Final

Zeqiu.Yu

2022-12-05

Final project

Part 1

1. The pairwise correlations are small, it doesn't mean there are no interaction between predictor variables. In addition, the existence of the interection term doesn't depend on the correlation between the predictor variables. Hence, We can't just conclude that the predictors do not interact with each other.

2. We will start from M_1. It means $X_2 = X_3 = X_4 = 0$. There are same response functions, when $X_2 = 1, X_3 = X_4 = 0\ldots$ Hence, testing if the response function is the same for all four tool models is to test $\beta_2, \beta_3, \beta_4$.
   Let
   $$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$
   , $H_a : \beta_2, \beta_3, \beta_4$ at least one of them is not zero.

3. When $X_3 = 1, Y = \beta_0 + \beta_1 X_1 + \beta_3 + \epsilon$,
   When $X_4 = 1, Y = \beta_0 + \beta_1 X_1 + \beta_4 + \epsilon$. If they are the same, it is to test:
   $$H_0 : \beta_3 = \beta_4$$
   , $H_a : \beta_3 \neq \beta_4$.

4. Yes.
   X_1 and X_2, X_3 may correlated. When he does regress between X_1 and X_2, X_3, it is to exclude the effect of X_2, X_3 on X_1. In the same way, When he does regression between Y and X_2, X_3, it is the Y without the effect of X_2, X_3. Hence, when he plots Residual1 VS Residual2, it is the true relationship between Y(under the effect of $X_1 1$) and X_1 itself. The relationship is linear and therefore he should include the first order.

Part 2

```
input1 <- read.table('./HomeSales.txt')
names(input1) <- c("IdNum","SalesPrice", "FSquaredFeet","NumBedrooms","NumBathrooms","AC",
                   "GarageSize","Pool","Year","Quality","Style","LotSize","AdHighway")
input1[1:5,]
```

```
##   IdNum SalesPrice FSquaredFeet NumBedrooms NumBathrooms AC GarageSize Pool
## 1     1     360000         3032           4            4  1          2    0
## 2     2     340000         2058           4            2  1          2    0
## 3     3     250000         1780           4            3  1          2    0
## 4     4     205500         1638           4            2  1          2    0
## 5     5     275500         2196           4            3  1          2    0
```

```
##   Year Quality Style LotSize AdHighway
## 1 1972       2     1   22221         0
## 2 1976       2     1   22912         0
## 3 1980       2     1   21345         0
## 4 1963       2     1   17342         0
## 5 1968       2     7   21786         0
```

1.)

```r
fit1 <- lm(SalesPrice~FSquaredFeet * Pool, data = input1)
summary(fit1)
```

```
##
## Call:
## lm(formula = SalesPrice ~ FSquaredFeet * Pool, data = input1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -247193  -40579   -7542   24476  384051
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -88538.996  12063.237  -7.340 8.34e-13 ***
## FSquaredFeet         161.910      5.168  31.331  < 2e-16 ***
## Pool              105909.972  47262.735   2.241   0.0255 *
## FSquaredFeet:Pool    -37.213     17.102  -2.176   0.0300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78890 on 518 degrees of freedom
## Multiple R-squared:  0.6747, Adjusted R-squared:  0.6728
## F-statistic: 358.1 on 3 and 518 DF,  p-value: < 2.2e-16
```

Hence, the statistical model is: E[Y_i] $= \beta_0 + \beta_1 X_{3i} + \beta_2 X_{8i} + \beta_3 X_{3i} X_{8i}$
The regression function is : $\hat{Y} = -88538.996 + 161.910 X_3 + 105909.972 X_8 - 37.213 X_3 X_8$.
When there is a pool, 1 unit change in X_3 (Finished Square Foot) will cause 124.697 increase in the response variable (Sales price).
When there is no pool, 1 unit change in X_3 (Finished Square Foot) will cause 161.910 increase in the response variable (Sales price).

(2.)

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$H_a : \beta_1, \beta_2, \beta_3$ at least one of them is not zero.
Decision Rule: Conclusion: when the p-value is larger than $\alpha = 0.05$, we conclude H_0, there is no regression relationship. Otherwise, we conclude H_a.

p-value is 2.2e-16 according to the summary table, we conclude H_a.

