Statistical Modeling with R, Homework #2

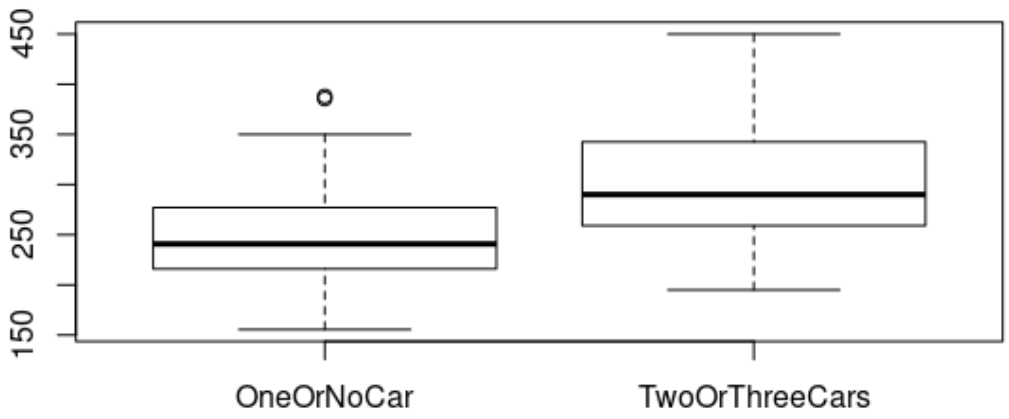
Group Members

- Karish Raja Rai
- Tom Wiesing
- Julius Nzaramba

```
> ### Load the dataset again
> load("./OregonHomes.Rdata")
> summary(homes)
     ID         Price           Floor            Lot              Bath
Bed
 Min.   : 1   Min.   :155.5   Min.   :1.440   Min.   : 1.000   Min.   :1.000   Min.
:2.000
 1st Qu.:20   1st Qu.:242.8   1st Qu.:1.861   1st Qu.: 3.000   1st Qu.:2.000   1st
Qu.:3.000
 Median :39   Median :276.0   Median :1.965   Median : 4.000   Median :2.000
Median :3.000
 Mean   :39   Mean   :285.8   Mean   :1.969   Mean   : 3.987   Mean   :2.206   Mean
:3.442
 3rd Qu.:58   3rd Qu.:336.8   3rd Qu.:2.106   3rd Qu.: 5.000   3rd Qu.:3.000   3rd
Qu.:4.000
 Max.   :77   Max.   :450.0   Max.   :2.896   Max.   :11.000   Max.   :3.100   Max.
:6.000
          NA's   :1
     Year            Age                Gar            Status       School
 Min.   :1905   Min.   :-6.50000   Min.   :0.000   Active :26   Adams  : 3
 1st Qu.:1958   1st Qu.:-1.20000   1st Qu.:1.000   Pending:13   Crest  : 6
 Median :1970   Median : 0.00000   Median :2.000   Sold   :38   Edison :12
 Mean   :1969   Mean   :-0.05195   Mean   :1.571                Harris :14
 3rd Qu.:1980   3rd Qu.: 1.00000   3rd Qu.:2.000                Parker :16
 Max.   :2005   Max.   : 3.50000   Max.   :3.000                Redwood:26

>
> ## and we need this down the road
> library(car)
```

Question 1

```
> ## Create a new factor $GarGroup
> homes$GarGroup <- NA # init with NA
> homes$GarGroup[homes$Gar <= 1] <- "OneOrNoCar" # one group for OneOrNoCars
> homes$GarGroup[homes$Gar >= 2] <- "TwoOrThreeCars" # and another group for the
other cases
> homes$GarGroup <- as.factor(homes$GarGroup) # make sure it is a factor
>
> ## make a boxplot
> boxplot(Price~GarGroup, data=homes)
```



(A) Yes we expect the mean of the prizes to differ significantly as the ranges for "TwoOrThreeCars" us significantly higher. Furthermore, the IQR is bigger for the second group.

