Statistical Modeling with R - Fall 2016
Homework 1

DUE IN: Tuesday, 20.09.2016 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)

```
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. Plot a box plot of the house prices using the school districts (variable `School`) as grouping factor.

```
options(width=70)
par(mfrow=c(1,1))
boxplot(Price~School,data=homes,main="Sales price of homes",
ylab="Sales price of homes in USD 1000's", xlab="School District",varwidth=TRUE)
```

   (a) (half a point) Are half of the houses in school district *Crest* priced at least as high as three quarters of the houses in School district *Adams*?
   Yes, as can be seen in Figure 1 by comparing the median line in the boxplot for school district *Crest* with the upper end of the box for school district *Adams*.

   (b) (half a point) As measured by the interquartile range, which school district shows the smallest spread in house prices?
   Parker, as can be seen in Figure 1 by comparing the height of all boxes.

   (c) (1 point) Looking at the boxplot does homoscedasticity hold for house prices in the six school districts? Give reasons for your answer!
   No, interquartile ranges are different; IQRs are small for houses in school districts Harris and Parker, larger in all others. The length of the whiskers are rather similar except the upper whisker for Harris. In Parker and Redwood there are outliers to the higher end of prices.

2. Using the variable `School` as a factor, run an ANOVA model to see whether the school district has a statistically significant impact on the average house price.

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

##                   Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(School)  5  60573   12115   3.992 0.00304 **
## Residuals         70 212434    3035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

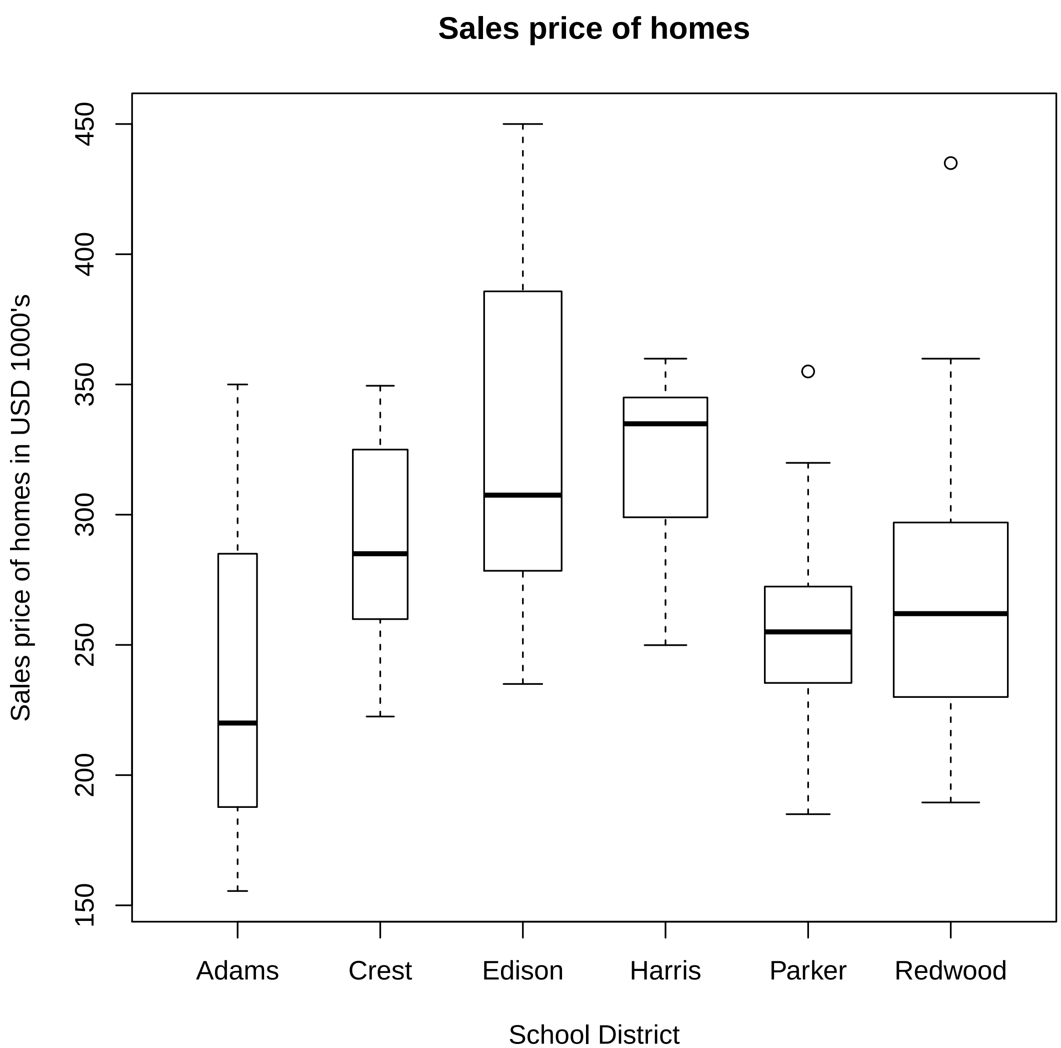
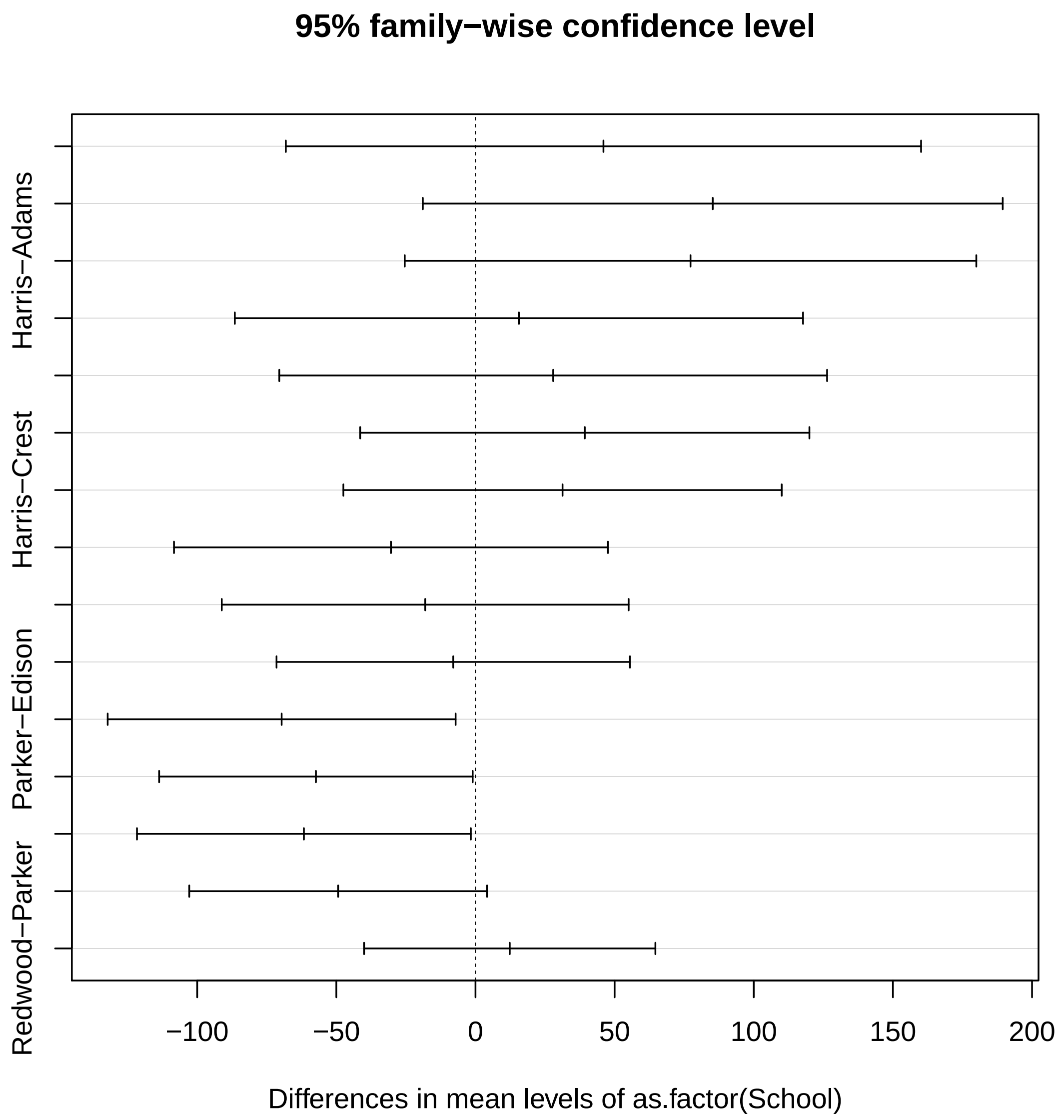


Figure 1: Box plot showing the different house prices in the six school districts.

