Statistical Modeling with R - Fall 2016
Homework 2

DUE IN: Tuesday, 27.09.2016 at 11:59,

HOW: electronically in pdf-format via submission to `www.turnitin.com`

Class id: 13494794

enrollment password: Ti20Ta16Nic

Please register for the class on turnitin ahead of time.

GROUP WORK: is allowed with a maximum of 3 persons per group. PLEASE stay within the same group throughout the semester. Only one solution is accepted and graded per group. Please include the names of all group members on each assignment.

HOW MANY: There will be a total of six homework assignments in this semester. We will do a random selection of questions to be graded. Each week a total of eight points can be gained. Only the five best homeworks will be counted.

DUE DATES: 20.09., 27.09., 04.10., 11.10., 18.10., 25.10. (tentatively, subject to change)

FORMAT: Please do the required analyses and provide answers in complete sentences. **Provide the R syntax for the commands.** Just report those statistics that are relevant; do not copy complete R output. Integrate requested figures or tables into your document and give a brief verbal comment/caption on them.

House Prices in Oregon

Economic theory tells us that house prices are based on a variety of features. The data file containing information on 77 single-family homes in Eugene, Oregon during 2005 was provided by Victoria Whitman, a Eugene realtor. We will model single-family home sale prices (Price, in thousands of dollars), which range from $155,000 to 450$,000, using some predictor variables.

Source Pardoe, I. (2012). *Applied Regression Modelling,* Wiley.

Variables Description of variables:

- **ID** identifier variable for each case
- **Price** house price in thousand US Dollars
- **Floor** floor size (thousands of square feet)
- **Lot** lot size category (categorized in groups from 1 (smallest) to 11 (largest))
- **Bath** number of bathrooms (with half-bathrooms counting as 0.1)
- **Bed** number of bedrooms (between 2 and 6)
- **Year** year in which home was built

Age age (standardized: (year built - 1970)/10)

Gar garage size (0, 1, 2, or 3 cars)

Status indicator with three categories: sold, pending, active

School elementary school districts (six categories: Adams, Crest, Edison, Harris, Parker, Redwood)

```
library(car)
library(MASS)
load("~/Data/OregonHomes.Rdata")
```

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., in `Rcmdr` you can find under the menu option `Data`, the menu `Manage variables in active data set` which comprises a function called `bin numeric variable`.]

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

   ```
   options(width=70)
   par(mfrow=c(1,1))
   homes$GarGroup <- recode(homes$Gar, "c(0,21)='Small'; else='Big'", as.factor.result=TRUE)
   #homes£GarGroup <- with(homes, bin.var(Gar, bins=2, method='intervals', labels=c('Small',
   boxplot(Price~GarGroup,data=homes,main="Sales price of homes",
   ylab="Sales price of homes in USD 1000's", xlab="Garage size",varwidth=TRUE)
   ```

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

   Yes, as can be seen in Figure 1 the medians differ substantially, and there is only a slight overlap of the boxes. Due to the small number of outliers and the tolerable skewness, the means will not be to far away from the medians. Since sample size is decent, the standard error of the mean will be much smaller than the visualised interquartile range.

   (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.

   ```
   housing.t<-t.test(Price~GarGroup, alternative='two.sided', conf.level=.95,
   var.equal=TRUE,  data=homes)
   ```

   There is a highly significant difference in average house prices between homes with a small garage (for one car at most) and homes with large garages (fro two and more cars) as derived by the independent-samples t-test (equal variances assumed) with a test-statistic of $t = 2.3861$ with 74 degrees of freedom yielding a p-value of $p < 0.0196$.

