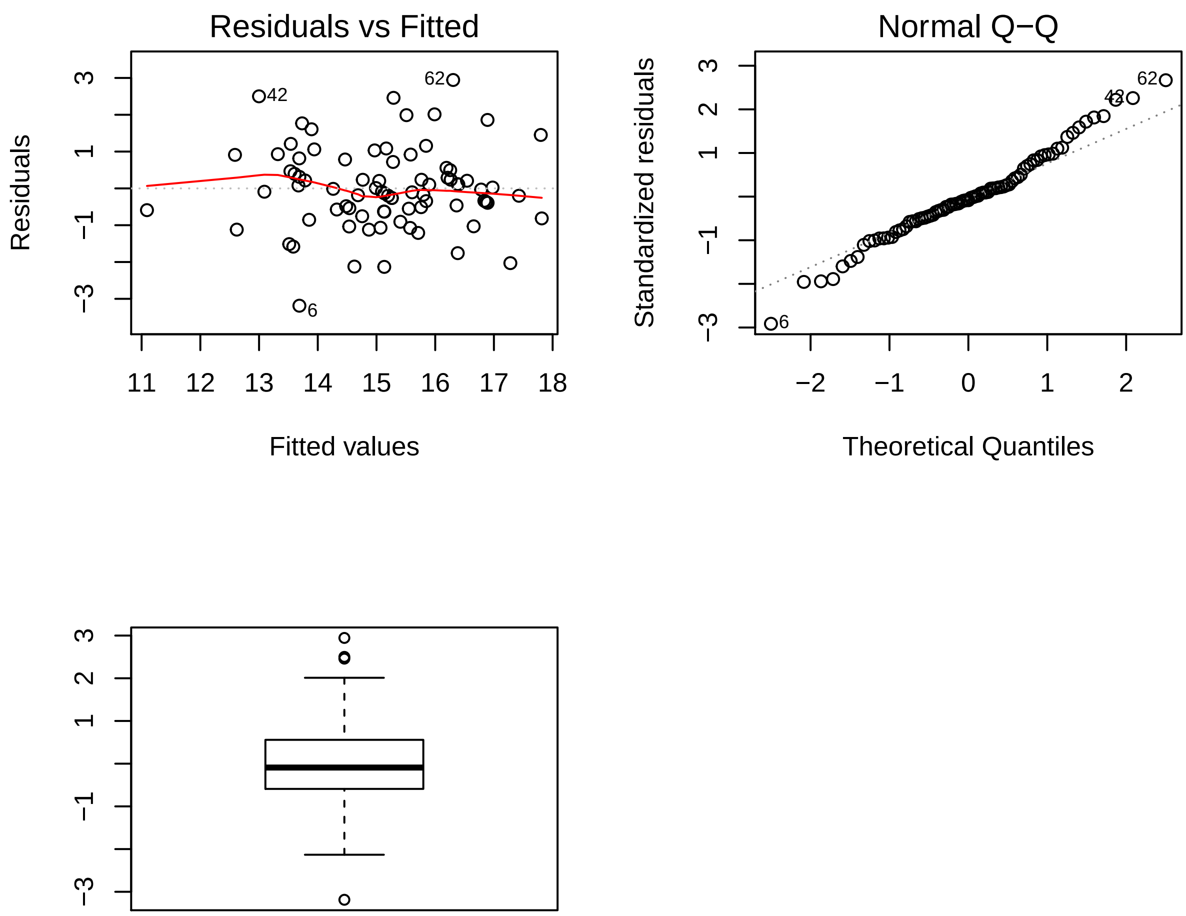
The fitted regression function:

$$Y = 12.201 - 0.142X_1 + 0.282X_2 + 0.619X_3 + 0.000X_4$$

$MSE = 1.136885^2 = 1.293$, $R^2 = 0.585$, $R_a^2 = 0.563$.

(c) Residuals vs fitted value plot:

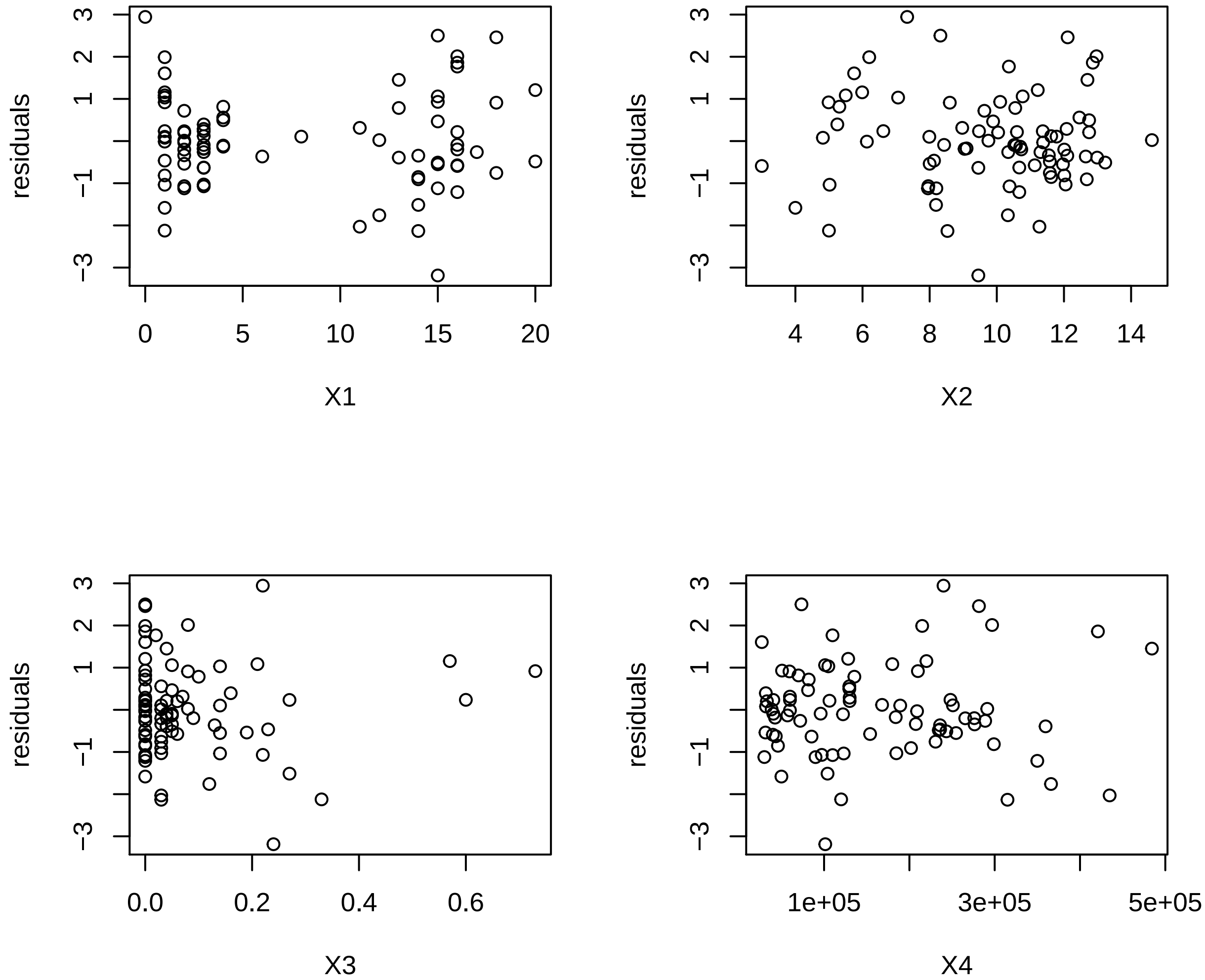
```r
opar <- par()
par(mfrow = c(2, 2))
plot(fit1, which = 1)
plot(fit1, which = 2)
boxplot(fit1$residuals)
```