```r
TukeyHSD(price.school)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ as.factor(School), data = homes)
##
## $`as.factor(School)`
##                      diff        lwr        upr     p adj
## Crest-Adams     45.983333  -68.15614 160.122811 0.8446444
## Edison-Adams    85.266667  -18.92794 189.461278 0.1711893
## Harris-Adams    77.273810  -25.42152 179.969140 0.2487557
## Parker-Adams    15.613333  -86.47612 117.702786 0.9976307
## Redwood-Adams   27.924359  -70.50000 126.348714 0.9606625
## Edison-Crest    39.283333  -41.42547 119.992132 0.7111287
## Harris-Crest    31.290476  -47.47325 110.054199 0.8521953
## Parker-Crest   -30.370000 -108.34211  47.602107 0.8623473
## Redwood-Crest  -18.058974  -91.16675  55.048801 0.9783770
## Harris-Edison   -7.992857  -71.49420  55.508486 0.9990708
## Parker-Edison  -69.653333 -132.17010  -7.136566 0.0202274
## Redwood-Edison -57.342308 -113.67563  -1.008983 0.0436513
## Parker-Harris  -61.660476 -121.64514  -1.675812 0.0404140
## Redwood-Harris -49.349451 -102.85886   4.159962 0.0875809
## Redwood-Parker  12.311026  -40.02618  64.648228 0.9825923

plot(TukeyHSD(price.school))
```



(a) (half a point) Does the test result indicate that school districts have a statistically significant impact on house prices?
Yes

(b) (half a point) Report the observed p-value for the overall ANOVA test!
0.00304

(c) (half a point) Which F-distribution is used to derive the p-value? F-distribution with 5 numerator and 70 denominator degrees of freedom, $F_{5,70}$-distribution.

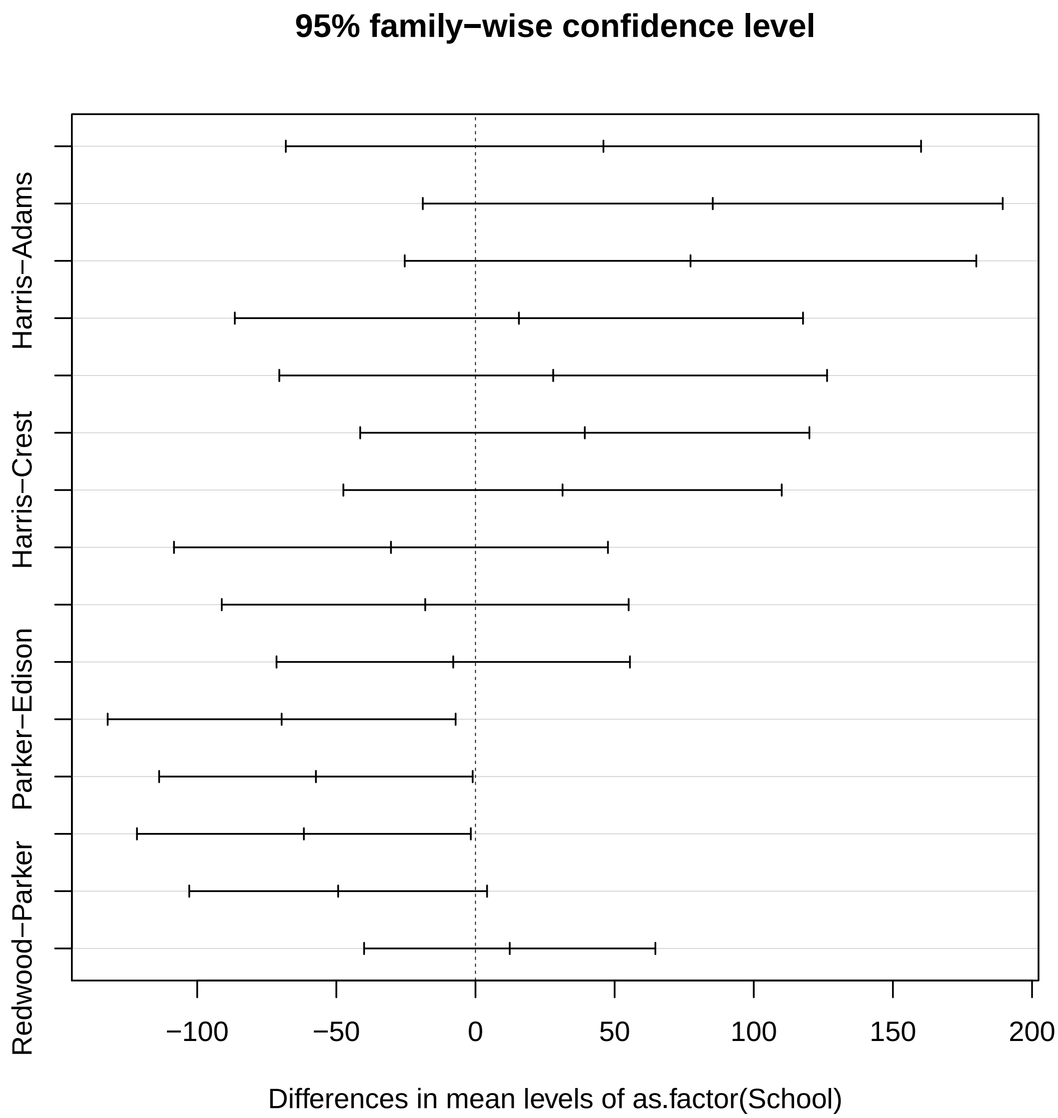
(d) (half a point) Which percentage of the total sum of squares is explained by the model? Divide the sum of squares for school by the sum taken over the sum of squares for the residual and the sum of squares fo school (to get pecentages you can also multiply by 100 or interpret the decimal number as percentage), i.e. $60573/(212434 + 60573) = 0.2219$

3. (2 points) Use the Tukey HSD post hoc test to determine which schools differ in average house prices at the 5% significance level. Visualise the result of the Tukey HSD test.

