Assignment 1 - Garage Influence

Aliaksandr Panko, Nadzeya Yakimchyk, Nikolai Agafonov, Ilia Kurennoi

May 22, 2018

Contents

1 Objectives

1. Determine does garage-dependent price effect is statistically significant.
2. If possible, distinguish among different types (attached, detached, etc.)
3. If possible, distinguish among different capacity of a garage (number of cars)
4. How much would you spend on building a garage before selling a house?

2 Choose the model

Let's start with the regression where all factors are statistically significant

2.1 R-code 1

```
train <- read.csv2(file.choose(), header=TRUE, sep=",", na.strings="")

train["NbrBrkSide"] <- train$Neighborhood == "BrkSide"
train["NbrCrawfor"] <- train$Neighborhood == "Crawfor"
train["NbrStoneBr"] <- train$Neighborhood == "StoneBr"

regr <- lm(log(SalePrice) ~  log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
NbrBrkSide +
Fireplaces +
as.factor(Zone) +
log(X1stFlrSF),train)

summary(regr)
```

2.2 Regression Result 1

```
Call:
lm(formula = log(SalePrice) ~ log(YearBuilt) + log(LotArea) +
    log(LivAreaSF) + OverallQual + OverallCond + NbrCrawfor +
    NbrStoneBr + NbrBrkSide + Fireplaces + as.factor(Zone) +
    log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48599 -0.06642  0.01128  0.07749  0.42240

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -59.206979   4.242214 -13.957  < 2e-16 ***
log(YearBuilt)      8.638165   0.555796  15.542  < 2e-16 ***
log(LotArea)        0.105533   0.014148   7.459 2.52e-13 ***
log(LivAreaSF)      0.349084   0.025878  13.490  < 2e-16 ***
OverallQual         0.093664   0.006727  13.923  < 2e-16 ***
OverallCond         0.053012   0.005833   9.089  < 2e-16 ***
NbrCrawforTRUE      0.138085   0.032873   4.201 3.00e-05 ***
NbrStoneBrTRUE      0.133464   0.044276   3.014 0.002666 **
NbrBrkSideTRUE      0.060656   0.030053   2.018 0.043933 *
Fireplaces          0.027368   0.010962   2.496 0.012767 *
as.factor(Zone)FV   0.346089   0.071102   4.868 1.39e-06 ***
as.factor(Zone)RH   0.292339   0.079803   3.663 0.000267 ***
as.factor(Zone)RL   0.277301   0.064378   4.307 1.88e-05 ***
as.factor(Zone)RM   0.257281   0.064681   3.978 7.67e-05 ***
log(X1stFlrSF)      0.146607   0.024812   5.909 5.33e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1529 on 715 degrees of freedom
Multiple R-squared:  0.847,     Adjusted R-squared:  0.844
F-statistic: 282.7 on 14 and 715 DF,  p-value: < 2.2e-16
```

3 Garage Existence Effect

Now let's add factor GarageType:

3.1 R-code 2

```
regr <- lm(log(SalePrice) ~  as.factor(GarageType) +
log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
```

```
OverallCond +
NbrCrawfor +
NbrStoneBr +
NbrBrkSide +
Fireplaces +
as.factor(Zone) +
log(X1stFlrSF),train)

summary(regr)
```

3.2 Regression Result 2

```
Call:
lm(formula = log(SalePrice) ~ as.factor(GarageType) + log(YearBuilt) +
    log(LotArea) + log(LivAreaSF) + OverallQual + OverallCond +
    NbrCrawfor + NbrStoneBr + NbrBrkSide + Fireplaces + as.factor(Zone) +
    log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46038 -0.06506  0.01255  0.07594  0.42626

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -57.006422   4.665672 -12.218  < 2e-16 ***
as.factor(GarageType)Detchd  -0.015481   0.016921  -0.915 0.360552
as.factor(GarageType)NA      -0.090747   0.027111  -3.347 0.000859 ***
as.factor(GarageType)Other   -0.026085   0.021957  -1.188 0.235229
log(YearBuilt)                8.364224   0.609536  13.722  < 2e-16 ***
log(LotArea)                  0.101581   0.014186   7.161 2.01e-12 ***
log(LivAreaSF)                0.351221   0.026496  13.255  < 2e-16 ***
OverallQual                   0.092018   0.006711  13.711  < 2e-16 ***
OverallCond                   0.052067   0.005820   8.945  < 2e-16 ***
NbrCrawforTRUE                0.134818   0.032744   4.117 4.28e-05 ***
NbrStoneBrTRUE                0.129193   0.044038   2.934 0.003457 **
NbrBrkSideTRUE                0.066586   0.030082   2.213 0.027181 *
Fireplaces                    0.024108   0.010994   2.193 0.028645 *
as.factor(Zone)FV             0.336162   0.071496   4.702 3.10e-06 ***
as.factor(Zone)RH             0.292857   0.079426   3.687 0.000244 ***
as.factor(Zone)RL             0.270569   0.064351   4.205 2.95e-05 ***
as.factor(Zone)RM             0.251082   0.064640   3.884 0.000112 ***
log(X1stFlrSF)                0.137155   0.025632   5.351 1.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1519 on 712 degrees of freedom
Multiple R-squared:  0.8496,    Adjusted R-squared:  0.846
F-statistic: 236.6 on 17 and 712 DF,  p-value: < 2.2e-16
```