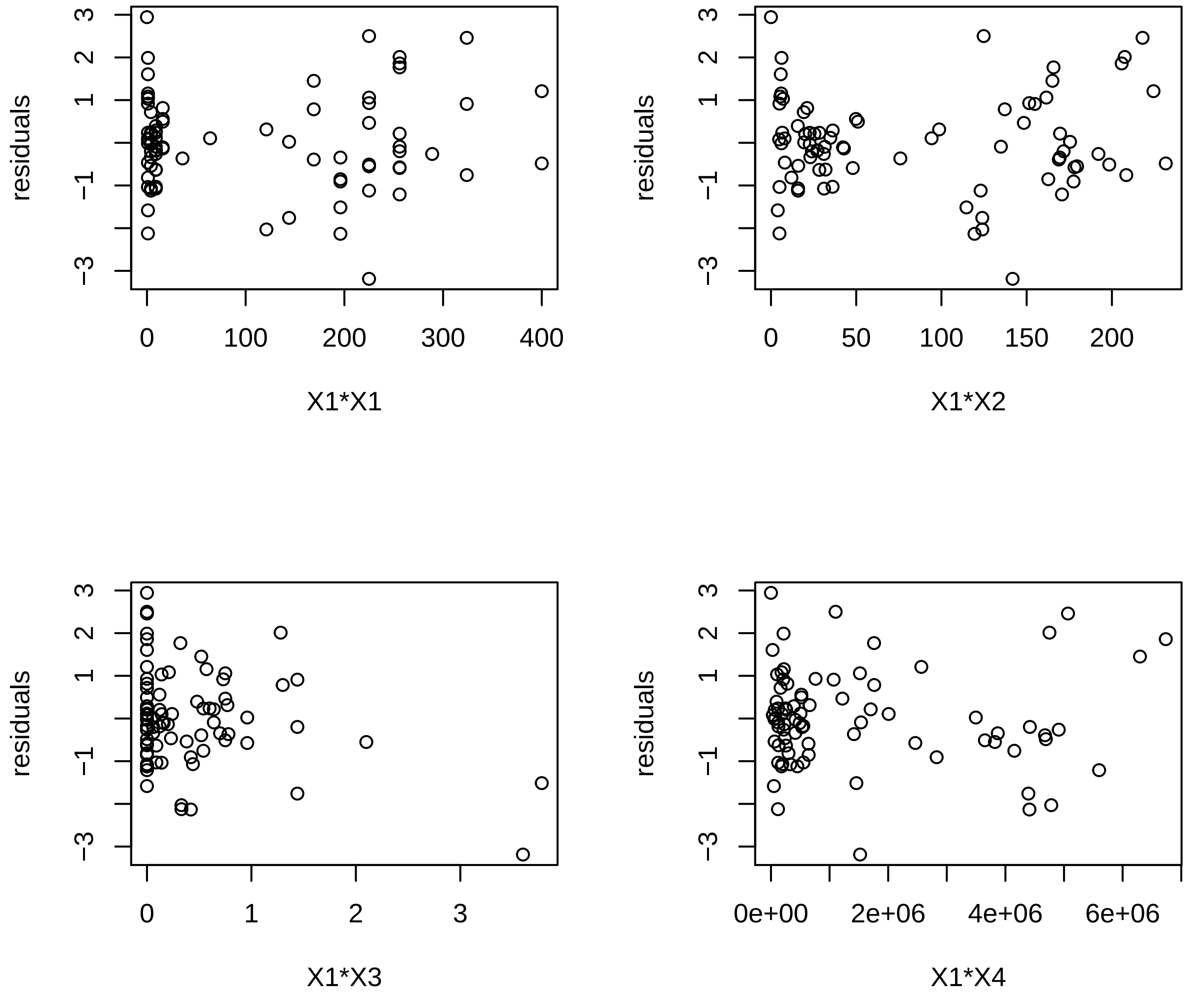
We can see the linear relationship is valid, but the qqplot is heavy tailed and there are some residulas not caught by the model since there are some outliers.

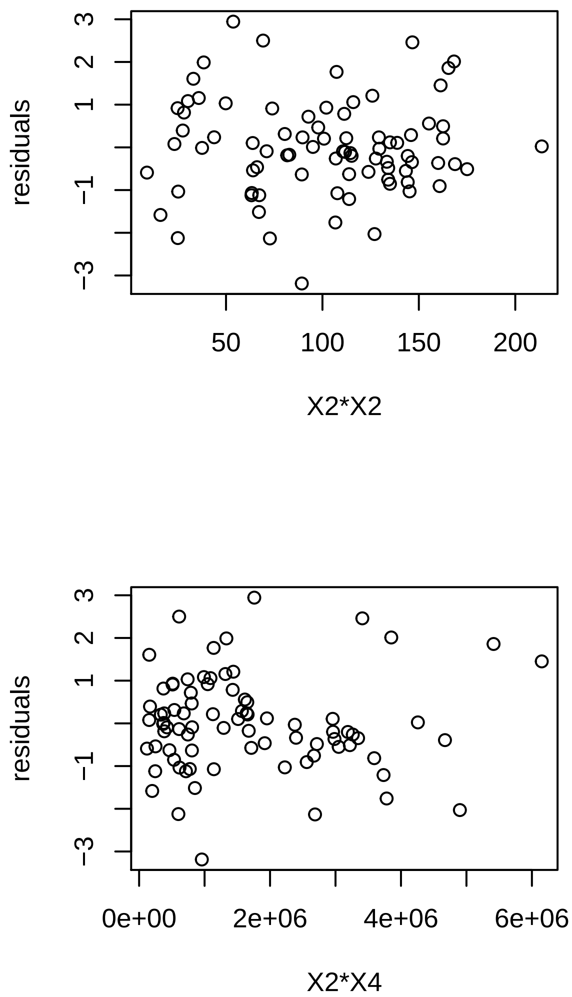
(d)