```r
TukeyHSD(price.school)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ as.factor(School), data = homes)
##
## $`as.factor(School)`
##                      diff        lwr        upr     p adj
## Crest-Adams     45.983333  -68.15614 160.122811 0.8446444
## Edison-Adams    85.266667  -18.92794 189.461278 0.1711893
## Harris-Adams    77.273810  -25.42152 179.969140 0.2487557
## Parker-Adams    15.613333  -86.47612 117.702786 0.9976307
## Redwood-Adams   27.924359  -70.50000 126.348714 0.9606625
## Edison-Crest    39.283333  -41.42547 119.992132 0.7111287
## Harris-Crest    31.290476  -47.47325 110.054199 0.8521953
## Parker-Crest   -30.370000 -108.34211  47.602107 0.8623473
## Redwood-Crest  -18.058974  -91.16675  55.048801 0.9783770
## Harris-Edison   -7.992857  -71.49420  55.508486 0.9990708
## Parker-Edison  -69.653333 -132.17010  -7.136566 0.0202274
## Redwood-Edison -57.342308 -113.67563  -1.008983 0.0436513
## Parker-Harris  -61.660476 -121.64514  -1.675812 0.0404140
## Redwood-Harris -49.349451 -102.85886   4.159962 0.0875809
## Redwood-Parker  12.311026  -40.02618  64.648228 0.9825923

plot(TukeyHSD(price.school))
```



At the 5% significance level, the house prices in the following School districts differ:
Parker and Edison
Redwood and Edison
Parker and Harris

4. (2 points) Report the adjusted p-values for the significant differences rounded to four digits.

   Parker and Edison: 0.0202
   Redwood and Edison: 0.0437
   Parker and Harris: 0.0404

5. Now run a linear model, i.e. use the R command `lm` to see whether the school district has a statistically significant impact on the average house price.