We clearly see that only the category GarageTypeNA (No Garage) is statistically significant. Keeping in mind that the reference category here is

Attchd (Attached to home) we conclude that the garage type itself is not statistically significant, it matters only whether there is a garage or not.

4 Garage Capacity Effect

Now let's try to add GarageCars:

4.1 Attempt 1 - Capacity and all types of a garage

Here let's first try to add capacity to all types of a garage

4.1.1 R-code 3

```
regr <- lm(log(SalePrice) ~  GarageCars +
as.factor(GarageType) +
log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
NbrBrkSide +
Fireplaces +
as.factor(Zone) +
log(X1stFlrSF),train)

summary(regr)
```

4.1.2 Regression Result 3

```
Call:
lm(formula = log(SalePrice) ~ GarageCars + as.factor(GarageType) +
    log(YearBuilt) + log(LotArea) + log(LivAreaSF) + OverallQual +
    OverallCond + NbrCrawfor + NbrStoneBr + NbrBrkSide + Fireplaces +
    as.factor(Zone) + log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33988 -0.06581  0.01126  0.07798  0.40450

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -4.735e+01  4.822e+00  -9.818  < 2e-16 ***
GarageCars                   7.372e-02  1.214e-02   6.071 2.07e-09 ***
as.factor(GarageType)Detchd -2.847e-02  1.665e-02  -1.710  0.08769 .
as.factor(GarageType)NA     -1.392e-04  3.037e-02  -0.005  0.99634
as.factor(GarageType)Other  -3.401e-02  2.146e-02  -1.585  0.11347
log(YearBuilt)               7.138e+00  6.281e-01  11.364  < 2e-16 ***
log(LotArea)                 9.056e-02  1.396e-02   6.487 1.64e-10 ***
log(LivAreaSF)               3.218e-01  2.630e-02  12.233  < 2e-16 ***
OverallQual                  8.533e-02  6.640e-03  12.850  < 2e-16 ***
OverallCond                  5.250e-02  5.680e-03   9.243  < 2e-16 ***
NbrCrawforTRUE               1.420e-01  3.197e-02   4.441 1.04e-05 ***
NbrStoneBrTRUE               1.295e-01  4.297e-02   3.014  0.00267 **
NbrBrkSideTRUE               6.822e-02  2.935e-02   2.324  0.02040 *
Fireplaces                   2.571e-02  1.073e-02   2.396  0.01685 *
as.factor(Zone)FV            3.613e-01  6.988e-02   5.170 3.05e-07 ***
as.factor(Zone)RH            3.163e-01  7.759e-02   4.077 5.09e-05 ***
as.factor(Zone)RL            2.995e-01  6.297e-02   4.757 2.38e-06 ***
as.factor(Zone)RM            2.680e-01  6.313e-02   4.245 2.48e-05 ***
log(X1stFlrSF)               1.145e-01  2.529e-02   4.528 6.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1482 on 711 degrees of freedom
Multiple R-squared:  0.857,     Adjusted R-squared:  0.8534
F-statistic: 236.7 on 18 and 711 DF,  p-value: < 2.2e-16
```

We see that all levels of GarageType become statistically insignificant, while GarageCars is significant. So it is impossible to distinguish among different types (attached, detached, etc.), and the capacity of a garage (number of cars).

4.2 Attempt 2 - Introduce a dummy for the garage existence

Here we include only existence of a garage instead of all garage types as a factor. Let's try to create a dummy variable GarageTypeYN which is assigned 1 if there is a garage and 0 if there is no garage.