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

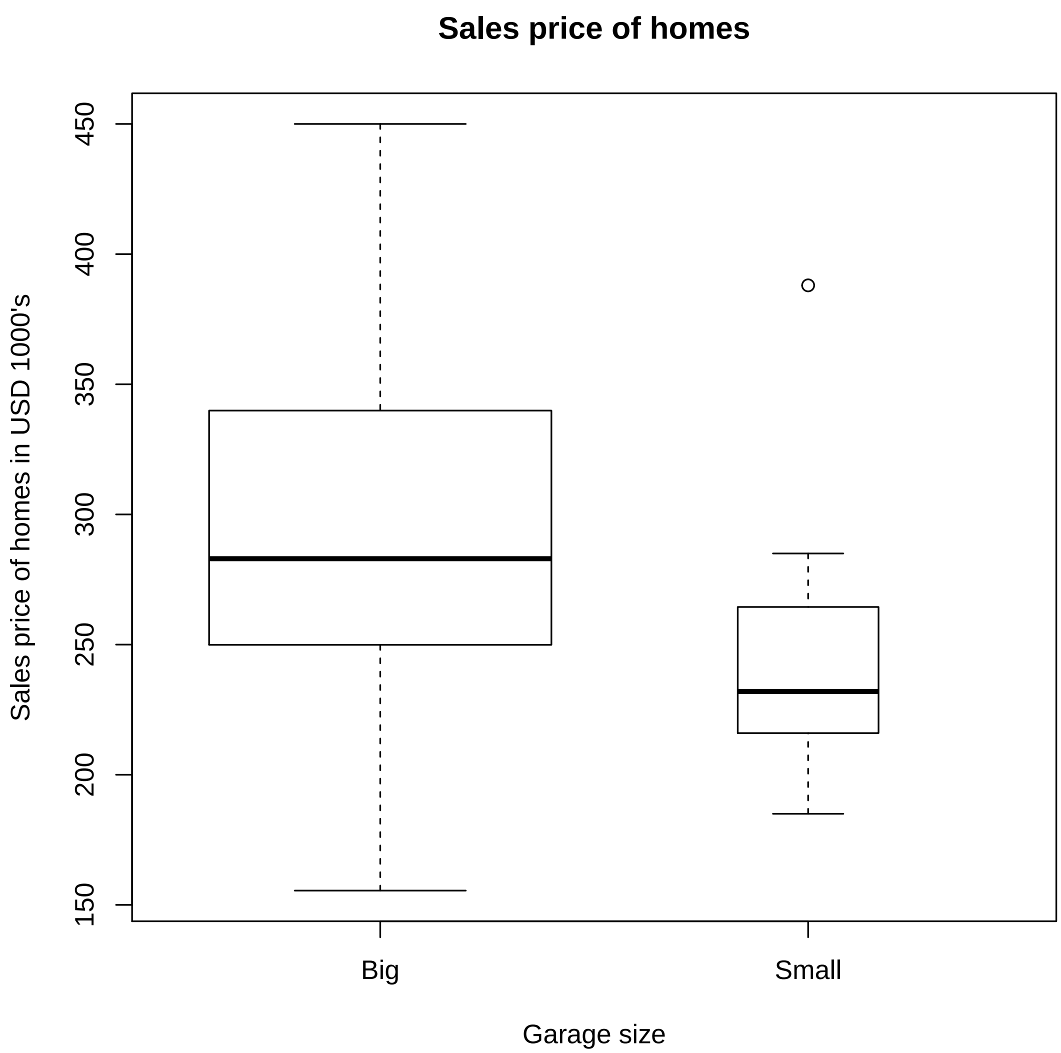


Figure 1: Box plot showing the different house prices in the two garage size groups.

```r
price.var<-tapply(homes$Price, homes$GarGroup, var, na.rm=TRUE)
lev.test<- leveneTest(homes$Price, homes$GarGroup, center=mean)
vartest<-var.test(Price ~ GarGroup, alternative='two.sided', conf.level=.95,
  data=homes)
```

The Levene's test for comparing the variances of house prices between between homes with a small garage (for one car at most) (var= 3462.4778) and homes with large garages (for two and more cars) (var= 3190.4927) yields a significant result with a p-value of 0.4554, NA.

[Alternatively: The variance test for comparing the variances of house prices homes with a small garage (for one car at most) (var= 3462.4778) and homes with large garages (for two and more cars) (var= 3190.4927) yields a significant result with a p-value of 0.9609.

The two groups can be assumed to have equal variance, hence we do not have to perform the Welch t-test.