(3.)
It is to test the coeffeicient of the interaction term is 0 or not.
For the statistical model mentioned in (1.).

$$H_0 : \beta_3 = 0, \quad H_a : \beta_3 \neq 0$$

```r
fit2 <- lm(SalesPrice~ FSquaredFeet + Pool, data = input1)
anova(fit2, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: SalesPrice ~ FSquaredFeet + Pool
## Model 2: SalesPrice ~ FSquaredFeet * Pool
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1    519 3.2536e+12
## 2    518 3.2241e+12  1 2.9469e+10 4.7347 0.03001 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

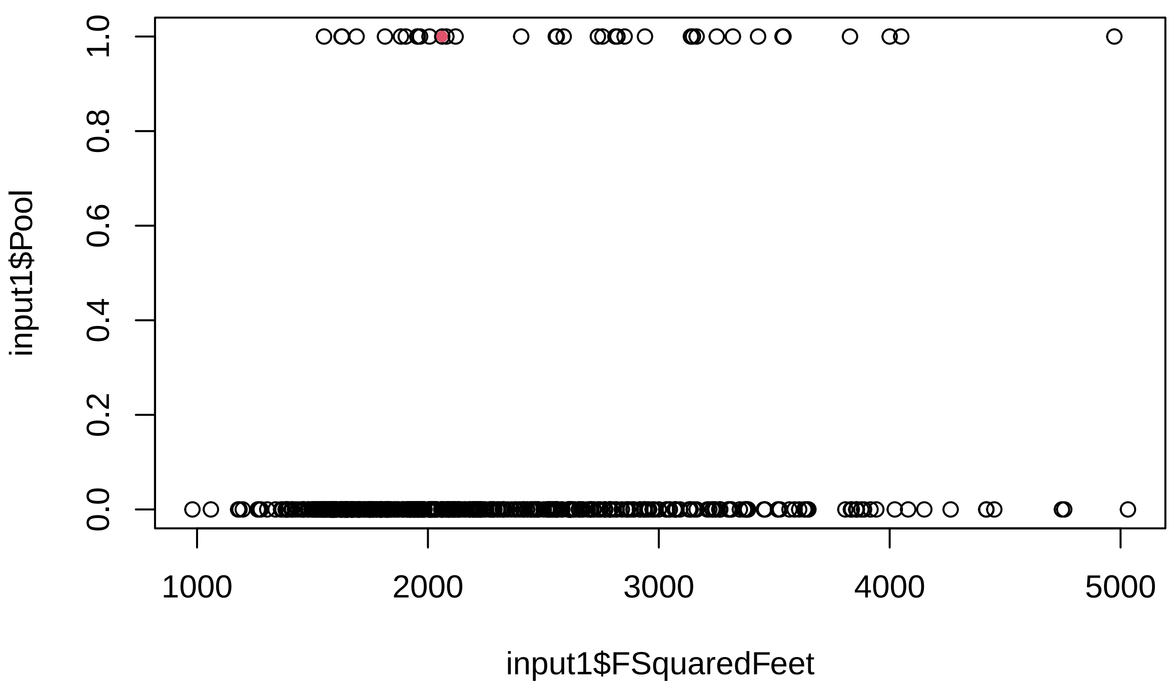
The p-value is 0.03001, which is smaller than $\alpha = 0.05$, which means we conclude H_a, we should include the interaction.

(4.)

```r
alpha = 0.05
new = data.frame(FSquaredFeet = 2061, Pool = 1)
predict(fit1, new, interval = "prediction", level = 1-alpha)
```

```
##        fit      lwr      upr
## 1 274371.7 115980.2 432763.1
```

```r
plot(input1$FSquaredFeet, input1$Pool)
points(2061,1, col= 2, pch =20)
```




```r
X.new <- c(1, 2061, 1)
n <- dim(input1)[1]
X <- cbind(rep(1,n), input1$FSquaredFeet, input1$Pool)
t(X.new)%*%solve(t(X)%*%X)%*%X.new
```

```
##            [,1]
## [1,] 0.02929351
```

```r
summary(hatvalues(fit1))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.002058 0.002407 0.003241 0.007663 0.004596 0.251416
```

0.02929351 is within the range of the leverage levels, we do not need to do extrapolation.

(5.)