(B) The p-value is very small, so we can assume there is a significant difference in house prices

```
> t.test(Price~GarGroup, data=homes, var.equal=TRUE) # P = 0.001547

      Two Sample t-test

data:  Price by GarGroup
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 OneOrNoCar mean in group TwoOrThreeCars
              254.3000                       300.3327
```

(C) This is not actually given, we see OneOrNoCar has Variance 3809, TwoOrThreeCars has 2952.802

```
> sapply(levels(homes$GarGroup), function(g){var(homes[homes$GarGroup == g,]$Price,
na.rm = TRUE)})
   OneOrNoCar TwoOrThreeCars
     3809.479       2952.802
```

Question 2

```
> ex2a <- aov(Price~GarGroup, data=homes)
```

```
> summary(ex2a)
            Df Sum Sq Mean Sq F value  Pr(>F)
GarGroup     1  34796   34796   10.81 0.00155 **
Residuals   74 238211    3219
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

(A) There is a significant difference, as F > 1 and p = 0.00155 < 0.01

(B) P =0.00155 < 0.01 => it is significant

```
> ex2b <- lm(Price~GarGroup, data=homes)
> summary(ex2b)

Call:
lm(formula = Price ~ GarGroup, 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 ***
GarGroupTwoOrThreeCars    46.03      14.00   3.288  0.00155 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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) The p-value for linear model and anova are the same. It differed slightly for the t test. All three compute a t statistic, so the value is obviously the same.

Question 3

```
> ex3 <- aov(Price ~ as.factor(Gar), data=homes)
> summary(ex3)
             Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Gar)  3  36682   12227   3.725  0.015 *
Residuals      72 236325    3282
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

(A) Yes it does. P = 0.015

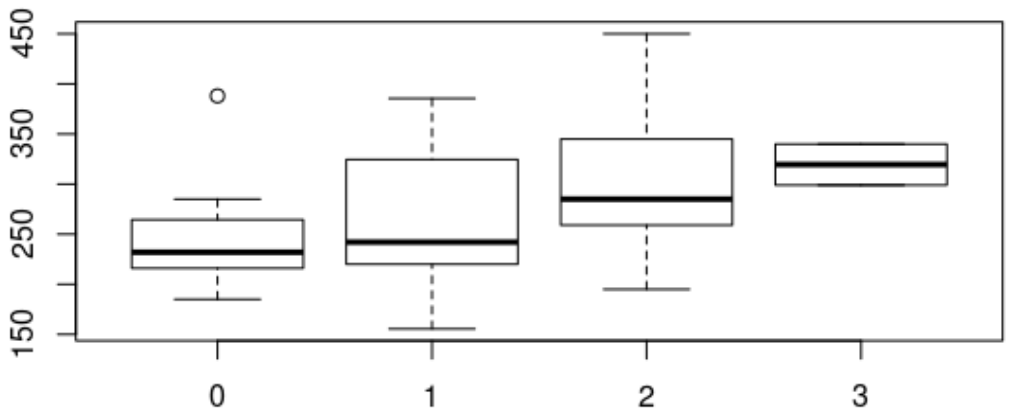
(B) Only 2-0

```
> TukeyHSD(ex3)
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
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) There is an outlier for houses with no garage (see boxplot).

```
> boxplot(Price ~ as.factor(Gar), data=homes)
```



Question 4

```
> ex4 <- lm(Price ~ Floor + Lot + Bath + Bed + Year + Age + Gar + Status + School,
 data=homes)
```

(A) Floor, Lot, Bed, Gar, School are significant.

```
> anova(ex4)
Analysis of Variance Table

Response: Price
          Df Sum Sq Mean Sq F value    Pr(>F)
Floor      1  11079 11078.6  5.4709  0.022574 *
Lot        1  15230 15229.8  7.5209  0.007962 **
Bath       1   5711  5711.5  2.8205  0.098103 .
Bed        1  23964 23963.5 11.8338  0.001046 **
Year       1    646   646.0  0.3190  0.574239
```

```
Gar        1   8429  8428.5  4.1622  0.045601 *
Status     2   9782  4890.9  2.4153  0.097707 .
School     5  72617 14523.4  7.1720 2.372e-05 ***
Residuals 62 125551  2025.0
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(B) The AIC gives us the model quality, which is 808.8182

```
> AIC(ex4)
[1] 808.8182
```

(C) As a numerical variable. We see this from the Df (degrees of freedom). For the explicit comparison, see the anova table in 4a) and the one below.