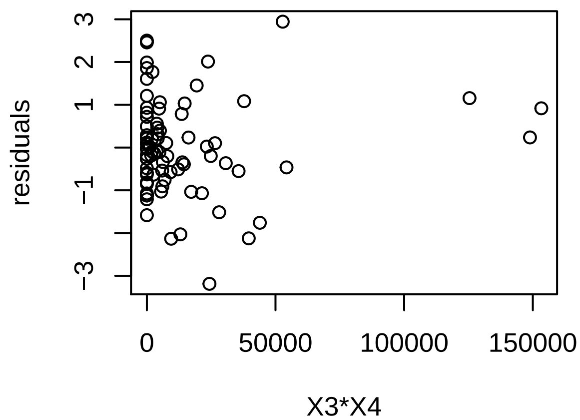
```r
par(mfrow = c(2, 2))
lapply(1:4, function(x) {
    plot(property[, paste0("X", x)], fit1$residuals, xlab = paste0("X",
        x), ylab = "residuals")
})
```

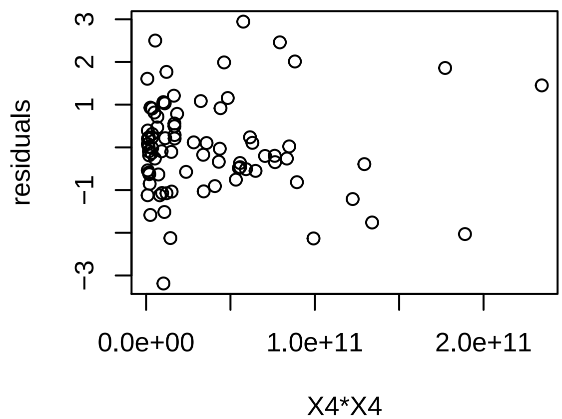



```r
par(mfrow = c(2, 2))
lapply(1:4, function(i) {
    lapply(i:4, function(j) {
        plot(property[, paste0("X", i)] * property[, paste0("X", j)], fit1$residuals,
            xlab = paste0("X", i, "*", "X", j), ylab = "residuals")
    })
})
```


residuals
X1*X1
X1*X2
X1*X3
X1*X4
3
2
1
−1
−3
0
100
200
300
400
50
150
0e+00
2e+06
4e+06
6e+06


residuals
X2*X2
X2*X3
X2*X4
X3*X3





There are ten interaction terms in total.

Interpretation: In all of the ten interaction terms, whenever X_3 is present, the points in the plot will be a little wired: the points should scatter, but now much of the points concentrate together near $x = 0$. So it suggests X_3 is uncorrelated with Y, should be excluded from the model.

(e) For β_1:

Null hypothesis (H_0): There is no relationship between Y and X_1 ($\beta_1 = 0$)

Alternative hypothesis (H_1): There is a relationship between Y and X_1 ($\beta_1 ! = 0$)

Test statistic: T-statistic: $T^* = \frac{\hat{\beta_1}}{se(\hat{\beta_1})}$ = -6.65493

Null distribution: Under H_0, $\beta_1 = 0$, $T^* \sim t(76)$

p-value: 0.0000000038943

For β_2:

Null hypothesis (H_0): There is no relationship between Y and X_2 ($\beta_2 = 0$)

Alternative hypothesis (H_1): There is a relationship between Y and X_2 ($\beta_2 ! = 0$)

Test statistic: T-statistic: $T^* = \frac{\hat{\beta_2}}{se(\hat{\beta_2})} = 4.46424$

Null distribution: Under H_0, $\beta_2 = 0$, $T^* \sim t(76)$

p-value: 0.0000274739604

For β_3:

Null hypothesis (H_0): There is no relationship between Y and X_3 ($\beta_3 = 0$)

Alternative hypothesis (H_1): There is a relationship between Y and X_3 ($\beta_3 ! = 0$)

Test statistic: T-statistic: $T^* = \frac{\hat{\beta_3}}{se(\hat{\beta_3})} = 0.56987$

Null distribution: Under H_0, $\beta_3 = 0$, $T^* \sim t(76)$

p-value: 0.57045

For β_4:

Null hypothesis (H_0): There is no relationship between Y and X_4 ($\beta_4 = 0$)

Alternative hypothesis (H_1): There is a relationship between Y and X_4 ($\beta_4 ! = 0$)

Test statistic: T-statistic: $T^* = \frac{\hat{\beta_4}}{se(\hat{\beta_4})} = 5.72245$

Null distribution: Under H_0, $\beta_4 = 0$, $T^* \sim t(76)$

p-value: 0.0000001975990

Conclusion: X_1, X_2 and X_4 are significant, while X_3 is not significant. It is consistent with the correlation table result.

