Shuting Chen

HW2

```
> library(leaps)
> library(MASS)
> library(lattice)
> library(ggplot2)
> library(PASWR2)
> setwd("/Users/shutingchen/Desktop/ISE 529            Data Analytics/HW")
> d0=read.csv("homes.csv",header = T)
>
```

> #1

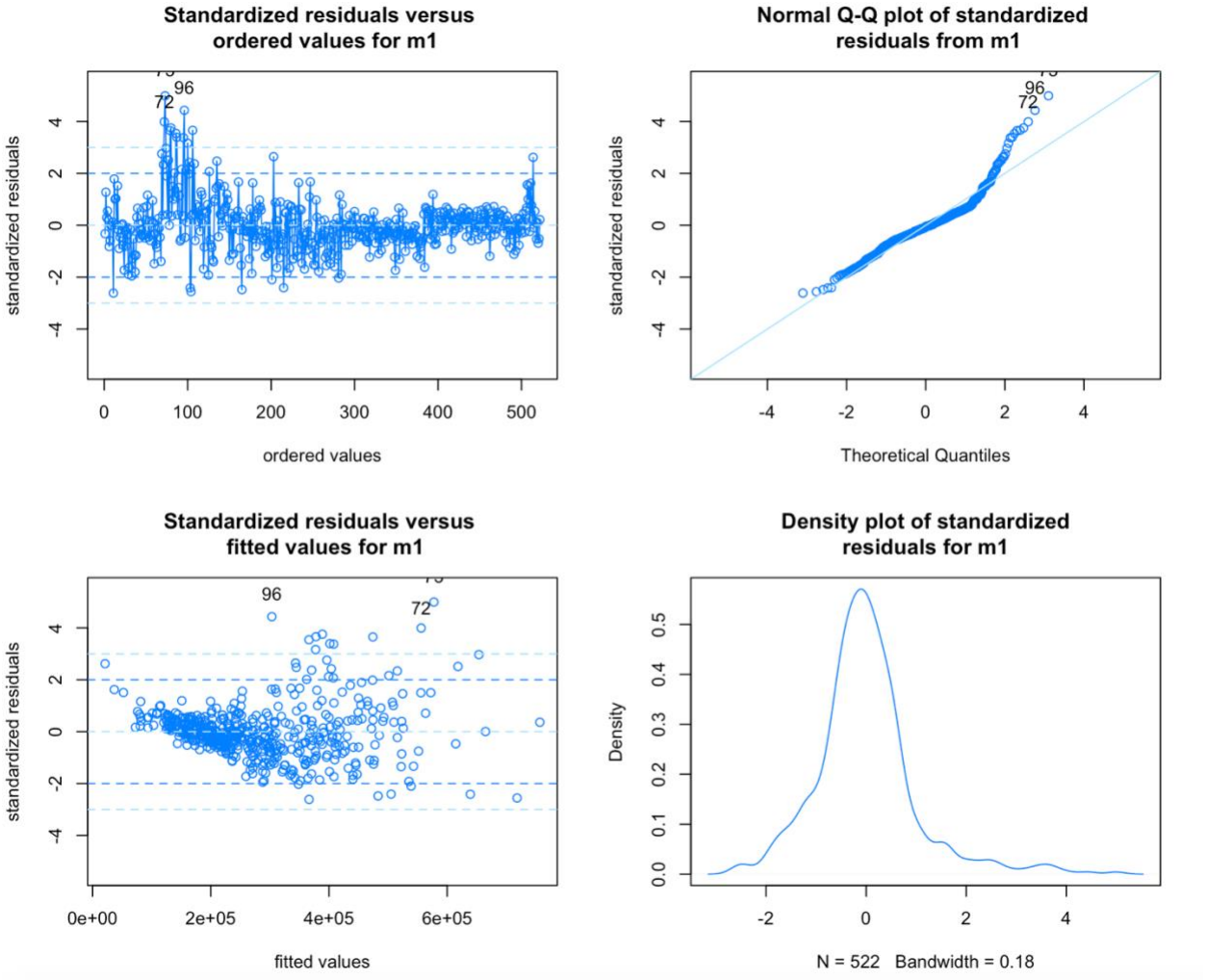
```
> d1=d0[,c(1,2,3,4,5,6,8)]
> cor(d1)
          price     area     beds    baths   garage     year    lotsize
price   1.0000000 0.8194701 0.4133239 0.6836854 0.5777863  0.5555164  0.2241685
area    0.8194701 1.0000000 0.5578378 0.7552729 0.5337665  0.4411967  0.1575247
beds    0.4133239 0.5578378 1.0000000 0.5834469 0.3168137  0.2686924  0.1265384
baths   0.6836854 0.7552729 0.5834469 1.0000000 0.4898981  0.5128410  0.1470066
garage  0.5777863 0.5337665 0.3168137 0.4898981 1.0000000  0.4617604  0.1522193
year    0.5555164 0.4411967 0.2686924 0.5128410 0.4617604  1.0000000 -0.1004519
lotsize 0.2241685 0.1575247 0.1265384 0.1470066 0.1522193 -0.1004519  1.0000000
> #  the area and baths are the predictors with highest correlation
>
```

