Shuting Chen

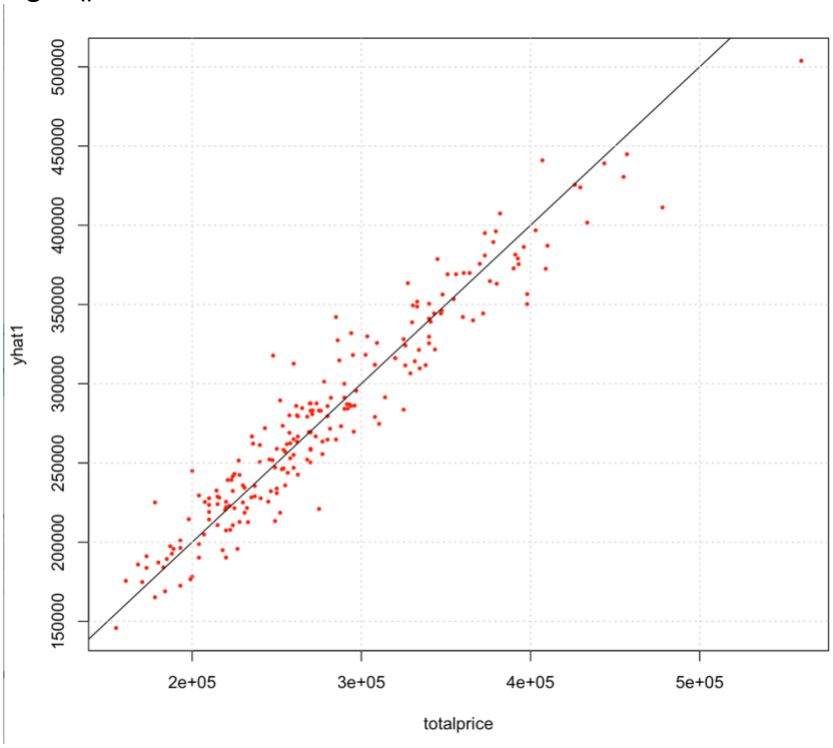
HW4

> #1

```
> #AIC
> d0=data.frame(VIT2005)
> m0=lm(totalprice~.,d0)
> step1=stepAIC(m0)
Start:  AIC=4412.62
totalprice ~ area + zone + category + age + floor + rooms + out +
    conservation + toilets + garage + elevator + streetcategory +
    heating + storage

                Df  Sum of Sq       RSS    AIC
- conservation    3 1.0031e+09 8.6894e+10 4409.2
- age             1 3.7563e+06 8.5895e+10 4410.6
- floor           1 3.8440e+07 8.5929e+10 4410.7
<none>                        8.5891e+10 4412.6
- rooms           1 1.0656e+09 8.6956e+10 4413.3
- storage         1 1.6433e+09 8.7534e+10 4414.8
- streetcategory  3 3.5550e+09 8.9446e+10 4415.5
- out             3 3.8946e+09 8.9785e+10 4416.3
- heating         3 4.3202e+09 9.0211e+10 4417.3
- toilets         1 4.7971e+09 9.0688e+10 4422.5
- category        6 9.5199e+09 9.5411e+10 4423.5
- elevator        1 5.4265e+09 9.1317e+10 4424.0
- garage          1 1.4771e+10 1.0066e+11 4445.2
- area            1 4.4519e+10 1.3041e+11 4501.7
- zone           22 1.1171e+11 1.9760e+11 4550.3

……..doesn't show all steps

Step:  AIC=4406.14
totalprice ~ area + zone + category + rooms + out + toilets +
    garage + elevator + streetcategory + heating + storage

                Df  Sum of Sq       RSS    AIC
<none>                        8.7288e+10 4406.1
- rooms           1 1.0246e+09 8.8312e+10 4406.7
```

```
- storage        1 1.6695e+09 8.8957e+10 4408.3
- streetcategory  3 3.5484e+09 9.0836e+10 4408.8
- heating        3 3.9987e+09 9.1286e+10 4409.9
- out            3 4.5287e+09 9.1816e+10 4411.2
- toilets        1 5.1432e+09 9.2431e+10 4416.6
- elevator       1 5.7882e+09 9.3076e+10 4418.1
- category       6 1.2678e+10 9.9966e+10 4423.7
- garage         1 1.5621e+10 1.0291e+11 4440.0
- area           1 4.4067e+10 1.3135e+11 4493.2
- zone          22 1.1785e+11 2.0514e+11 4548.4
```

```
> m1=glm(totalprice ~ area + zone + category + rooms + out + toilets + garage + elevator +
streetcategory + heating + storage,data=d0)
> yhat1=predict(m1,d0)
> plot(yhat1~totalprice,d0,cex=0.4,pch=19,col="red")
> abline(0,1)
> grid()
```

