# Philadelphia houses

Francesca Ghidini (Matricola 5007699)

## 1. Dataset

The analyzed dataset provides information regarding houses in Philadelphia sold by the online real estate marketplace company Zillow in the period between August 2016 and March 2017.

The dataset contains 565 observations and 17 variables that are:

- **Address**: Address of the house
- **Sale_Price**: Dollar money paid to buy the house
- **Postal_Code**: Postal code of the houses in Philadelphia from 19104 to 19153
- **Advertising**: Dollar Money spent to advertise the house
- **AvgWalkTransitscore**: Average between the Walk Score and the Transit Score provided by Zillow. Walk Score measures how walkable an address is based on the distance to nearby amenities and is calculated by Zillow. It's a value within the range 0-100. The Transit Score measures how well a location is served by public transportation and it is a value within the range 0-100
- **ViolentCrimeRate**: It calculates the crime risk near an address and it is provided by Zillow on a scale 0-2
- **SchoolScore**: School rating provided by Zillow on a scale from 0 to 100
- **Zillow_Estimate**: Estimation of a home's market value made by Zillow taking into account public and user-submitted data
- **Rent_Estimate**: Zillow's estimated monthly rent price
- **TaxAssessment**: It is a dollar value attached to your real property and business personal property by the local government, specifically for the purpose of levying and collecting tax money that is used to support your community. (From https://www.crowdreason.com/blog/tax-assessment-vs-property-tax)
- **YearBuilt**
- **Sqft**: Livable area in sqft
- **Bathrooms**: Number of bathrooms, where 0.5 indicates a bathroom without bathtub or shower.
- **Bedrooms**
- **PropType**: Property types that can be: Condominium, Single Family (up to 2 people), Multi Family or Townhouse
- **Average_comps**: A real estate estimation term (in dollar) referring to properties with similar characteristics
- **River**: A qualitative variable with two level describing if a house is near or far from Delaware River

To make the analysis more interpretable the quantitative variables of the dataset have been centered as follow:

| Variables | Center |
|---|---|
| `Advertising` | 1,739.24 |
| `AvgWalkTransitscore` | 71.16 |
| `ViolentCrimeRate` | 0.68 |
| `SchoolScore` | 13.57 |
| `Zillow_Estimate` | 118,402.1 |

| Variables | Center |
|---|---|
| Rent_Estimate | 1,221.73 |
| TaxAssessment | 103,573.1 |
| YearBuilt | 1934 |
| Sqft | 1,290.84 |
| Bathrooms | 1 |
| Bedrooms | 3 |
| Average_comps | 113,112.2 |

Therefore, the subsequent analysis will consider "regular houses" defined as the houses with the features listed above.

## 2. Goal

The dataset provides reliable information that are collected from the website Zillow. The only strange data is the one concerning the `Sale_Price`. In fact, the prices of the houses seems to be particularly low: they vary over an interval of $\$5,200$ and $\$350,000$. Anyway, also checking on the website the displayed prices at the period of the sale are exactly the ones of the dataset. On the whole, the dataset can be considered reliable because the data are stable and precise considering the city of Philadelphia and the period between August 2016 and March 2017. Also, the sample selection is appropriate because the Zillow website sells houses of Philadelphia of any type, so it can be considered a representative sample. However, it is essential to take into account the fact that the analysis regards houses already sold in Philadelphia and in a precise period, therefore the conclusion may be not accurate for houses sold in other cities and in other periods.

The goal of this analysis is to explore associations between explanatory variables and the response variable `Sale_Price`. The data are suitable to address this aim because they offered different kind of information regarding the houses that may not be directly associated to an increase in the price of a house. For instance, it would be interesting to study how the `SchoolScore` of an area or the `ViolentCrimeRate` impact on the price of a house.