4.2.1 R-code 4

```
train["GarageTypeYN"] <- train$GarageType != "NA"

regr <- lm(log(SalePrice) ~  GarageCars +
as.factor(GarageTypeYN) +
log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
NbrBrkSide +
Fireplaces +
as.factor(Zone) +
log(X1stFlrSF),train)

summary(regr)
```

4.2.2 Regression Result 4

```
Call:
lm(formula = log(SalePrice) ~ GarageCars + as.factor(GarageTypeYN) +
    log(YearBuilt) + log(LotArea) + log(LivAreaSF) + OverallQual +
    OverallCond + NbrCrawfor + NbrStoneBr + NbrBrkSide + Fireplaces +
    as.factor(Zone) + log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35797 -0.06483  0.01272  0.07750  0.40938

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -49.776893   4.367687 -11.397  < 2e-16 ***
GarageCars                  0.070463   0.012056   5.845 7.72e-09 ***
as.factor(GarageTypeYN)TRUE -0.014934   0.029197  -0.512  0.60915
log(YearBuilt)              7.447949   0.569779  13.072  < 2e-16 ***
log(LotArea)                0.091807   0.013894   6.608 7.64e-11 ***
log(LivAreaSF)              0.315204   0.025670  12.279  < 2e-16 ***
OverallQual                 0.086444   0.006627  13.045  < 2e-16 ***
OverallCond                 0.053592   0.005667   9.457  < 2e-16 ***
NbrCrawforTRUE              0.138427   0.031983   4.328 1.72e-05 ***
NbrStoneBrTRUE              0.133449   0.043007   3.103  0.00199 **
NbrBrkSideTRUE              0.062448   0.029202   2.139  0.03281 *
Fireplaces                  0.028276   0.010669   2.650  0.00822 **
as.factor(Zone)FV           0.365773   0.069178   5.287 1.65e-07 ***
as.factor(Zone)RH           0.321723   0.077633   4.144 3.82e-05 ***
as.factor(Zone)RL           0.304742   0.062713   4.859 1.45e-06 ***
as.factor(Zone)RM           0.271269   0.062881   4.314 1.83e-05 ***
log(X1stFlrSF)              0.128986   0.024288   5.311 1.46e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1485 on 713 degrees of freedom
Multiple R-squared:  0.8561,    Adjusted R-squared:  0.8528
F-statistic: 265.1 on 16 and 713 DF,  p-value: < 2.2e-16
```

We see that if we include both factors, GarageTypeYN is insignificant. Let's evaluate whether GarageTypeYN can be statistically significant if we use only this factor, and if so, evaluate which factor - GarageType or GarageCars - provides better model.

4.3 Attempt 3 - Compare models: Existence VS Capacity

Let's analyze regressions where only one factor is included: Existence or Capacity.

4.3.1 R-code 5

```
regr <- lm(log(SalePrice) ~  as.factor(GarageTypeYN) +
```

```
log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
NbrBrkSide +
Fireplaces +
as.factor(Zone) +
log(X1stFlrSF),train)

summary(regr)
```

4.3.2 Regression Result 5

```
Call:
lm(formula = log(SalePrice) ~ as.factor(GarageTypeYN) + log(YearBuilt) +
    log(LotArea) + log(LivAreaSF) + OverallQual + OverallCond +
    NbrCrawfor + NbrStoneBr + NbrBrkSide + Fireplaces + as.factor(Zone) +
    log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46642 -0.06163  0.01231  0.07468  0.42630

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -57.945076   4.233074 -13.689  < 2e-16 ***
as.factor(GarageTypeYN)TRUE  0.079818   0.024840   3.213 0.001371 **
log(YearBuilt)               8.471861   0.554627  15.275  < 2e-16 ***
log(LotArea)                 0.101687   0.014107   7.208 1.45e-12 ***
log(LivAreaSF)               0.344709   0.025747  13.388  < 2e-16 ***
OverallQual                  0.092643   0.006692  13.845  < 2e-16 ***
OverallCond                  0.052805   0.005795   9.112  < 2e-16 ***
NbrCrawforTRUE               0.132802   0.032702   4.061 5.43e-05 ***
NbrStoneBrTRUE               0.131764   0.043994   2.995 0.002839 **
NbrBrkSideTRUE               0.063459   0.029871   2.124 0.033980 *
Fireplaces                   0.025681   0.010904   2.355 0.018787 *
as.factor(Zone)FV            0.342710   0.070651   4.851 1.51e-06 ***
as.factor(Zone)RH            0.297878   0.079306   3.756 0.000187 ***
as.factor(Zone)RL            0.276535   0.063963   4.323 1.75e-05 ***
as.factor(Zone)RM            0.255528   0.064266   3.976 7.72e-05 ***
log(X1stFlrSF)               0.146709   0.024651   5.951 4.17e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1519 on 714 degrees of freedom
Multiple R-squared:  0.8492,    Adjusted R-squared:  0.846
F-statistic:   268 on 15 and 714 DF,  p-value: < 2.2e-16
```