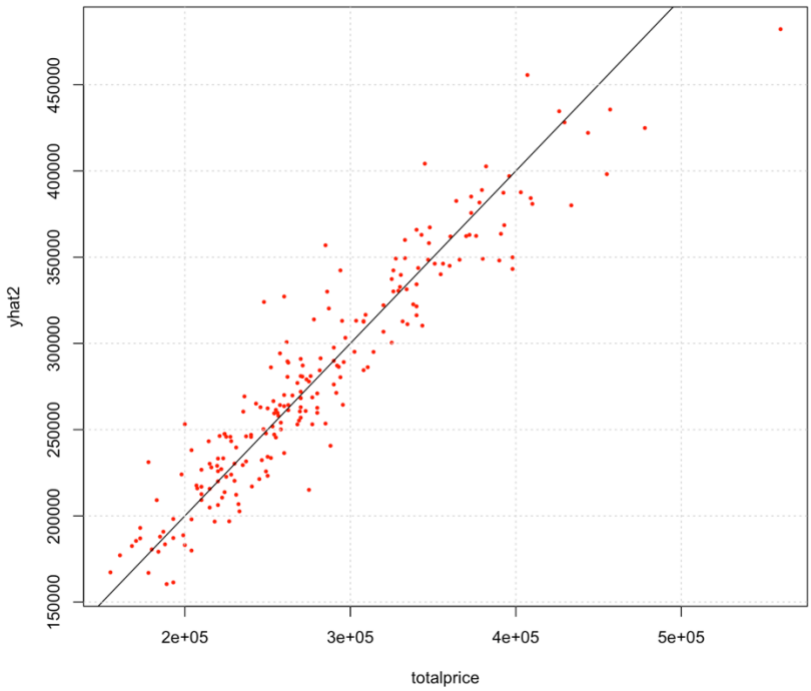



```
> #BIC
> n1=nrow(d0)
> step2=stepAIC(m0,k=log(n1))
Start:  AIC=4578.46
totalprice ~ area + zone + category + age + floor + rooms + out +
    conservation + toilets + garage + elevator + streetcategory +
    heating + storage
```

```
               Df  Sum of Sq        RSS    AIC
- conservation    3 1.0031e+09 8.6894e+10 4564.8
- category        6 9.5199e+09 9.5411e+10 4569.1
- streetcategory  3 3.5550e+09 8.9446e+10 4571.2
- out             3 3.8946e+09 8.9785e+10 4572.0
- heating         3 4.3202e+09 9.0211e+10 4573.0
- age             1 3.7563e+06 8.5895e+10 4573.1
- floor           1 3.8440e+07 8.5929e+10 4573.2
- rooms           1 1.0656e+09 8.6956e+10 4575.8
- storage         1 1.6433e+09 8.7534e+10 4577.2
<none>                        8.5891e+10 4578.5
- toilets         1 4.7971e+09 9.0688e+10 4584.9
- elevator        1 5.4265e+09 9.1317e+10 4586.4
- garage          1 1.4771e+10 1.0066e+11 4607.7
- zone           22 1.1171e+11 1.9760e+11 4641.6
- area            1 4.4519e+10 1.3041e+11 4664.1
```

……..doesn't show all steps

```
Step:  AIC=4526.97
totalprice ~ area + zone + toilets + garage + elevator + storage

          Df  Sum of Sq        RSS    AIC
<none>                   1.1393e+11 4527.0
- storage   1 3.5733e+09 1.1750e+11 4528.3
- toilets   1 1.2489e+10 1.2642e+11 4544.3
- elevator  1 1.2766e+10 1.2669e+11 4544.7
- garage    1 1.7837e+10 1.3177e+11 4553.3
- zone     22 1.2501e+11 2.3894e+11 4570.0
- area      1 7.0588e+10 1.8452e+11 4626.7
> m2=glm(totalprice ~ area + zone + toilets + garage + elevator + storage,data=d0)
> yhat2=predict(m2,d0)
> plot(yhat2~totalprice,d0,cex=0.4,pch=19,col="red")
> abline(0,1)
> grid()
```




```
> #5-fold cross validation MSPE of the models m1 and m2
> set.seed(1)
> cverrors1=cv.glm(d0,m1,K=5)$delta[1]
> cverrors1
[1] 662589324
> cverrors2=cv.glm(d0,m2,K=5)$delta[1]
> cverrors2
[1] 715765090
>
>
```