## 3. Relation among all variables

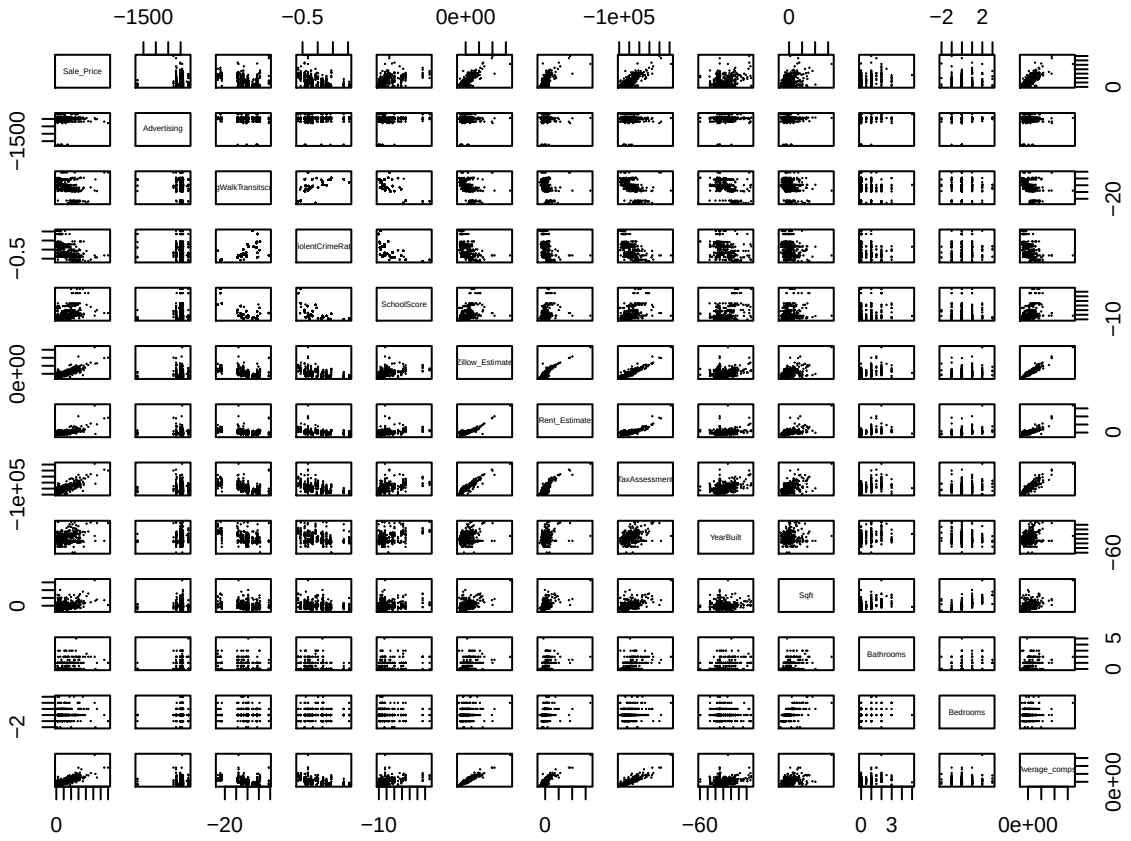



Figure 1: Scatterplot quantitative variables

From Figure 1 it is possible to notice that the response variable (`Sale_Price`) has a clear positive relationship with the variables `Zillow_Estimate`, `Rent_Estimate`, `TaxAssessment` and `Average_comps`. These variables seem to have also a strong positive relationship among them considered in pairs. This was something foreseeable due to the fact that all these values are estimates made by Zillow regarding the value of the house depending on their features. Also, the `Sqft` variable has a positive relationship with `Average_comps`. Regarding all the others variables there aren't any particular or clear relationships.