4.3.3 R-code 6

```
regr <- lm(log(SalePrice) ~  GarageCars +
log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
NbrBrkSide +
Fireplaces +
as.factor(Zone) +
log(X1stFlrSF),train)

summary(regr)
```

4.3.4 Regression Result 6

```
Call:
lm(formula = log(SalePrice) ~ GarageCars + log(YearBuilt) + log(LotArea) +
    log(LivAreaSF) + OverallQual + OverallCond + NbrCrawfor +
    NbrStoneBr + NbrBrkSide + Fireplaces + as.factor(Zone) +
    log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36071 -0.06617  0.01244  0.07713  0.40870

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -50.010498   4.341499 -11.519  < 2e-16 ***
GarageCars          0.067039   0.010021   6.690 4.52e-11 ***
log(YearBuilt)      7.476181   0.566806  13.190  < 2e-16 ***
log(LotArea)        0.091790   0.013887   6.610 7.52e-11 ***
log(LivAreaSF)      0.316072   0.025601  12.346  < 2e-16 ***
OverallQual         0.086613   0.006615  13.093  < 2e-16 ***
OverallCond         0.053527   0.005662   9.453  < 2e-16 ***
NbrCrawforTRUE      0.137470   0.031912   4.308 1.88e-05 ***
NbrStoneBrTRUE      0.133147   0.042981   3.098  0.00203 **
NbrBrkSideTRUE      0.062860   0.029175   2.155  0.03153 *
Fireplaces          0.027931   0.010642   2.625  0.00886 **
as.factor(Zone)FV   0.364215   0.069075   5.273 1.78e-07 ***
as.factor(Zone)RH   0.321281   0.077588   4.141 3.87e-05 ***
as.factor(Zone)RL   0.303272   0.062615   4.843 1.57e-06 ***
as.factor(Zone)RM   0.270277   0.062819   4.302 1.92e-05 ***
log(X1stFlrSF)      0.129860   0.024215   5.363 1.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1484 on 714 degrees of freedom
Multiple R-squared:  0.856,     Adjusted R-squared:  0.853
F-statistic:   283 on 15 and 714 DF,  p-value: < 2.2e-16
```

It is clear that in both cases a garage factor becomes statistically significant. However, model with GarageCars has larger R^2. Therefore, we can conclude, that if we know the size of garage in car capacity, it makes more sense to use it in the model rather than just the information whether there is a garage or not. Or, simply speaking, the car capacity really matters. In a nutshell, our results also fit some common sense that the utility of a garage is really determined by how many cars can be stored there, and it is way more important than whether the garage is attached or detached to home, etc.
Taking into account our model specification, the answer to the question, how much we would spend on building a garage before selling a house, at first sight would be:

$$actual\_price * (e^{0.067039 * planned\_car\_capacity} - 1)$$

It is clear that it depends on which car capacity we would like to make.

5 Investment Safety

However, in order to make our investments safer, we have to take into account that GarageCars coefficient is itself a random variable, so we need to know its 95% confidence interval:

5.1 R-code 7

```
confint(regr, 'GarageCars', level=0.95)

qt(0.975, 714)*summary(regr)$coefficients[2,2]/summary(regr)$coefficients[2,1]*
```

5.2 Interpretation of the Confidence Interval

```
> confint(regr, 'GarageCars', level=0.95)
                2.5 %     97.5 %
GarageCars 0.04736421 0.08671417
>
>
>
> qt(0.975, 714)*summary(regr)$coefficients[2,2]/summary(regr)$coefficients[2,1]*100
[1] 29.34848
```

Note that the coefficient can fluctuate considerably - by approximately +-29% from its estimated value:

So, it would be more reasonable to calculate maximum spending using the lower bound of the 95% confidence level of the coefficient:

$$actual\_price * (e^{0.04736421 * planned\_car\_capacity} - 1)$$

Interpretation: taking 95% confidence interval, if garage capacity increases by 1 car, the expected increase of the house price is at least approximately 4.736421%.