Homework 2

Shun-Lung Chang, Dilip Hiremath

```r
load("data/OregonHomes.Rdata") # read the data into R
```

1. First of all, read the data file OregonHomes.Rdata (the data frame is called homes) and load the libraries you typically use. Create a new variable that groups the garage size information into two classes: one for garage size for no or one car, the second one for garage sizes for two or more cars.[hint: There are multiple ways to do this. E.g., using the cut command or the command recode from the car package.]

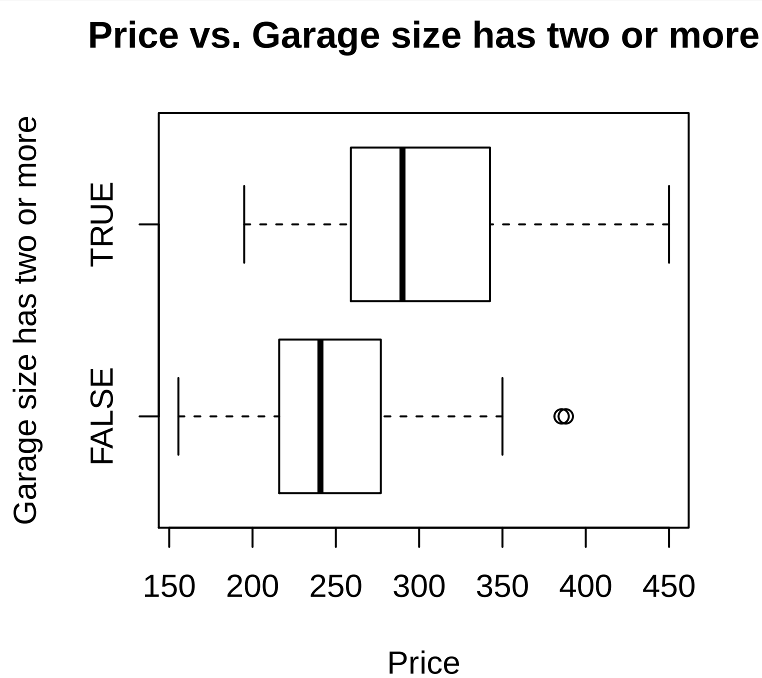
Generate a boxplot for the house prices grouped by the newly created garage size groups.

```r
homes$two_cars_more <- homes$Gar >= 2
```

(a) (1 point) Based on the box plot, do you expect that the mean house price differs significantly between the two groups?

As the plot indicates, the mean house price between two groups may differ significantly given the high differece (approximately 45).

```r
boxplot(Price ~ two_cars_more,
        data = homes,
        main = "Price vs. Garage size has two or more",
        xlab = "Price",
        ylab = "Garage size has two or more",
        horizontal = TRUE)
```



(b) (half a point) Using a t-test assuming equal variances assess whether there is a significant difference in house prices between the two groups.

Considering a significance level of 5%, the p-value in the test is small enough to conclude that the house price in two groups are not equal.

```r
t.test(Price ~ two_cars_more, data = homes, var.equal = TRUE)
```

```

	Two Sample t-test

data:  Price by two_cars_more
t = -3.2878, df = 74, p-value = 0.001547
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -73.93064 -18.13474
sample estimates:
mean in group FALSE  mean in group TRUE 
           254.3000            300.3327 
```

(c) (half a point) Check whether equality of variance is actually given?

Given the large p-value, we can say the assumption of equal variance holds.

```r
var.test(Price ~ two_cars_more, data = homes)
```

```

	F test to compare two variances

data:  Price by two_cars_more
F = 1.2901, num df = 23, denom df = 51, p-value = 0.4427
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6652745 2.7534679
sample estimates:
ratio of variances 
          1.290124 
```

2. Run a one-way ANOVA-test (command aov to assess whether there is a significant difference in house prices between the two groups.

```r
aov_fit <- aov(Price ~ two_cars_more, data = homes)
summary(aov_fit)
```

```
              Df Sum Sq Mean Sq F value  Pr(>F)
two_cars_more  1  34796   34796   10.81 0.00155
Residuals     74 238211    3219
1 observation deleted due to missingness
```

(a) (half a point) Based on the one-way ANOVA-test is there a significant difference in house prices between the two groups.