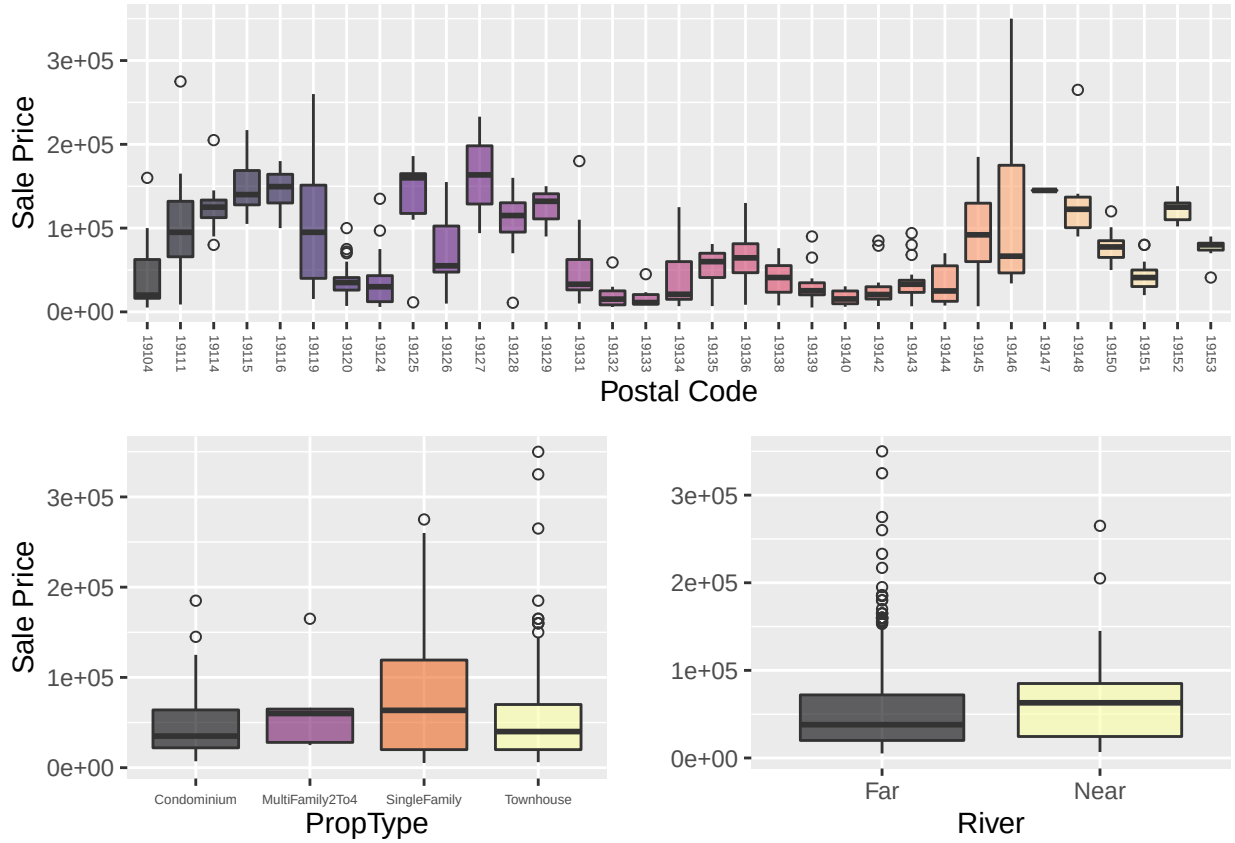



Figure 2: Boxplot qualitative variables vs Sale_Price

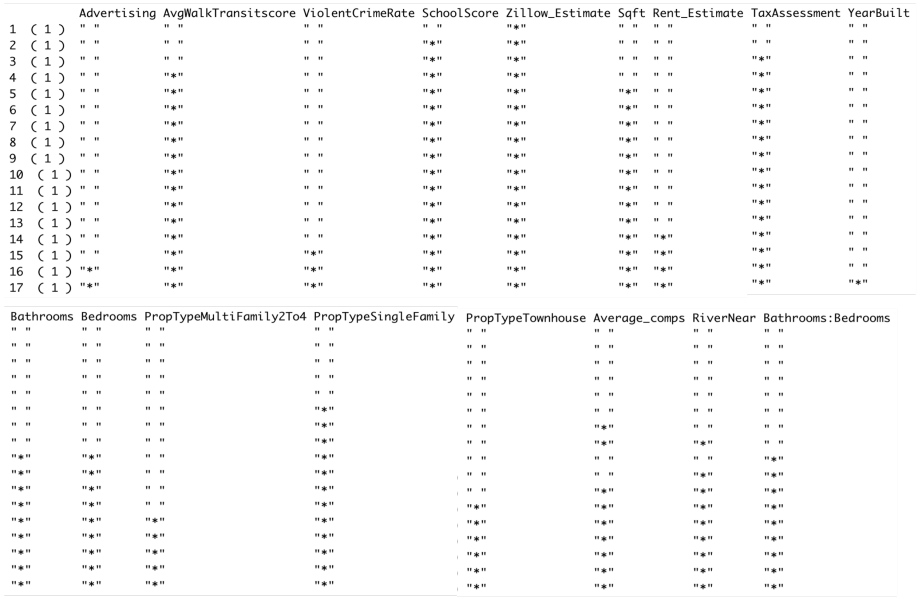
In Figure 2 are represented the box plots of the categorical variables in relation to the response variable (`Sale_Price`). The location of a house defined by its `Postal_Code` has a huge impact on the price as it is possible to notice from the graph where the box plots are very different from each other. Considering the box plot of the `PropType` variable it is shown that the categories Condominium“,”MultiFamily2To4" and “Townhouse” have quite similar box plots while the only category that seems to have huge difference in the prices of the houses is “SingleFamily”. To conclude, the proximity to the Delaware River doesn't seem to be a discriminating factor in defining the price of a house.

## 4. Best Subset Selection

To compute the Best Subset selection are considered all the variables except of `Postal_Code` due to the number of its levels. It is also included an interaction term given by the product between the variables `Bathrooms` and `Bedrooms`. This interaction term has been created due to the fact that there is relationship among these variables that should grow in a similar way. Therefore, the p parameters are 17 and n is 565. In Figure 3 are shown the result of the `regsubsets` function showing the variables included in the best linear model for each number of parameters.

```
ols_ss <-
  regsubsets(
    Sale_Price ~ Advertising + AvgWalkTransitscore + ViolentCrimeRate + SchoolScore +
      Zillow_Estimate + Sqft + Rent_Estimate + TaxAssessment + YearBuilt + Bathrooms * Bedrooms +
      PropType + Average_comps + River, house.prices, nvmax = p)
summ <- summary(ols_ss)
```