```
> aov(lm(Price ~ Floor + Lot + Bath + Bed + Year + Age + as.factor(Gar) + Status +
School, data=homes))
Call:
  aov(formula = lm(Price ~ Floor + Lot + Bath + Bed + Year + Age +
   as.factor(Gar) + Status + School, data = homes))

Terms:
                Floor       Lot      Bath       Bed      Year as.factor(Gar)
Status
Sum of Squares   11078.57  15229.83   5711.46  23963.51    646.00       11728.25
9751.54
Deg. of Freedom         1         1         1         1         1              3
2
              School Residuals
Sum of Squares   73174.98 121723.12
Deg. of Freedom         5        60

Residual standard error: 45.0413
1 out of 17 effects not estimable
Estimated effects may be unbalanced
1 observation deleted due to missingness
```

Question 5

The variable age is linearly dependent on year (we obviously have age + year == constant). Hence during the liner model analysis, the computation encounters a singularity and thus needs to pick one of the variables to ignore.

Question 6

```
> ex4

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

```
Coefficients:
  (Intercept)          Floor            Lot           Bath            Bed
Year
   -131.8769        72.5267        10.5655         5.4472       -12.0432
0.1271
        Age            Gar   StatusPending     StatusSold    SchoolCrest
SchoolEdison
         NA         7.6951       -19.0634       -37.0352         1.2157
84.5645
 SchoolHarris   SchoolParker   SchoolRedwood
     50.3965       -14.3984         4.6459
```

We go over the significant variables.

- Floor: coefficient +72, i.e. better floor → higher costs (makes sense)
- Lot: coefficient +10, i.e. better lot → higher costs (makes sense)
- Bed: coefficient -12, i.e. more bedrooms → lower price (DOES NOT make sense)
- Gar: coefficient +7, i.e. more garage space → higher price (makes sense)
- School: different coefficients per district, only one negative. This makes sense if the Parker district is one of the worse districts.