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

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

```r
housing.aov <- aov(Price ~ GarGroup, data=homes)
summary(housing.aov)

##             Df Sum Sq Mean Sq F value Pr(>F)
## GarGroup     1  19504   19504   5.693 0.0196 *
## Residuals   74 253504    3426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

   The ANOVA yields a highly significant difference in average house prices between homes with a small garage (for one car at most) and homes with large garages (for two and more cars) showing a F-statistic of $F = 5.6933277$ with 74 residual degrees of freedom yielding a p-value of $p = 0.0195881, NA$.

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

```r
housing.lm<-lm(Price~GarGroup, data=homes)
summary(housing.lm)

##
## Call:
## lm(formula = Price ~ GarGroup, data = homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -136.886  -38.661   -9.986   44.364  157.614
```

```
##
## Coefficients:
##                Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      292.39       7.26  40.275 <0.0000000000000002 ***
## GarGroupSmall    -45.53      19.08  -2.386              0.0196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.53 on 74 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.07144,Adjusted R-squared:  0.05889
## F-statistic: 5.693 on 1 and 74 DF,  p-value: 0.01959
```

There is a highly significant difference in house prices between homes with a small garage (for one car at most) and homes with large garages (for two and more cars) showing a F-statistic of $F = 6$ with 1 numerator and 74 denominator degrees of freedom yielding a p-value of $p = 0.0196$.

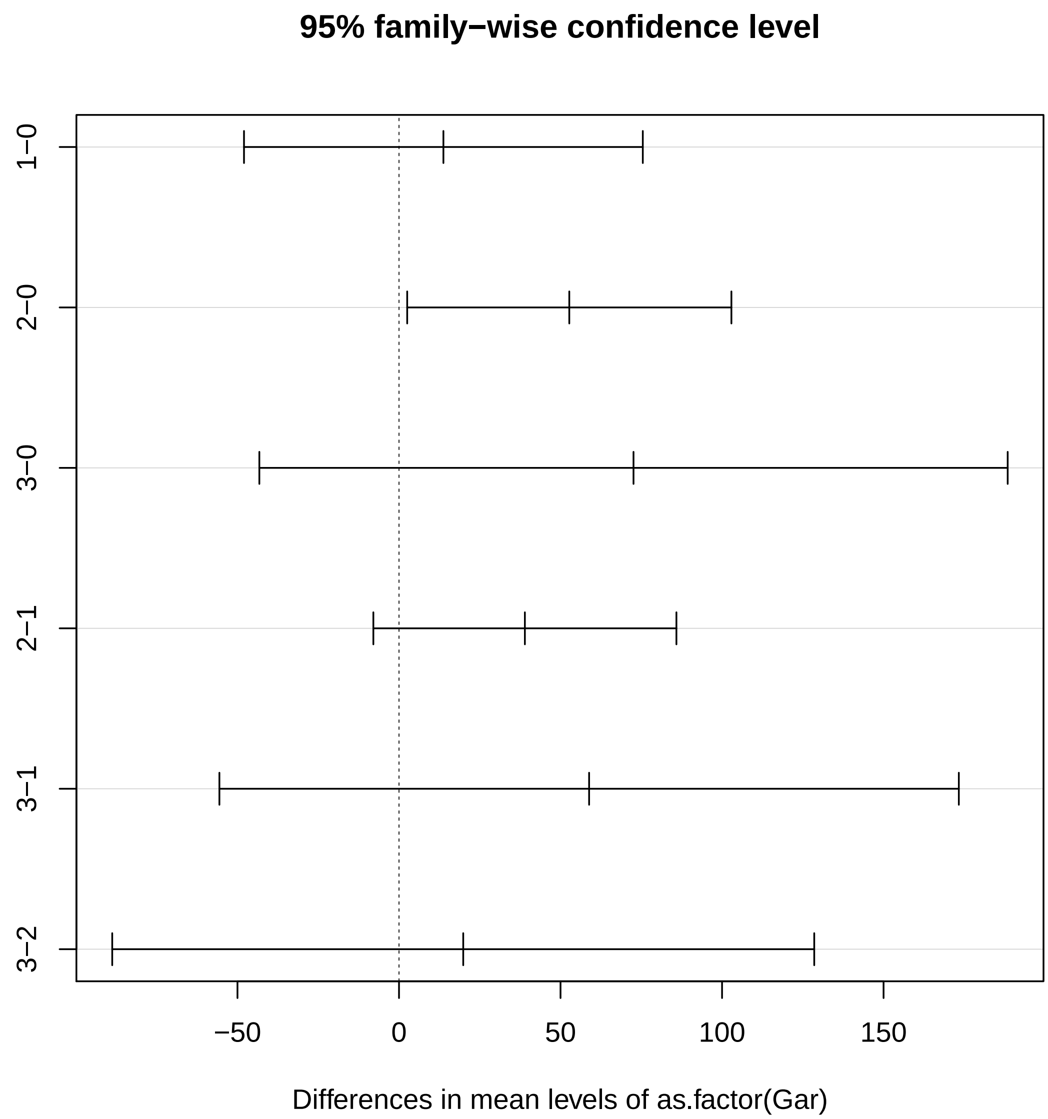
(c) (1 point) Compare the results of the t-test, the linear model and the ANOVA. How do the p-values of the three tests relate to each other? How do the test statistics of the three tests relate to each other?

The results of the t-test, the linear model and the ANOVA are identical.The p-values are the same. The F-value of the ANOVA and the linear model is just the square of the t-statistic from the t-test. The absolute value of the t-test statistic is identical to the t-value for the predictor in the linear model.

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.