> #2

```
> subset(d1,area,subset=price==max(d0$price))
   area
73 3857
> #the area of the most expensive house is 3857
> #Fit the full model.
> m1=lm(price~.,d1)
> coef(m1)
  (Intercept)         area         beds        baths       garage         year      lotsize
-3.567709e+06  1.257386e+02 -1.304139e+04  7.987552e+03  2.253038e+04  1.779611e+03
1.554990e+00
> #full model: yhat=-3.567709e+06 + 1.257386e+02*area+ -1.304139e+04*beds+
7.987552e+03*baths+  2.253038e+04*garage +  1.779611e+03*year+  1.554990e+00*lotsize
>
```

> #3

> checking.plots(m1)



> #assumptions hold
> #outliers(72,73,96)
> d1[c(73),]
price area beds baths garage year lotsize
73 920000 3857 4 5 3 1997 32793
> #the largest outlier is in row 73
>

> #4

> confint(m1,level = 0.99)

	0.5 %	99.5 %
(Intercept)	-4.649785e+06	-2.485632e+06
area	1.077357e+02	1.437415e+02
beds	-2.281527e+04	-3.267520e+03
baths	-4.471425e+03	2.044653e+04
garage	7.497920e+03	3.756284e+04

```
year        1.222488e+03  2.336734e+03
lotsize     8.517208e-01  2.258260e+00
> #99% confidence interval for area is(1.077357e+02,1.437415e+02)
>
```

> #5

```
> newval=data.frame(area=2650,beds=3,baths=3,garage=2,year=1990,lotsize=24500)
> predict(m1,newval,interval = "conf",level = 0.95)
      fit     lwr     upr
1 374920.5 362128.4 387712.6
> # 95% confidence interval for the mean price is(362128.4，387712.6)
>
```

> #6

```
> #best subset of predictors (in terms of adj-R 2 )
> models=regsubsets(price~.,d1,nvmax=12)
> summary(models)
Subset selection object
Call: regsubsets.formula(price ~ ., d1, nvmax = 12)
6 Variables  (and intercept)
       Forced in Forced out
area       FALSE     FALSE
beds       FALSE     FALSE
baths      FALSE     FALSE
garage     FALSE     FALSE
year       FALSE     FALSE
lotsize    FALSE     FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         area beds baths garage year lotsize
1  ( 1 ) "*"  " "  " "   " "    " "  " "
2  ( 1 ) "*"  " "  " "   " "    "*"  " "
3  ( 1 ) "*"  " "  " "   " "    "*"  "*"
4  ( 1 ) "*"  " "  " "   "*"    "*"  "*"
5  ( 1 ) "*"  "*"  " "   "*"    "*"  "*"
6  ( 1 ) "*"  "*"  "*"   "*"    "*"  "*"
> a=summary(models)$adjr2
> which.max(a)
[1] 6
> #best model is in row 6
> #best model includes  area, beds，baths，garage, year, lotsize
> #these variables are highly correlated with price
>
newval2=data.frame(area=mean(d1$area),beds=mean(d1$beds),baths=mean(d1$baths),garag
e=mean(d1$garage),year=mean(d1$year),lotsize=mean(d1$lotsize))
```

```
> predict(m1,newval2)
      1
277894.1
>
> #full model: yhat=-3.567709e+06 + 1.257386e+02*area+ -1.304139e+04*beds+
7.987552e+03*baths+  2.253038e+04*garage +  1.779611e+03*year+  1.554990e+00*lotsize
>
```