Question 7

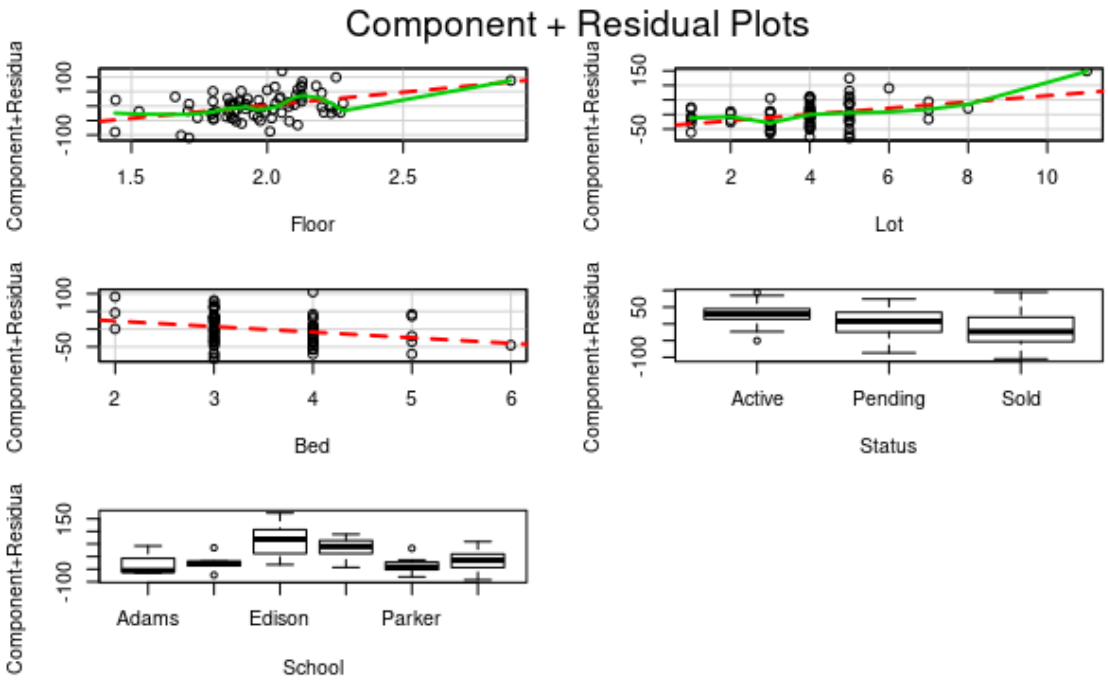
The final model uses Floor, Lot, Bed, Status, School.

```
> ex7f <- step(ex7, direction="both")
Start:  AIC=591.14
Price ~ Floor + Lot + Bath + Bed + Age + Gar + Status + School

       Df Sum of Sq    RSS    AIC
- Age     1       295 125846 589.32
- Bath    1       436 125986 589.40
- Gar     1      1338 126888 589.95
<none>                125551 591.14
- Bed     1      3542 129092 591.25
- Status  2     14279 139830 595.33
- Floor   1     11117 136667 595.59
- Lot     1     16862 142412 598.72
- School  5     72617 198168 615.83

Step:  AIC=589.32
Price ~ Floor + Lot + Bath + Bed + Gar + Status + School

       Df Sum of Sq    RSS    AIC
- Bath    1       658 126504 587.71
- Gar     1      2066 127912 588.56
<none>                125846 589.32
- Bed     1      4501 130347 589.99
+ Age     1       295 125551 591.14
- Status  2     14557 140403 593.64
- Floor   1     11371 137216 593.89
```

```
- Lot     1     16590 142435 596.73
- School  5     78087 203933 616.01

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

       Df Sum of Sq    RSS    AIC
- Gar     1      2432 128936 587.16
<none>                126504 587.71
- Bed     1      4148 130652 588.17
+ Bath    1       658 125846 589.32
+ Age     1       518 125986 589.40
- Status  2     15174 141678 592.32
- Lot     1     15995 142499 594.76
- Floor   1     16864 143368 595.23
- School  5     80042 206546 614.97

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

       Df Sum of Sq    RSS    AIC
<none>                128936 587.16
+ Gar     1      2432 126504 587.71
+ Age     1      1550 127386 588.24
+ Bath    1      1024 127912 588.56
- 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
```

Question 8

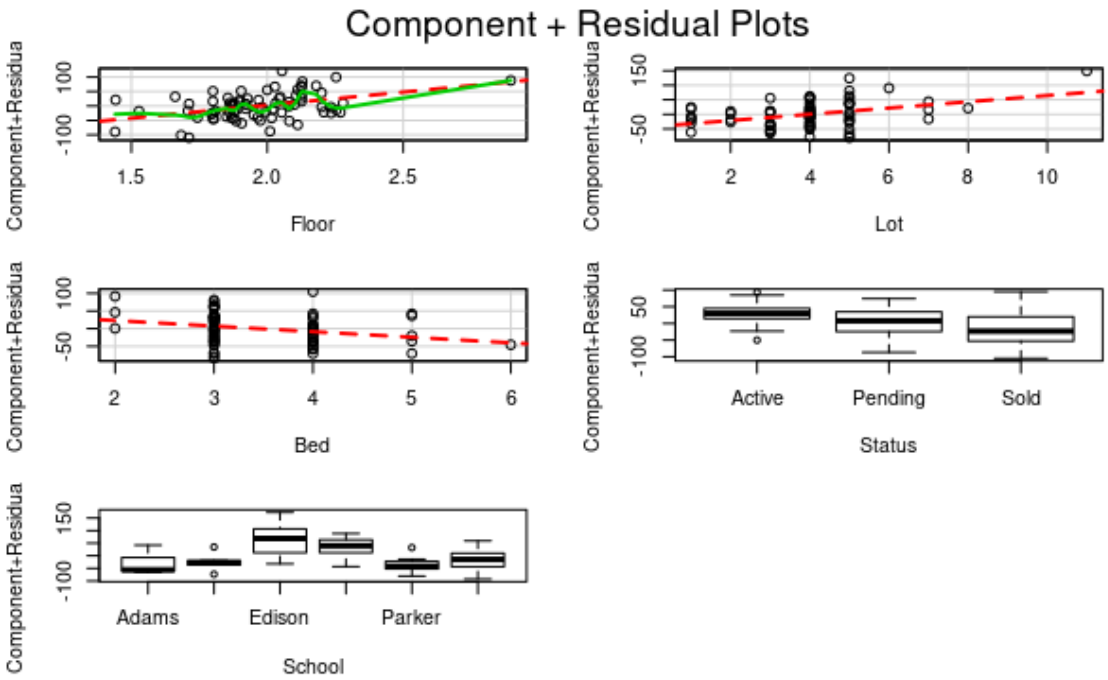
```
> crPlots(ex7f)
Warning message:
In smoother(.x, partial.res[, var], col = col.lines[2], log.x = FALSE,  :
  could not fit smooth
```