```
price.gar2 <- aov(Price~as.factor(Gar), data=homes)
summary(price.gar2)

##                  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

TukeyHSD(price.gar2)

##   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

plot(TukeyHSD(price.gar2))
```



(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!
Yes, $p = 0.015$.

(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.

```
TukeyHSD(price.school)

## Error in TukeyHSD(price.school):  object 'price.school' not found

plot(TukeyHSD(price.school))

## Error in TukeyHSD(price.school):  object 'price.school' not found
```

At the 5% significance level, the house prices in the following garage sizes differ:
2 cars and 0 cars

(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.

```
numSummary(homes[,"Price"], groups=as.factor(homes$Gar), statistics=c("mean", "sd"))

## Error in eval(expr, envir, enclos):  could not find function "numSummary"

homes.means <- tapply(homes$Price, homes$Gar, mean, na.rm=TRUE)
homes.sd <- tapply(homes$Price, homes$Gar, sd, na.rm=TRUE)
homes.n <- tapply(homes$Price, homes$Gar, length)
cbind(homes.means, homes.sd, homes.n)

##   homes.means homes.sd homes.n
## 0    246.8545 56.48445      11
## 1    260.6000 67.44587      13
## 2    299.5680 55.14291      51
## 3    319.4500 28.92067       2
```

The comparisons depend not only on the difference in means but also on the standard error of the mean difference between the two groups. The standard error of the mean difference depends on the standard deviation and the samples sizes in each group. Since there are only two houses with a garage for three cars, the standard error for this group mean difference is rather large. In opposite, there are 50 homes with garages for 2 cars in the data set and 11 homes without garage. This mean difference has a ways smaller standard error.

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).

```
housing.all <- lm(Price ~ . - GarGroup, data=homes)
summary(housing.all)
```

```
##
## Call:
## lm(formula = Price ~ . - GarGroup, 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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

```
Anova(housing.all, type="II")
```

```
## Anova Table (Type II tests)
##
## Response: Price
##           Sum Sq Df F value     Pr(>F)
## ID          1837  1  0.9059   0.344950
## Floor      12742  1  6.2827   0.014873 *
## Lot        16095  1  7.9360   0.006521 **
## Bath         286  1  0.1412   0.708416
## Bed         4077  1  2.0104   0.161315
## Year         464  1  0.2290   0.633962
## Age                0
```

```
## Gar           832  1  0.4104   0.524144
## Status      14577  2  3.5937   0.033466 *
## School      74329  5  7.3300 0.00001959 ***
## Residuals 123713 61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Looking at the ANOVA table, we see that `Floor`, `Lot`, `Status` and `School` are significant at the 5% level.

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

The model fits fairly poorly as indicated by *adjusted R-squared*, which means that the model explains just about 44 percent of the variability in house prices.

(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?

As a numeric one, since there is only 1 df for `Gar` indicated in the ANOVA-table.

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!

`Age` is a variable that is linearly dependent on `Year`. Hence, only one of the two can be included in any linear model. By default, the one that occurs earlier in the model specification is included, the othe rone omitted.

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

Out of the significant predictors three have a positive sign, namely `Floor`, `Lot` and `SchoolEdison`, while one has a negative sign: `StatusSold`.

For the cateogircal predictors, the coefficients indicate the differenc to the reference category. So, according to the model house prices in school district Edison are significantly higher than in school district Adam. Not knowing the details, I can't asses plausibility, but it seems plausible in general.

The negative sign for the coefficient of `StatusSold` means that on average house prices of homes that are sold ar elower than prices for the homes that are still active on the market. A very reasonable result, since i assume that everything else being the same, buyers will opt for the cheaper houses first, more expensive ones will only be sold later.

It also appears plausible that house prices increase by floor and lot size.

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!