```
price.school.lm <- lm(Price~School, data=homes)
summary(price.school.lm)

##
## Call:
## lm(formula = Price ~ School, data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.100 -42.187  -1.602  30.205 165.242
##
## Coefficients:
##                 Estimate Std. Error t value       Pr(>|t|)
## (Intercept)       241.83      31.81   7.603 0.000000000098 ***
## SchoolCrest        45.98      38.95   1.180         0.2418
## SchoolEdison       85.27      35.56   2.398         0.0192 *
## SchoolHarris       77.27      35.05   2.205         0.0308 *
## SchoolParker       15.61      34.84   0.448         0.6554
## SchoolRedwood      27.92      33.59   0.831         0.4086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.09 on 70 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2219,Adjusted R-squared:  0.1663
## F-statistic: 3.992 on 5 and 70 DF,  p-value: 0.003041
```

(a) (1 point) According to this model, which school districts differ in average house prices at the 5% significance level?

Adams and Edison
Adams and Harris

(b) (1 point) Which percentage of the total variability is explained by this model? Compare this result to your response to Question 2d!

According to multiple R-Squared we get the same result as in Question 2d, namely $R^2 = 0.2219$.

6. Now create the ANOVA table for the linear model computed in Question 5

```
anova(price.school.lm)

## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value   Pr(>F)
## School     5  60573 12114.6  3.9919 0.003041 **
```

```
## Residuals 70 212434  3034.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (1 point) According to this table, do school districts have a statistically significant impact on house prices?

Yes, there is a significant impact. The table shows exactly the same numbers as the ANOVA model in Question 2.

(b) (1 point) The F-test performed in the ANOVA table is identical to the overall model test in the ANOVA model. The same test results are displayed also in the coefficient table created in Question 5. Where can you find it and which hypothesis is tested there?

In the bottom line of the table testing the hypothesis that the linear model with `School` as predictor is not different from the naive model just using mean house prices.

7. Now compute a linear model for the house prices using three predictors, namely, school districts, floor size and age.

```
price.sfa.lm <- lm(Price~School+ Floor + Age, data=homes)
summary(price.sfa.lm)

##
## Call:
## lm(formula = Price ~ School + Floor + Age, data = homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.303  -29.206   -2.259   26.305  145.408
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     153.755     60.216   2.553   0.0129 *
## SchoolCrest      27.984     37.985   0.737   0.4638
## SchoolEdison     89.130     34.891   2.555   0.0129 *
## SchoolHarris     47.615     34.643   1.374   0.1738
## SchoolParker     -5.155     34.123  -0.151   0.8804
## SchoolRedwood     9.111     32.692   0.279   0.7813
## Floor            53.443     29.826   1.792   0.0776 .
## Age               7.021      3.102   2.263   0.0268 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.31 on 68 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3186,Adjusted R-squared:  0.2484
## F-statistic: 4.541 on 7 and 68 DF,  p-value: 0.0003335
```

```r
anova(price.sfa.lm)

## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value   Pr(>F)
## School     5  60573 12114.6  4.4280 0.001501 **
## Floor      1  12383 12382.6  4.5260 0.037013 *
## Age        1  14012 14012.3  5.1217 0.026825 *
## Residuals 68 186039  2735.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (1 point) Use the ANOVA table to decide which predictors have a statistically significant impact at the 5% level?

According to the ANOVA table all three predictors are significant at the 5% level.

(b) (1 point) Is the model better than the one just using school districts as predictor?

Yes, adjusted R-squared increased.

8. Now add the interaction term between age and school district to the model from Question 7.

```r
price.sfa.int <- lm(Price~School*Age +Floor, data=homes)
summary(price.sfa.int)

##
## Call:
## lm(formula = Price ~ School * Age + Floor, data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.088 -30.144   0.684  18.109 136.764
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         153.66      59.41   2.586  0.01202 *
## SchoolCrest          49.34      37.42   1.319  0.19208
## SchoolEdison         83.55      35.90   2.327  0.02319 *
## SchoolHarris         69.02      34.22   2.017  0.04799 *
## SchoolParker         13.04      33.23   0.392  0.69603
## SchoolRedwood        24.90      31.84   0.782  0.43709
## Age                 -18.70      12.50  -1.496  0.13952
## Floor                43.61      29.80   1.463  0.14832
## SchoolCrest:Age      29.72      19.20   1.548  0.12675
## SchoolEdison:Age     16.40      13.55   1.210  0.23081
```