We can conclude that there exist a significant difference between the two groups since the p-value is small enough under significance level of 5%.

(b) (half a point) Assess by using a linear model whether there is a significant difference in house prices between the two groups.

The small p-value indicates there is a significant difference in house prices between the two groups.

```r
mod_1 <- lm(Price ~ two_cars_more, data = homes)
summary(mod_1)
```

```
Call:
lm(formula = Price ~ two_cars_more, data = homes)

Residuals:
    Min      1Q  Median      3Q     Max
-105.33  -39.81  -13.55   39.59  149.67

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         254.30      11.58  21.958  < 2e-16
two_cars_moreTRUE    46.03      14.00   3.288  0.00155

Residual standard error: 56.74 on 74 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1275,    Adjusted R-squared:  0.1157
F-statistic: 10.81 on 1 and 74 DF,  p-value: 0.001547
```

(c) (1 point) Compare the results of the t-test, the linear model and the ANOVA. How do the p-values of the three tests relate to each other? How do the test statistics of the three tests relate to each other?

The p-values in the results are the same, and the square of t value is equal to the F value.

3. Using the variable Gar as a factor, run an ANOVA model to see whether the garage size has a statistically significant impact on the average house price.

```r
aov_fit <- aov(Price ~ as.factor(Gar), data = homes)
summary(aov_fit)
```

```
               Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Gar)  3  36682   12227   3.725  0.015
Residuals      72 236325    3282
1 observation deleted due to missingness
```

(a) (half a point) Does the test result indicate that garage size has a statistically significant impact on house prices? Report the observed p-value for the overall ANOVA test!

Under significant level of 5%, the average house price is significantly different among different garage sizes.

(b) (half a point) Use the Tukey HSD post hoc test to determine for which garage sizes average house prices differ significantly at the 5% significance level.

The average house price is different between the garage size of 0 and garage size of 2, as the table below indicates.

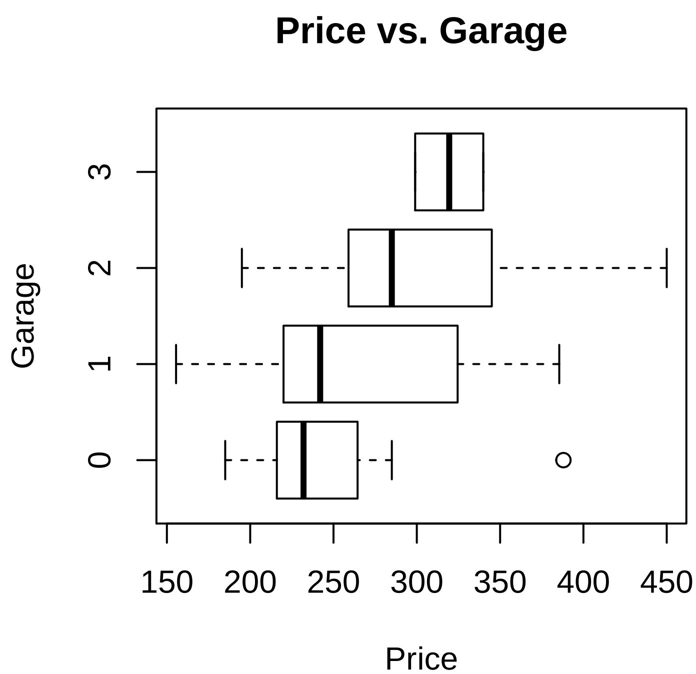
```
TukeyHSD(aov_fit, ordered = TRUE, conf.level = 0.95)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = Price ~ as.factor(Gar), data = homes)

$`as.factor(Gar)`
        diff        lwr       upr     p adj
1-0 13.74545 -47.983945  75.47485 0.9361062
2-0 52.71345   2.532603 102.89431 0.0357791
3-0 72.59545 -43.232873 188.42378 0.3585166
2-1 38.96800  -7.942282  85.87828 0.1372769
3-1 58.85000 -55.599370 173.29937 0.5330866
3-2 19.88200 -88.774603 128.53860 0.9630134
```