> #2

```
> setwd("/Users/shutingchen/Desktop/ISE 529          Data Analytics/L5")
> d1=read.csv("chlorine.csv")
> head(d1)
  Pct.Dep Temperature PH.Level Weather
1    32.6          60     6.6       2
2    40.4          65     6.6       2
3    39.4          70     6.6       2
4    37.3          75     6.6       2
5    45.1          80     6.6       2
6    40.6          85     6.6       2
> m1=lm(Pct.Dep~Temperature+PH.Level+I(PH.Level^2)+Weather,data=d1)
> summary(m1)

Call:
```

```
lm(formula = Pct.Dep ~ Temperature + PH.Level + I(PH.Level^2) +
   Weather, data = d1)

Residuals:
    Min      1Q  Median      3Q     Max
-11.4840  -2.9816   0.1387   3.0292  10.1392

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1001.73548   56.03128  17.878  < 2e-16 ***
Temperature      0.19376    0.02903   6.674 2.28e-10 ***
PH.Level      -265.61190   14.99115 -17.718  < 2e-16 ***
I(PH.Level^2)   17.75794    0.99884  17.779  < 2e-16 ***
Weather          0.53429    0.35557   1.503    0.134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.207 on 205 degrees of freedom
Multiple R-squared:  0.6404,  Adjusted R-squared:  0.6334
F-statistic: 91.28 on 4 and 205 DF,  p-value: < 2.2e-16

> # p-value=2.28e-10 so we can refer weather is a factor have relationship with chlorine and the slope of temperture=0.19376 is positive  so we can infer that higher temperatures deplete chlorine more quickly
> #p-value of PH.Level and (PH.Level^2)<2e-16 is too small so we cannot accept H0, and conclude that the belief about the relationship between chlorine depletion and pH level is correct
> d2=d1
> d2$Weather=as.factor(d2$Weather)
> m2=lm(Pct.Dep~Temperature+PH.Level+I(PH.Level^2)+Weather,data=d2)
> #or use m2=glm(Pct.Dep~Temperature+poly(PH.Level,2,raw=T)+Weather,data=d2)
> anova(m2)
Analysis of Variance Table

Response: Pct.Dep
             Df Sum Sq Mean Sq  F value    Pr(>F)
Temperature    1  788.4   788.4  46.0201 1.242e-10 ***
PH.Level       1   39.7    39.7   2.3189  0.129360
I(PH.Level^2)  1 5594.5  5594.5 326.5498 < 2.2e-16 ***
Weather        2  173.5    86.7   5.0629  0.007146 **
Residuals    204 3494.9    17.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> #p-value of weather= 0.007146, is smaller than 0.01 we reject H0, so the weather have relationship with chlorine depletion.
>

> #3

```
> #a)
> setwd("/Users/shutingchen/Desktop/ISE 529              Data Analytics/L5")
> d3=read.csv("commercial.csv")
> str(d3)
'data.frame':   60 obs. of  3 variables:
 $ Test  : int  24 20 16 11 10 4 24 18 16 15 ...
 $ Length: int  52 40 36 28 44 16 48 52 60 44 ...
 $ Type  : int  1 2 2 1 3 1 1 2 3 2 ...
> m4=lm(Test~.,d)
> summary(m4)

Call:
lm(formula = Test ~ ., data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-15.6470  -3.7245   0.0442   4.1159  13.6990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0218     3.2396   2.167   0.0344 *
Length        0.2495     0.0560   4.456 3.96e-05 ***
Type         -1.3518     0.9470  -1.427   0.1589
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.837 on 57 degrees of freedom
Multiple R-squared:  0.3138,  Adjusted R-squared:  0.2897
F-statistic: 13.03 on 2 and 57 DF,  p-value: 2.183e-05

> #p-value of type =0.1589 so we cannot conclude that the memory test score is related to the type of commercial
> #b)
> d4=d3
> d4$Type=as.factor(d4$Type)
> str(d4)
'data.frame':   60 obs. of  3 variables:
 $ Test  : int  24 20 16 11 10 4 24 18 16 15 ...
 $ Length: int  52 40 36 28 44 16 48 52 60 44 ...
```

```
 $ Type  : Factor w/ 3 levels "1","2","3": 1 2 2 1 3 1 1 2 3 2 ...
> m5=lm(Test~.,d4)
> anova(m5)
Analysis of Variance Table

Response: Test
          Df  Sum Sq Mean Sq F value    Pr(>F)
Length     1  818.50  818.50 26.4780 3.546e-06 ***
Type       2  280.01  140.01  4.5292   0.01502 *
Residuals 56 1731.09   30.91
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #p-value of Type is 0.01502 so we can conclude that the memory test score is related to the type2-musical (2), and type3-serious (3) of commercial
> #but we cannot tell whether memory test score is related to type1-humorous (1).
> #c)
> #Multiple R-squared:  0.3138,         Adjusted R-squared:  0.2897
> #Multiple R-squared:  0.3882,         Adjusted R-squared:  0.3554
```