```
housing.base <- lm(Price ~ . - GarGroup - Year, data=homes)
#housing.best <- stepwise(housing.base, direction='backward/forward', criterion='AIC')
housing.best <- stepAIC(housing.base, scope=list(upper=housing.base,lower=~1),direction='b
```

```
## Start:  AIC=592.02
## Price ~ (ID + Floor + Lot + Bath + Bed + Year + Age + Gar + Status +
##     School + GarGroup) - GarGroup - Year
##
##          Df Sum of Sq    RSS    AIC
## - Bath    1       286 124000 590.19
## - Age     1       464 124178 590.30
## - Gar     1       832 124546 590.53
## - ID      1      1837 125551 591.14
## <none>                123713 592.02
## - Bed     1      4077 127791 592.48
## - Status  2     14577 138290 596.48
## - Floor   1     12742 136455 597.47
## - Lot     1     16095 139808 599.31
## - School  5     74329 198042 617.78
##
## Step:  AIC=590.19
## Price ~ ID + Floor + Lot + Bed + Age + Gar + Status + School
##
##          Df Sum of Sq    RSS    AIC
## - Age     1       700 124700 588.62
## - Gar     1       855 124855 588.72
## - ID      1      1987 125986 589.40
## <none>                124000 590.19
## - Bed     1      3828 127828 590.51
## + Bath    1       286 123713 592.02
## - Status  2     14920 138919 594.83
## - Lot     1     15830 139829 597.33
## - Floor   1     17186 141185 598.06
## - School  5     78656 202656 617.53
##
## Step:  AIC=588.62
## Price ~ ID + Floor + Lot + Bed + Gar + Status + School
##
##          Df Sum of Sq    RSS    AIC
## - Gar     1      1795 126495 587.71
## - ID      1      1804 126504 587.71
## <none>                124700 588.62
## - Bed     1      4917 129618 589.56
## + Age     1       700 124000 590.19
## + Bath    1       522 124178 590.30
## - Status  2     15473 140173 593.51
## - Lot     1     15135 139835 595.33
```

```
## - Floor   1     18607 143307 597.19
## - School  5     81375 206075 616.80
##
## 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
## + Gar     1      1795 124700 588.62
## + Age     1      1641 124855 588.72
## + Bath    1       776 125719 589.24
## - 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
## + 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
```

```
summary(housing.best)
```

```
##
## 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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

The backward/forward strategy and the AIC as criterion remove the unsignificant predictors `ID, Gar, Bath` and `Age` from the model. While *Floor*, *Lot*, *Status* and `School` are significant at the 1% level, `Bed` falls above the 5% threshold by only a very small margin. As indicated by *adjusted R-squared*, the model fits mediocrely with explaining about 45% variability in the house prices. The AIC resulting in 1576.68 as compared to 1580.63 for the starting model, i.e. an improvement of 0.25%.

8. Draw component/residual plots for all predictors in the final model resulting in the previous task.

```
par(mfrow=c(4,3))
crPlots(housing.best)

## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x = FALSE,
:  could not fit smooth
```

(a) (half a point) Check whether some quadratic effects should be included.
For floor size a quadratoc effect seems reasonable. Also for lot size one can include one, although the variable is categorized and hence the points are aligned on paralell lines.

(b) (half a point) Vary the smoothing parameter to 0.25 and to 0.75. Which parameter setting indicates the quadratic effects more clearly?

```
par(mfrow=c(4,3))
crPlots(housing.best, span = 0.25)

## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x =
FALSE, :  could not fit smooth
## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x =
FALSE, :  could not fit smooth
```