(c) (1 point) Can you explain why the average house price for homes with garages for 2 cars is signficantly different from the average house price for homes without garage (garage with car size 0) while the average house price for homes with garages for 3 cars is NOT signficantly different from the average house price for homes without garage (garage with car size 0) despite the fact that the average house price for homes with garages for three cars is larger than the one for homes with garages for two cars.

The boxplot shows that an outlier price exists in the garage size of 0, and value of the outlier (388) is larger than the maximum value (339.9) in the garage size of 3. Therefore there is no signficant difference between size of 0 and size of 3, even though the averge house price for homes with garages for three cars is the largest.

```
boxplot(Price ~ Gar, data = homes,
        main = 'Price vs. Garage',
        xlab = 'Price',
        ylab = 'Garage',
        horizontal = TRUE)
```



4. Now, you build a linear model for the house price based on all predictor variables in the original data set (So, please do not include the newly created grouping variable for the garage size).

```
mod_2 <- lm(Price ~ . - two_cars_more, data = homes)
```

(a) (1 point) According to this model and using the ANOVA table, which predictors have a signfifcant impact on the average house price at the 5% significance level?

The table indicates that the signfifcant predictors are *Floor*, *Lot*, *Status (if Sold)* and *School (if Edison)*.

```
summary(mod_2)
```

```
Call:
lm(formula = Price ~ . - two_cars_more, data = homes)

Residuals:
    Min      1Q  Median      3Q     Max 
-94.978 -28.849  -0.511  24.350  94.094 

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -202.1596   660.5569  -0.306  0.76061
ID              -0.2662     0.2796  -0.952  0.34495
Floor           80.3028    32.0373   2.507  0.01487
Lot             10.3434     3.6717   2.817  0.00652
Bath             4.4336    11.7998   0.376  0.70842
Bed            -12.9997     9.1684  -1.418  0.16131
Year             0.1604     0.3352   0.479  0.63396
```

```
Age                  NA          NA      NA       NA
Gar              6.1577      9.6116   0.641  0.52414
StatusPending  -17.8892     16.5880  -1.078  0.28508
StatusSold     -37.3573     13.9762  -2.673  0.00963
SchoolCrest     12.4545     36.1419   0.345  0.73158
SchoolEdison    91.7660     31.7622   2.889  0.00534
SchoolHarris    61.9000     33.0168   1.875  0.06561
SchoolParker    -6.9931     31.0327  -0.225  0.82246
SchoolRedwood   13.0448     30.4176   0.429  0.66954

Residual standard error: 45.03 on 61 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5468,    Adjusted R-squared:  0.4428
F-statistic: 5.258 on 14 and 61 DF,  p-value: 0.000002202
```

(b) (half a point) How good does the model fit?

The R-squared values means that the total variance of house can be explained by this model by 54.68%.

(c) (half a point) In which form is the variable Gar included in this model? As a factor or as a numeric variable? How do you see the difference in the output?