```r
par(mfrow=c(1,1))
n = dim(input1)[1]
p = 4
crit <- qt(1-alpha/2/n, n-p-1)
which(abs(rstudent(fit1)) >= crit)
```

```
## 73 80 96
## 73 80 96
```

Yes. there are usually high sales price. The houses with identification number 73, 80, 96.

(6.)

```r
which.max(pf(cooks.distance(fit1), p, n-p))
```

```
## 104
## 104
```

```r
pf(cooks.distance(fit1)[c(104)], p, n-p)
```

```
##        104
## 0.03889433
```

Its CD is below 20%, it indicates little influence on the fitted values.

(7.)
Let air_conditioning, pool, quality, Style and Adjacent_to_highway be qualitative variables.

```r
library(leaps)
fit.Full <- lm(SalesPrice~ FSquaredFeet + NumBedrooms + NumBathrooms + as.factor(AC) +
                 GarageSize + as.factor(Pool) + Year + as.factor(Quality) + as.factor(Style) +
                 LotSize + as.factor(AdHighway), input1)
Base <- lm(SalesPrice~1, data = input1)
step(fit.Full, scope = list(upper = fit.Full, lower = Base), direction = "both", trace = FALSE)
```

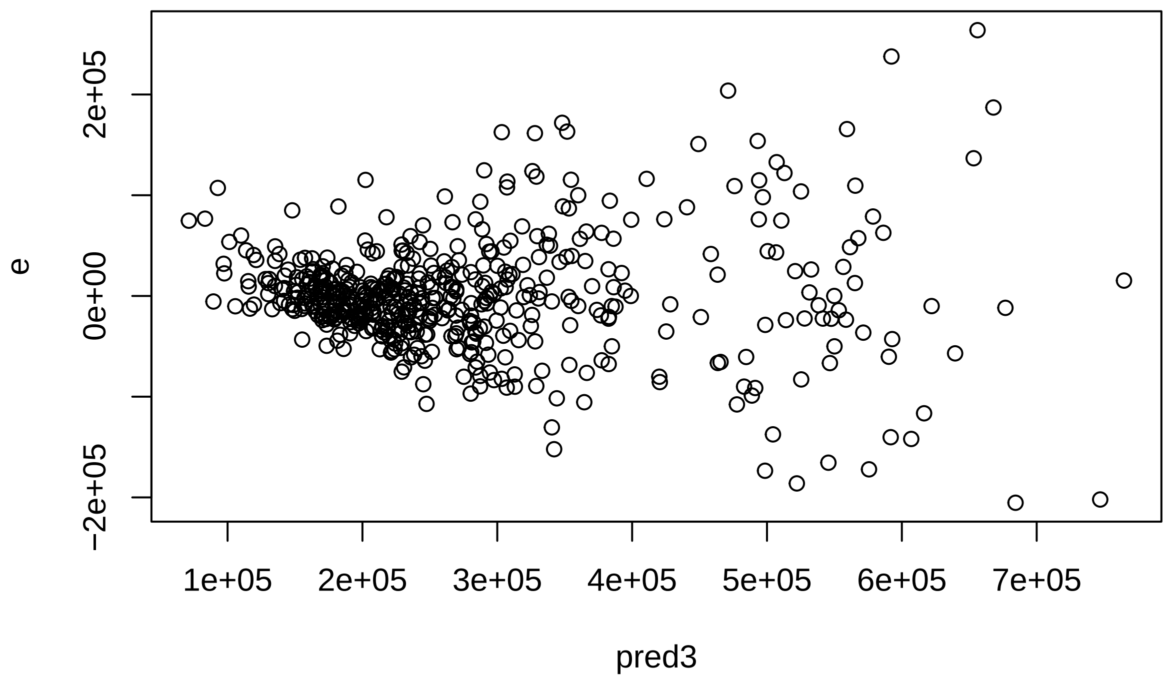
```
##
## Call:
## lm(formula = SalesPrice ~ FSquaredFeet + NumBathrooms + GarageSize +
##     Year + as.factor(Quality) + as.factor(Style) + LotSize +
##     as.factor(AdHighway), data = input1)
##
## Coefficients:
##          (Intercept)          FSquaredFeet          NumBathrooms
##           -2.625e+06             9.809e+01             8.997e+03
##           GarageSize                  Year   as.factor(Quality)2
##            9.278e+03             1.396e+03            -1.344e+05
##  as.factor(Quality)3     as.factor(Style)2     as.factor(Style)3
##           -1.460e+05            -2.762e+04            -1.502e+04
##    as.factor(Style)4     as.factor(Style)5     as.factor(Style)6
##            1.387e+04            -2.801e+04            -7.459e+03
##    as.factor(Style)7     as.factor(Style)9    as.factor(Style)10
```