```
## SchoolHarris:Age     24.64      13.82   1.783  0.07938 .
## SchoolParker:Age     35.91      13.84   2.594  0.01179 *
## SchoolRedwood:Age    42.49      15.06   2.822  0.00638 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.02 on 63 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4455,Adjusted R-squared:  0.3399
## F-statistic: 4.218 on 12 and 63 DF,  p-value: 0.00007452
```

```
anova(price.sfa.int)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df Sum Sq Mean Sq F value    Pr(>F)
## School      5  60573 12114.6  5.0419 0.0005995 ***
## Age         1  17611 17611.3  7.3295 0.0087202 **
## Floor       1   8784  8783.6  3.6556 0.0604315 .
## School:Age  5  34664  6932.7  2.8853 0.0207967 *
## Residuals  63 151376  2402.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (1 point) Use the ANOVA table to decide which predictors have a statistically significant impact at the 5% level?

According to the ANOVA table now only School and age as well as their interaction are significant at the 5% level.

(b) (1 point) Which slope is estimated for age in the coefficient table? To which school district does this refer to?

The estimated slope for age is -18.7. It refers to the School district *Adam*.

9. (2 points) For which school districts is the impact of age on house prices negative, for which is it positive? To answer this question one has to add the coefficient of the corresponding interaction term to the coefficient of age. Hence we get:

```
slopes <- price.sfa.int$coef[7]+c(0,price.sfa.int$coef[9:13])
names(slopes)<-levels(homes$School)
slopes
```

```
##      Adams      Crest     Edison     Harris     Parker    Redwood
## -18.703990  11.012173  -2.305884   5.936948  17.201139  23.785377
```

So, slopes are negative for school districts *Adams* and *Edison*, positive for all others.

10. Now add a quadratic term of age to the linear model created in Question 8.

```
price.sfa.int.sq <- lm(Price~School*Age +Floor + I(Age^2), data=homes)
summary(price.sfa.int.sq)
```

```
##
## Call:
## lm(formula = Price ~ School * Age + Floor + I(Age^2), data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.155 -29.260   0.782  18.355 136.642
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       153.63989   59.88955   2.565  0.01274 *
## SchoolCrest        49.05226   38.14282   1.286  0.20322
## SchoolEdison       83.43400   36.25764   2.301  0.02476 *
## SchoolHarris       69.08928   34.52322   2.001  0.04975 *
## SchoolParker       12.93198   33.56281   0.385  0.70133
## SchoolRedwood      24.61902   32.55834   0.756  0.45242
## Age               -18.72012   12.60266  -1.485  0.14250
## Floor              43.80471   30.26689   1.447  0.15286
## I(Age^2)           -0.05688    1.10018  -0.052  0.95894
## SchoolCrest:Age    29.56206   19.58465   1.509  0.13626
## SchoolEdison:Age   16.19368   14.22173   1.139  0.25923
## SchoolHarris:Age   24.66900   13.94029   1.770  0.08171 .
## SchoolParker:Age   35.97354   14.01646   2.567  0.01270 *
## SchoolRedwood:Age  42.55728   15.23388   2.794  0.00693 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.41 on 62 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4455,Adjusted R-squared:  0.3293
## F-statistic: 3.832 on 13 and 62 DF,  p-value: 0.0001608
```

```
anova(price.sfa.int.sq)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df Sum Sq Mean Sq F value    Pr(>F)
## School      5  60573 12114.6  4.9621 0.0006941 ***
## Age         1  17611 17611.3  7.2135 0.0092730 **
## Floor       1   8784  8783.6  3.5977 0.0625198 .
## I(Age^2)    1   5345  5345.4  2.1894 0.1440249
```

```
## School:Age  5  29325  5864.9  2.4022 0.0468535 *
## Residuals  62 151369  2441.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (half a point) Is the quadratic effect of age statistically significant?

The quadratic effect of age is statistically not significant.

(b) (1.5 points) Which of the two models (Question 8 and Question 10) do you prefer? Explain your preference?

I prefer the one without the quadratic effect of age, since it has the higher adjusted R-quared value.