The variable *Gar* is used as a numeric variable. Furthermore, the model below shows that how *Gar* is used as a factor variable. We therefore can see the *Gar* is treated in a dicrete manner, and there are three levels in *Gar* (the level of 0 in *Gar* is included in the intercept term).

```r
homes$Gar <- as.factor(homes$Gar)
mod_3 <- lm(Price ~ . - two_cars_more, data = homes)
summary(mod_3)
```

```

Call:
lm(formula = Price ~ . - two_cars_more, data = homes)

Residuals:
     Min       1Q   Median       3Q      Max 
-101.644  -26.129    0.759   25.531   94.487 

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -259.6789   662.9167  -0.392  0.69668
ID              -0.2588     0.2807  -0.922  0.36033
Floor           81.7675    32.1374   2.544  0.01359
Lot             10.8002     3.7055   2.915  0.00502
Bath            -0.4432    12.7816  -0.035  0.97246
Bed            -12.3504     9.5822  -1.289  0.20247
Year             0.2044     0.3373   0.606  0.54687
Age                  NA          NA      NA       NA
Gar1           -20.5876    22.7198  -0.906  0.36854
Gar2             5.7959    20.0789   0.289  0.77385
Gar3             0.8246    46.5716   0.018  0.98593
StatusPending  -16.1670    17.7091  -0.913  0.36500
StatusSold     -37.7953    14.1371  -2.673  0.00969
```

```
SchoolCrest      -9.1262    39.9655  -0.228  0.82016
SchoolEdison     76.5241    33.8480   2.261  0.02747
SchoolHarris     48.1638    34.5973   1.392  0.16911
SchoolParker    -25.1866    34.4369  -0.731  0.46744
SchoolRedwood    -4.6595    33.3122  -0.140  0.88924

Residual standard error: 45.1 on 59 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5605,    Adjusted R-squared:  0.4413
F-statistic: 4.702 on 16 and 59 DF,  p-value: 0.000005733
```

5. (2 points) In the linear model from Question 4 either the line for variable Age or the one for variable Year is empty in the ANOVA table and in the coefficient table all corresponding numbers are marked as NA. Explain why!

The variable *Age* and variable *Year* have a perfect collinearity, their relationship can be represented as $Age = (Year - 1970) \times 0.1$. From the perspective of linear algebra, the matrix of predictors is not full rank, and hence the inverse of the matrix cannot be obtained. To circumvent this issue, the variable *Age* was excluded when the model was built, and its coefficient is shown as NA.

6. (2 points) Looking at the sign of the (significant) regression coefficients, do the empirically present relationships make sense?

The sign of significant variables seem to be reasonably indicate the relationships among the variables and price. For instance, the larger a house is, the higher its price is. Moreover, when a house is still in the market, the seller might try to drive the price up to reap profits. After the house is sold, its price then drops to its true value. Lastly, Edison may be a prestigious school, so the houses in its district can have a better price.

7. (2 points) Starting with a model using all predictors in the data set (except the grouped garage size and the variable Year) use the stepwise automatic model procedure to find the best linear model. Use the backward/forward strategy and the AIC as criterion. Briefly summarize the resulting model!

The AIC gradually declines in the process, as the tables below demonstrate. The final model consists of the variables *Floor*, *Lot*, *Bed*, *Status* and *School*. As expected the variables, except for *Bed*, are the signicant predictors in the full model in Task 4.

```r
mod_4 <- lm(Price ~ . - two_cars_more - Year, data = homes)
mod_5 <- step(mod_4, direction = "both")
```

```
Start:  AIC=593.7
Price ~ (ID + Floor + Lot + Bath + Bed + Year + Age + Gar + Status +
    School + two_cars_more) - two_cars_more - Year

         Df Sum of Sq    RSS    AIC
- Gar     3      4551 124546 590.53
- Bath    1         2 119997 591.70
- Age     1       747 120741 592.17
- ID      1      1729 121723 592.79
<none>                119995 593.70
- Bed     1      3379 123373 593.81
```

```
- Status  2     14920 134914 598.61
- Floor   1     13166 133160 599.61
- Lot     1     17278 137272 601.92
- School  5     74596 194590 620.44

Step:  AIC=590.53
Price ~ ID + Floor + Lot + Bath + Bed + Age + Status + School

         Df Sum of Sq    RSS    AIC
- Bath    1       309 124855 588.72
- Age     1      1174 125719 589.24
- ID      1      2343 126888 589.95
<none>                124546 590.53
- Bed     1      5347 129892 591.72
+ Gar     3      4551 119995 593.70
- Status  2     18476 143022 597.04
- Floor   1     14946 139492 597.14
- Lot     1     18735 143280 599.18
- School  5     77256 201802 617.21

Step:  AIC=588.72
Price ~ ID + Floor + Lot + Bed + Age + Status + School

         Df Sum of Sq    RSS    AIC
- Age     1      1641 126495 587.71
- ID      1      2531 127386 588.24
<none>                124855 588.72
- Bed     1      5065 129920 589.74
+ Bath    1       309 124546 590.53
+ Gar     3      4858 119997 591.70
- Status  2     18924 143779 595.44
- Lot     1     18447 143302 597.19
- Floor   1     20464 145319 598.25
- School  5     82345 207200 617.21

Step:  AIC=587.71
Price ~ ID + Floor + Lot + Bed + Status + School

         Df Sum of Sq    RSS    AIC
- ID      1      2441 128936 587.16
<none>                126495 587.71
+ Age     1      1641 124855 588.72
+ Bath    1       776 125719 589.24
+ Gar     3      5723 120773 590.19
- Bed     1      8539 135035 590.67
- Lot     1     17414 143909 595.51
- Status  2     22492 148988 596.15
- Floor   1     25538 152033 599.69
- School  5     81931 208427 615.66

Step:  AIC=587.16
Price ~ Floor + Lot + Bed + Status + School

         Df Sum of Sq    RSS    AIC
```