```
##            -4.389e+04            -8.783e+04            -8.318e+04
##    as.factor(Style)11              LotSize  as.factor(AdHighway)1
##            -9.131e+04             1.300e+00            -3.692e+04
```

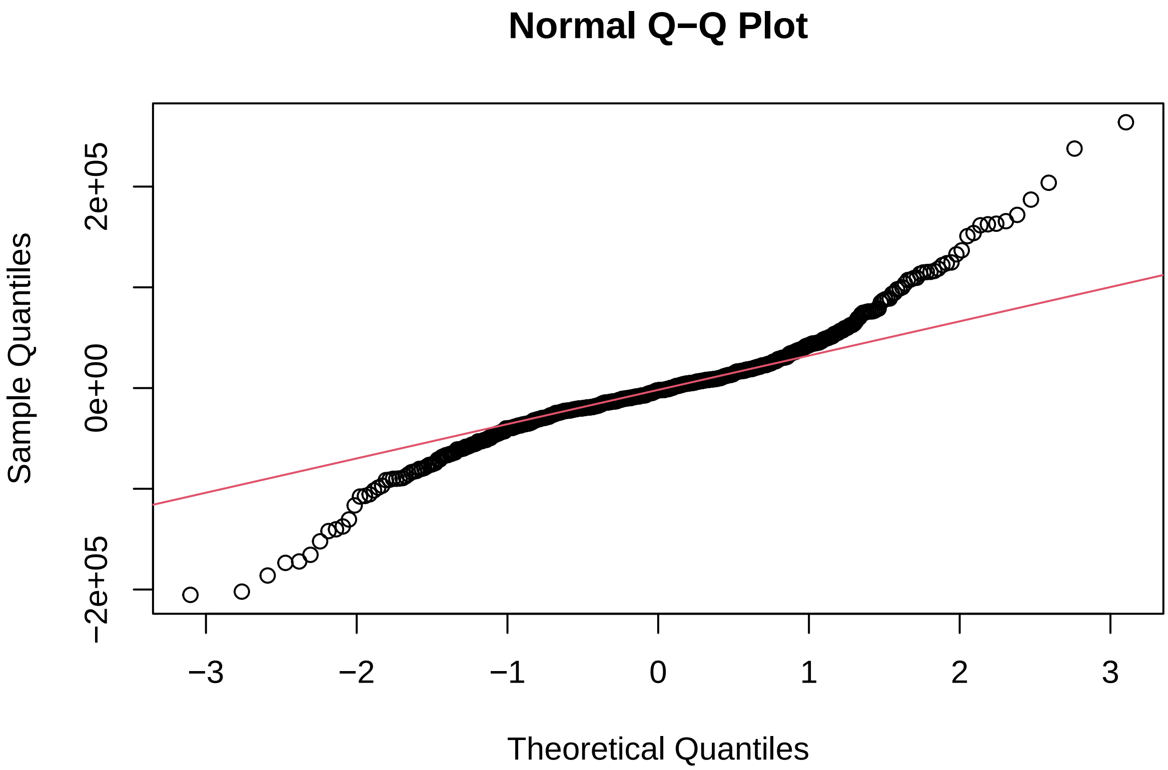
As shown in the table, Finished_Squared_Feet, Number_of_Bathrooms, Garage_Size, Year, Quality, Style, LotSize and Adjacent_to_Highway are included.

(8.)

```r
library(leaps)
selected.model = lm(SalesPrice~ FSquaredFeet + NumBathrooms +
                  GarageSize + Year + as.factor(Quality) + as.factor(Style) +
                  LotSize + as.factor(AdHighway), input1)
e <- selected.model$residuals
pred3 <- selected.model$fitted.values
plot(e~pred3)
```




```r
qqnorm(e)
qqline(e, col = 2)
```



From the first plot, I find a funnel shape, which mneans the variance is not constant. From the qq plot, It is not located along the line, which means the errors are not normally distributed.
As for the remedial measures, I prefer to do Box-Cox transformation.