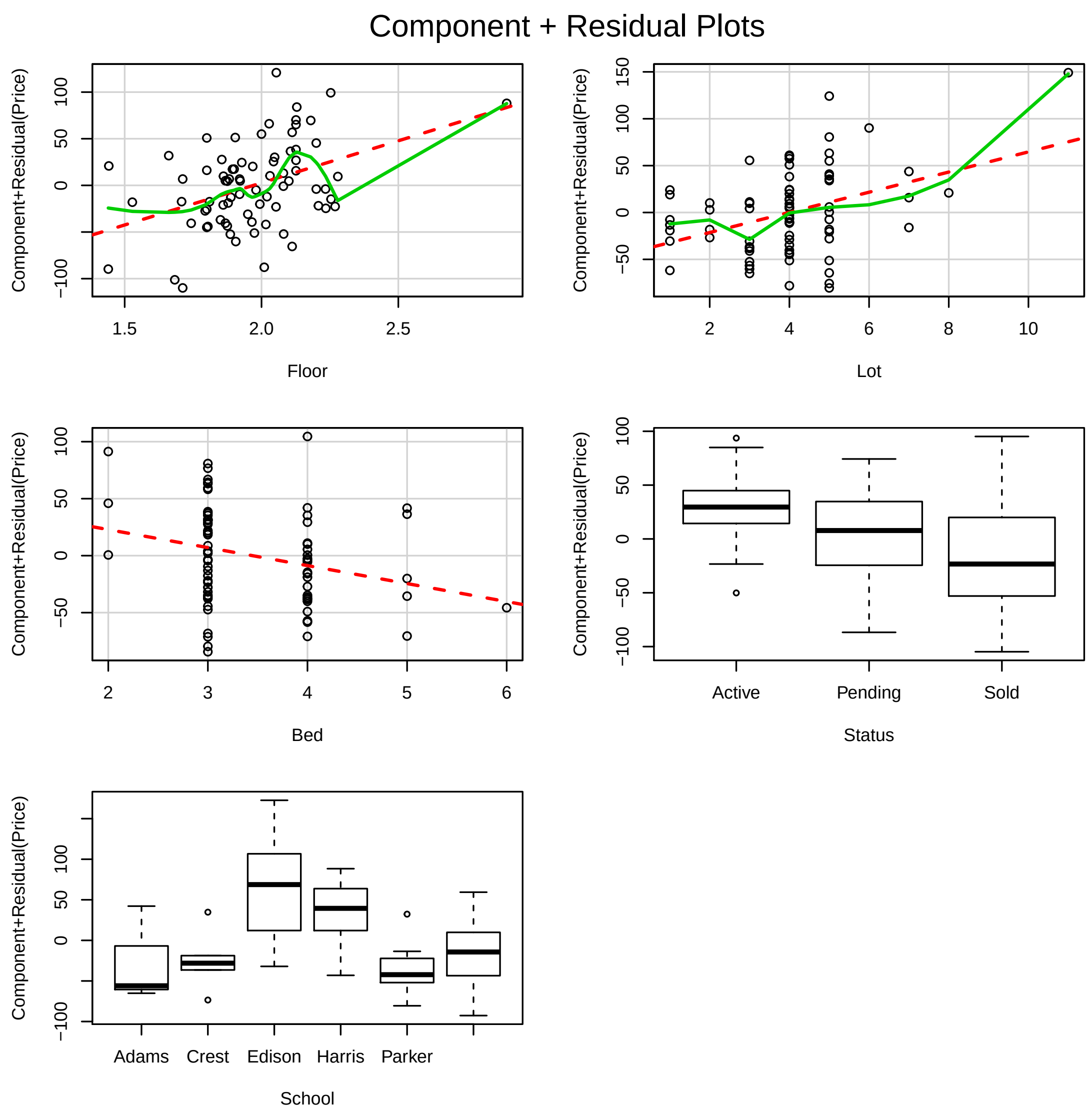


Figure 2: Component residual plots indicating the possibility of quadratic effects for floor and lot size.