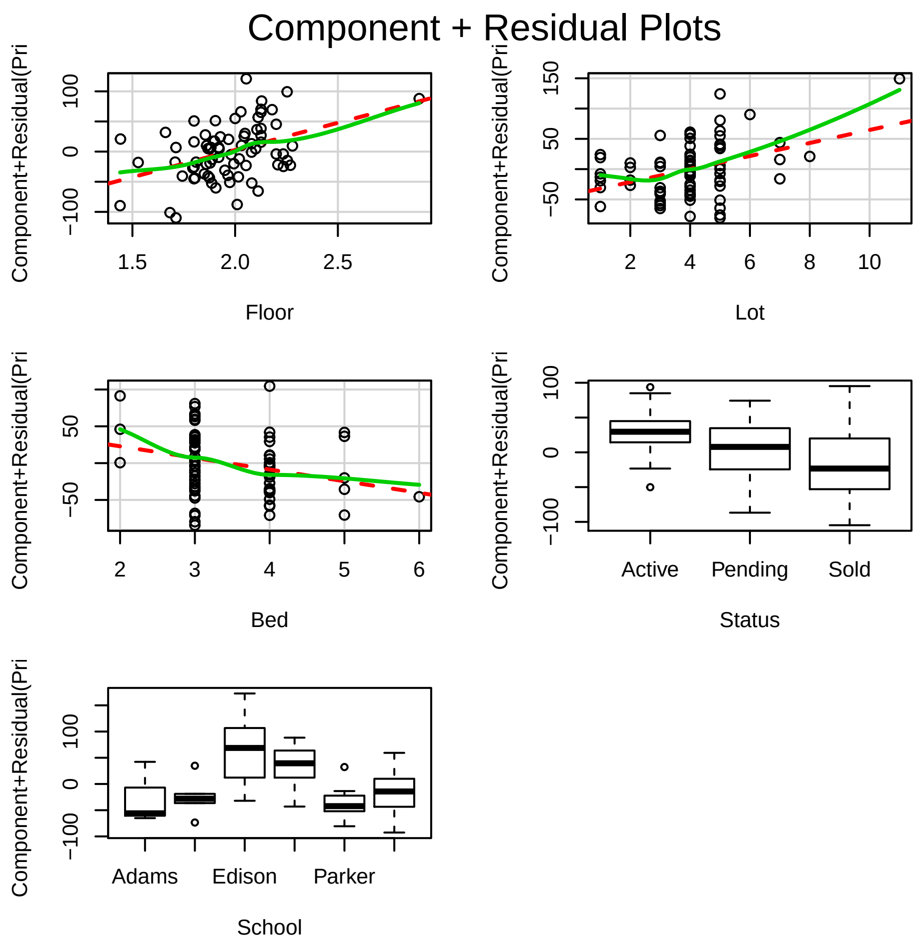
```
<none>                128936 587.16
+ ID      1      2441 126495 587.71
+ Age     1      1550 127386 588.24
+ Bath    1      1024 127912 588.56
+ Gar     3      6622 122314 589.15
- Bed     1      7690 136626 589.56
- Status  2     22760 151696 595.52
- Lot     1     18945 147881 595.58
- Floor   1     23307 152242 597.79
- School  5     80237 209172 613.93
```

```r
summary(mod_5)
```

```
Call:
lm(formula = Price ~ Floor + Lot + Bed + Status + School, data = homes)

Residuals:
    Min      1Q  Median      3Q     Max 
-91.369 -29.241  -0.683  24.312 113.296 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    118.459     54.739   2.164  0.03414
Floor           90.214     26.319   3.428  0.00106
Lot             10.754      3.480   3.090  0.00294
Bed            -15.787      8.018  -1.969  0.05323
StatusPending  -20.545     16.227  -1.266  0.20999
StatusSold     -43.293     12.890  -3.359  0.00131
SchoolCrest      1.259     33.763   0.037  0.97036
SchoolEdison    85.616     29.470   2.905  0.00501
SchoolHarris    58.508     29.680   1.971  0.05295
SchoolParker   -11.034     29.546  -0.373  0.71003
SchoolRedwood   11.902     28.145   0.423  0.67377

Residual standard error: 44.54 on 65 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5277,    Adjusted R-squared:  0.4551 
F-statistic: 7.263 on 10 and 65 DF,  p-value: 0.0000001392
```