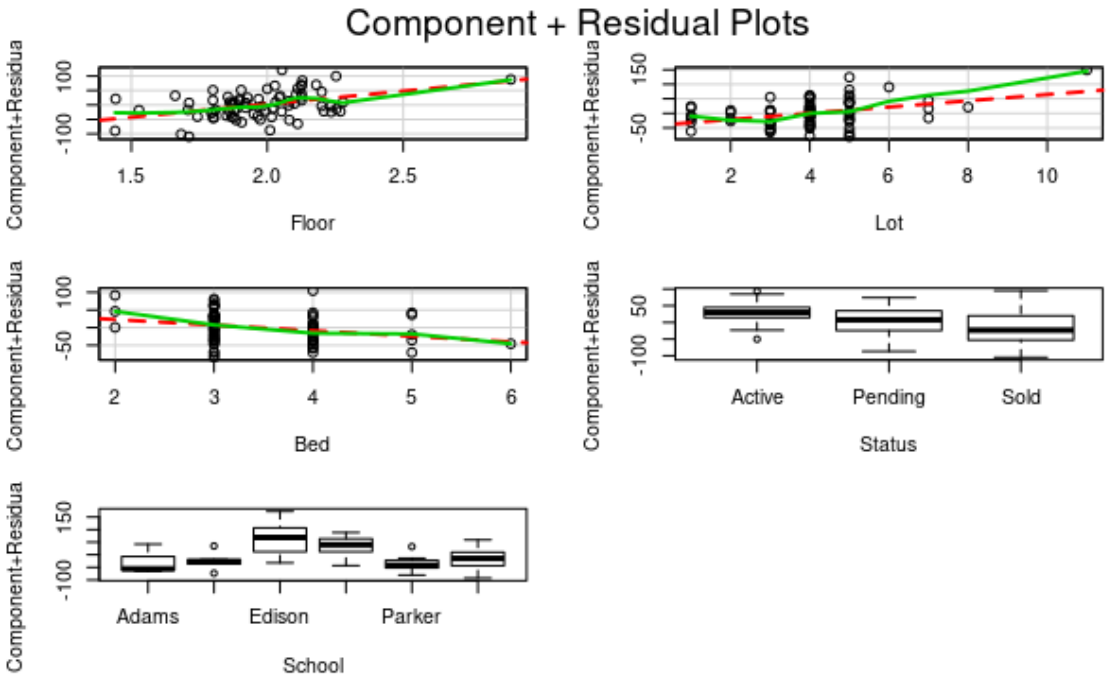
(A) Some square terms should be included for the Lot and Floor components

(B) Neither of shows a better smoothing.

```
> crPlots(ex7f, smoother.args=list(span=0.25))
Warning messages:
1: In smoother(.x, partial.res[, var], col = col.lines[2], log.x = FALSE,  :
  could not fit smooth
2: In smoother(.x, partial.res[, var], col = col.lines[2], log.x = FALSE,  :
  could not fit smooth
```




```
> crPlots(ex7f, smoother.args=list(span=0.75))
```



(C) the AIC got lower to 802.0374, so it does not neccessarily make sense.

```
> ex8c <- lm(formula = Price ~ Floor + Lot + Bed + Status + School + I(Lot^2), data
= homes)
```

```
> AIC(ex8c)
[1] 802.0374
```