> #7

```
> #best predictor is area, worst predictor is baths
>
> m2=lm(price~beds,d1)
> coef(m2)
(Intercept)       beds
  82808.80    56200.08
> #price=82808.80 + 56200.08*beds
```

> #8

```
> #Interpret the slope value b:This means with number of bedroom increasing by 1, the price
increases by $56200.08, on average
>
> d3=subset(d1,subset = beds == 2|beds == 3|beds == 4)
> m3=lm(price~.,d3)
> coef(m3)
 (Intercept)         area         beds        baths       garage         year      lotsize
-3.240850e+06  1.327428e+02 -1.274040e+04  8.812354e+03  2.207092e+04  1.603830e+03
1.581008e+00
> #full model for houses having between two to four bedrooms
> #yhat2=-3.240850e+06  1.327428e+02*area -1.274040e+04*beds  8.812354e+03*baths
2.207092e+04*garage  1.603830e+03*year  1.581008e+00 *lotsize
>
```

> #9

```
> anova(m3)
Analysis of Variance Table

Response: price
           Df     Sum Sq    Mean Sq   F value    Pr(>F)
area        1 5.2642e+12 5.2642e+12 1241.1540 < 2.2e-16 ***
beds        1 4.6931e+09 4.6931e+09    1.1065    0.2934
baths       1 1.1787e+11 1.1787e+11   27.7901 2.128e-07 ***
garage      1 1.4762e+11 1.4762e+11   34.8041 7.302e-09 ***
year        1 1.6438e+11 1.6438e+11   38.7575 1.125e-09 ***
lotsize     1 1.3158e+11 1.3158e+11   31.0231 4.457e-08 ***
Residuals 438 1.8577e+12 4.2413e+09
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #MSE=4.2413e+09, is the estimate of the variance
>
> summary(m3)

Call:
lm(formula = price ~ ., data = d3)

Residuals:
   Min     1Q  Median     3Q    Max
-173208  -34304   -3198   26486  334853

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.241e+06  4.214e+05  -7.690 9.80e-14 ***
area         1.327e+02  7.537e+00  17.613  < 2e-16 ***
beds        -1.274e+04  5.078e+03  -2.509 0.012474 *
baths        8.812e+03  5.049e+03   1.745 0.081622 .
garage       2.207e+04  5.949e+03   3.710 0.000234 ***
year         1.604e+03  2.171e+02   7.387 7.68e-13 ***
lotsize      1.581e+00  2.839e-01   5.570 4.46e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65130 on 438 degrees of freedom
Multiple R-squared:  0.7584,  Adjusted R-squared:  0.7551
F-statistic: 229.1 on 6 and 438 DF,  p-value: < 2.2e-16

> #R^2=0.7584, 75.84% of variation of prices is explained by variation of model factors
>
```

> #10

```
> newval3=data.frame(area=3150,beds=2,baths=3,garage=2,year=1996,lotsize=26250)
> predict(m3,newval3,interval = "pred",level = 0.95)
      fit     lwr     upr
1 465134.5 334969.9 595299.1
> #95% prediction interval for the price is （334969.9，595299.1）
```