8. Draw component/residual plots for all predictors in the final model resulting in the previous task. [hint: the package car contains a command crPlots to draw these plots.]

```r
library(car)
crPlots(mod_5)
```



(a) (half a point) Check whether some quadratic effects should be included.

The plot shows that there exist a slight quadratic effect in the variable *Lot*.

(b) (half a point) Vary the smoothing parameter to 0.25 and to 0.75. Which parameter setting indicates the quadratic effects more clearly? [hint: Look at the help pages for car::crPlots and check the examples of using the smoothing parameter.]

(c) (1 point) Add at least one quadratic effect to the model and compare the resulting model with the previous one. Is there a sufficient improvement in the model that justifies inclusion of the quadratic effect?

If the quadratic effect of the variable *Lot* is included, the new model is shown as follows. The adjusted R-squared is increased by 2.5%, thus we would say that adding the quadratic effect of *Lot* is legitimate.

```r
mod_6 <- lm(Price ~ Floor + Lot + Bed + Status + School + I(Lot ^ 2), data = homes)
summary(mod_6)
```

```
Call:
lm(formula = Price ~ Floor + Lot + Bed + Status + School + I(Lot^2),
    data = homes)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-86.520 -27.011  -1.905  27.445 118.260 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   168.3582    58.7639   2.865 0.005635
Floor          79.4582    26.2322   3.029 0.003536
Lot            -7.1028     9.3763  -0.758 0.451513
Bed           -14.7420     7.8456  -1.879 0.064797
StatusPending -21.9238    15.8591  -1.382 0.171651
StatusSold    -45.3581    12.6264  -3.592 0.000637
SchoolCrest    10.2465    33.2594   0.308 0.759024
SchoolEdison   93.5169    29.0337   3.221 0.002009
SchoolHarris   58.1964    28.9808   2.008 0.048857
SchoolParker   -3.4065    29.0906  -0.117 0.907147
SchoolRedwood  18.2384    27.6559   0.659 0.511956
I(Lot^2)        1.8599     0.9103   2.043 0.045145

Residual standard error: 43.49 on 64 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5566,    Adjusted R-squared:  0.4804 
F-statistic: 7.305 on 11 and 64 DF,  p-value: 0.00000006514
```