(f)

```r
(fit1_anova <- anova(fit1))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 14.819  14.819 11.4649  0.001125 **
## X2         1 72.802  72.802 56.3262 9.699e-11 ***
## X3         1  8.381   8.381  6.4846  0.012904 *
## X4         1 42.325  42.325 32.7464 1.976e-07 ***
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the anova table.

So $SSE = 98.230594$, dof $= 76$

$SSR = 14.818520 + 72.802011 + 8.381417 + 42.324958 = 138.326906$, dof $= 4$

$SSTO = SSE + SSR = 236.5575$, dof $= 80$

Null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

Alternative hypothesis H_1: At least one $\beta_i, (i = 1, 2, 3, 4)$ is not 0.

Test statistic: F-test: $F^* = \frac{SSR/4}{SSE/76} = 26.7555$.

Null distribution: Under H_0, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, $F^* \sim F_{0.99,4,76}$

Decision rule: if $F^* > F_{0.99,4,76}$, reject H_0, meaning at least one $\beta_i, (i = 1, 2, 3, 4)$ is not 0

Conclusion: since $F_{0.99,4,76} = 3.576520071$, so $F^* > F_{0.99,4,76}$, reject H_0, so at least one $\beta_i, (i = 1, 2, 3, 4)$ is not 0.

(g) Since β_3 is not significant, so I decide to exclude X_3 from the model:

```
fit2 <- lm(Y ~ . - X3, data = property)
summary(fit2)
```

```
##
## Call:
## lm(formula = Y ~ . - X3, data = property)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0620 -0.6437 -0.1013  0.5672  2.9583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
## X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
## X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.132 on 77 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5667
## F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

So the least square estimators are: $\beta_0 = 12.371$, $\beta_1 =$ -0.144, $\beta_2 = 0.267$, $\beta_4 = 0.000$.

The fitted regression function:

$$Y = 12.371 - 0.144X_1 + 0.267X_2 + 0.000X_4$$

$MSE = 1.131889^2 = 1.281$, $R^2 = 0.583$, $R_a^2 = 0.567$.

The MSE and R_a^2 is smaller than model 1.

(h) The standard errors for model 2 are:

$sd(\beta_1) = 0.021$, $sd(\beta_2) = 0.057$, $sd(\beta_4) = 0.000$

The standard errors for model 2 are:

$sd(\beta_1) = 0.021$, $sd(\beta_2) = 0.063$, $sd(\beta_3) = 1.087$, $sd(\beta_4) = 0.000$

So it is a little smaller for β_2.

The confident intervals for both models:

```r
lapply(list(fit1 = fit1, fit2 = fit2), confint)
```

```
## $fit1
##                     2.5 %        97.5 %
## (Intercept)  1.104949e+01  1.335169e+01
## X1          -1.845411e-01 -9.952615e-02
## X2           1.561979e-01  4.078352e-01
## X3          -1.545232e+00  2.783919e+00
## X4           5.166283e-06  1.068232e-05
##
## $fit2
##                     2.5 %        97.5 %
## (Intercept)  1.138920e+01  1.335197e+01
## X1          -1.858219e-01 -1.025074e-01
## X2           1.530784e-01  3.812557e-01
## X4           5.578873e-06  1.077755e-05
```

```r
sapply(lapply(list(fit1 = fit1, fit2 = fit2), confint), function(x) {
    sapply(1:dim(x)[1], function(y) x[y + dim(x)[1]] - x[y])
})
```

```
## $fit1
## [1] 2.302199e+00 8.501498e-02 2.516373e-01 4.329151e+00 5.516038e-06
##
## $fit2
## [1] 1.962767e+00 8.331458e-02 2.281773e-01 5.198675e-06
```

The confident intervals for X_1, X_2 and X_4 are wider in model 1. There is no X_3 in the second model, so we cannot compare it to the first model.

(i) The prediction interval is:

```r
predictions <- lapply(list(fit1 = fit1, fit2 = fit2), function(x) {
    predict(x, data.frame(X1 = 4, X2 = 10, X3 = 0.1, X4 = 80000), level = 0.99,
        interval = "prediction")
})
predictions
```

```
## $fit1
##       fit     lwr      upr
## 1 15.1485 12.1027 18.19429
##
## $fit2
##        fit      lwr      upr
## 1 15.11985 12.09134 18.14836
```

The prediction interval for model 2 is smaller than that of model 1.

(j) I would prefer model 2, since it excludes an insignificant variable, and gets a more larger R_a^2, more narrower prediction interval.