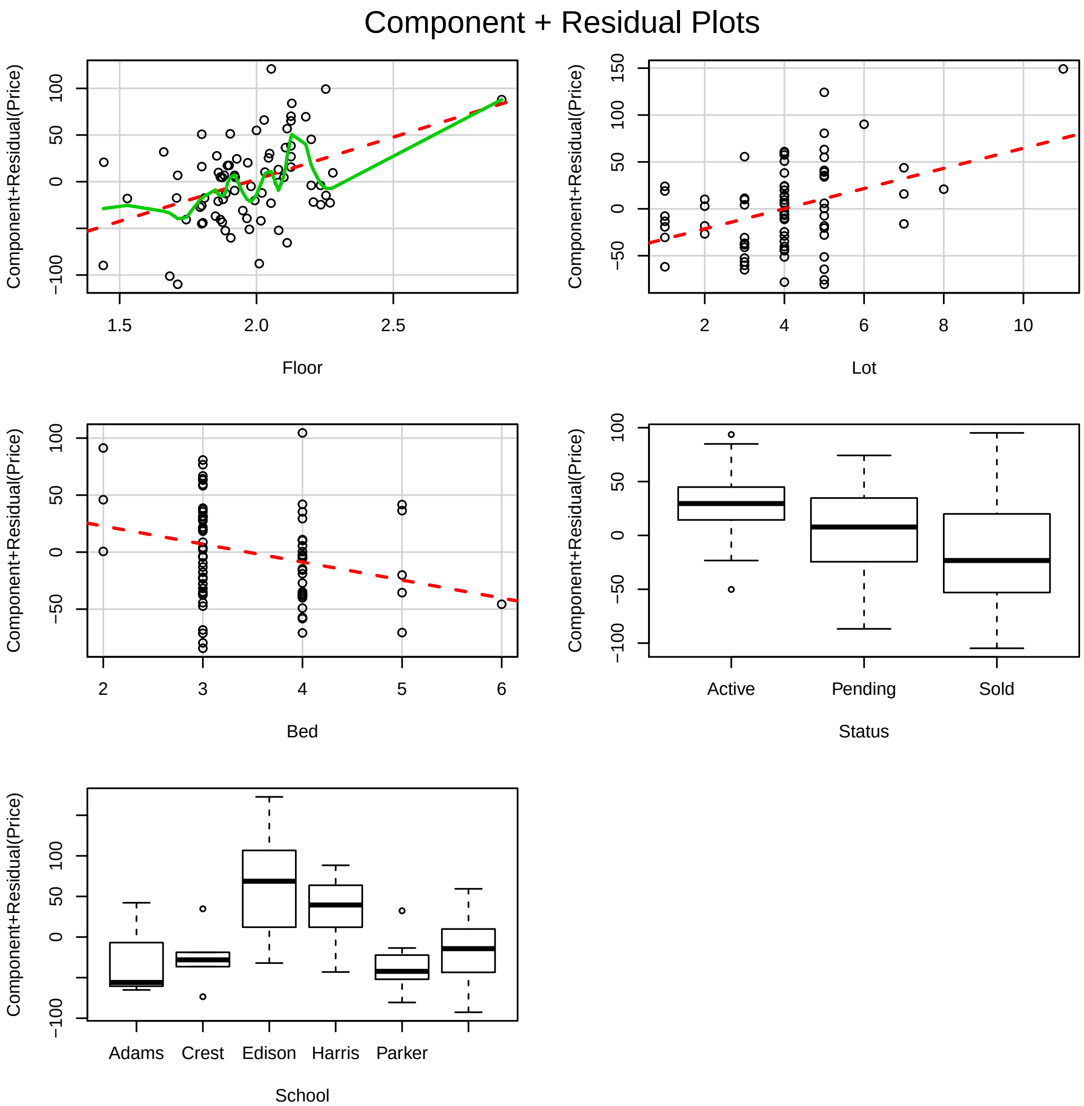


Figure 3: Component residual plots using span = 0.25 for smoothing

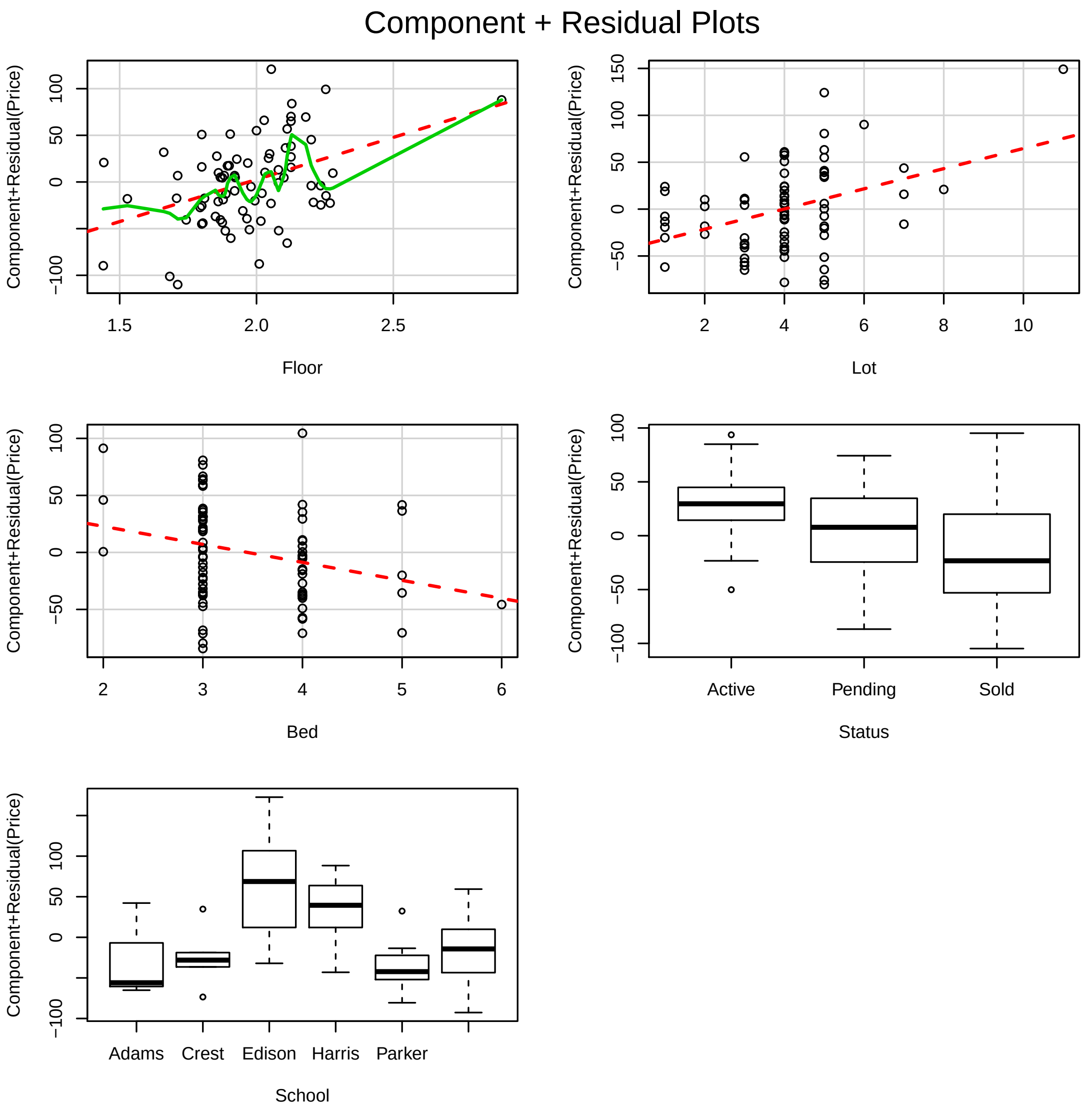


Figure 4: Component residual plots using span = 0.75 for smoothing

```
par(mfrow=c(4,3))
crPlots(housing.best, span = 0.25)

## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x =
FALSE, : could not fit smooth
## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x =
FALSE, : could not fit smooth
```

(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?

```
housing.best.sq <- lm(Price ~ Floor + Lot + Bed + Status + School + I(Floor^2) + I(Lo
summary(housing.best.sq)
```

```
##
## Call:
## lm(formula = Price ~ Floor + Lot + Bed + Status + School + I(Floor^2) +
##     I(Lot^2), data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.572 -27.268  -1.292  26.776 119.231
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   230.0635   242.7884   0.948  0.34696
## Floor          17.2037   239.0224   0.072  0.94285
## Lot            -7.1741     9.4492  -0.759  0.45055
## Bed           -14.9096     7.9291  -1.880  0.06468 .
## StatusPending -21.0228    16.3416  -1.286  0.20299
## StatusSold    -44.2770    13.3714  -3.311  0.00154 **
## SchoolCrest    10.4169    33.5104   0.311  0.75694
## SchoolEdison   94.6923    29.5892   3.200  0.00215 **
## SchoolHarris   60.2284    30.2061   1.994  0.05049 .
## SchoolParker   -1.2396    30.4489  -0.041  0.96766
## SchoolRedwood  20.0570    28.7107   0.699  0.48738
## I(Floor^2)     15.1187    57.6915   0.262  0.79413
## I(Lot^2)        1.8671     0.9174   2.035  0.04603 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.81 on 63 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5571,Adjusted R-squared:  0.4728
## F-statistic: 6.604 on 12 and 63 DF,  p-value: 0.0000001781
```

```
anova(housing.best, housing.best.sq)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Floor + Lot + Bed + Status + School
## Model 2: Price ~ Floor + Lot + Bed + Status + School + I(Floor^2) + I(Lot^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     65 128936
## 2     63 120908  2    8027.9 2.0915  0.132
```

Using the anova-test to compare the two models we see that there is no signficant improvement. We can hence stick with the simpler model.