In Figure 3 it is shows that the best model with only one variable is the one with `Zillow_Estimate` and this result was quite foreseeable because as showed in Figure 1 the two variables are really correlated. Another interesting result is the



| | Advertising | AvgWalkTransitscore | ViolentCrimeRate | SchoolScore | Zillow_Estimate | Sqft | Rent_Estimate | TaxAssessment | YearBuilt |
|---|---|---|---|---|---|---|---|---|---|
| 1 ( 1 ) | " " | " " | " " | " " | "*" | " " | " " | " " | " " |
| 2 ( 1 ) | " " | " " | " " | "*" | "*" | " " | " " | " " | " " |
| 3 ( 1 ) | " " | " " | " " | "*" | "*" | " " | " " | "*" | " " |
| 4 ( 1 ) | " " | "*" | " " | "*" | "*" | " " | " " | "*" | " " |
| 5 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 6 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 7 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 8 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 9 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 10 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 11 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 12 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 13 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | " " | "*" | " " |
| 14 ( 1 ) | " " | "*" | " " | "*" | "*" | "*" | "*" | "*" | " " |
| 15 ( 1 ) | " " | "*" | "*" | "*" | "*" | "*" | "*" | "*" | " " |
| 16 ( 1 ) | "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" | " " |
| 17 ( 1 ) | "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" |

| Bathrooms | Bedrooms | PropTypeMultiFamily2To4 | PropTypeSingleFamily | PropTypeTownhouse | Average_comps | RiverNear | Bathrooms:Bedrooms |
|---|---|---|---|---|---|---|---|
| " " | " " | " " | " " | " " | " " | " " | " " |
| " " | " " | " " | " " | " " | " " | " " | " " |
| " " | " " | " " | " " | " " | " " | " " | " " |
| " " | " " | " " | " " | " " | " " | " " | " " |
| " " | " " | " " | "*" | " " | " " | " " | " " |
| " " | " " | " " | "*" | " " | "*" | " " | " " |
| " " | " " | " " | "*" | " " | "*" | "*" | " " |
| "*" | "*" | " " | "*" | " " | " " | " " | "*" |
| "*" | "*" | " " | "*" | " " | " " | "*" | "*" |
| "*" | "*" | " " | "*" | " " | "*" | "*" | "*" |
| "*" | "*" | " " | "*" | "*" | "*" | "*" | "*" |
| "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" |
| "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" |
| "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" |
| "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" |
| "*" | "*" | "*" | "*" | "*" | "*" | "*" | "*" |

Figure 3: Best Subset Selection

fact that the level `SingleFamily` of the predictor `PropType` is included from the model with 6 predictors, while the others levels are included only in the latest models. An unexpected result is the fact that `YearBuilt` predictor is the last added to the model even if the fact that a house is new or old should be an important factor to define the price of a house.

## 5. Best Overall Model

To identify the best overall model are employed the calculation of $BIC$, $Adjusted\ R^2$, $C_p\ Mallows$ and $Cross-validation\ error$ for all the models individuated in the best subset selection. $Cross-validation\ error$ is computed with the k-fold approach considering 10 folds, and it is computed 3 times considering different samples to see if the results are consistent.

As it is possible to notice in Figure 4 the result of the 6 plots are quite different:

- The Bayesian Information Criterion identify as best model the one that includes 6 predictors that are `AvgWalkTransitscore`, `SchoolScore`, `Zillow_Estimate`, `Sqft`, `TaxAssessment` and `PropType=SingleFamily`. This model has BIC = -704.5226 that is the lowest among all the others.
- $Adjusted\ R^2$ of the best models are shown in the second plot in Figure 4. The model with the highest $Adjusted\ R^2$ is the one with 11 predictors with $Adjusted\ R^2 = 0.7353561$
- $C_p\ Mallows$ of the best models defined in Figure 3 are the ones plotted in the third graph. The best model according to $C_p\ Mallows$ is the one with the lowest value that is the one with 11 predictors ($C_p\ Mallows = 8.468142$)
- $Cross-validation\ error$ is computed for three different samples with `set.seed = 1, 2, 3`. The result is that in the first and in the third plot (second row) the best model is the one with 6 predictors, while in the second graph the model with the lowest $Cross-validation\ error$ is the one with 7 predictors.

Therefore, the models that seem to be the best are the one with 6 predictors and the one with 11 predictors. For Occam's razor principle the simplest explanation is usually the right one therefore the model with 6 predictors should be used. To support this choice it is possible to see from the graph that even if the $C_p\ Mallows$ and the $Adjusted\ R^2$ methods suggested the model with 11 predictors their values don't get worse excessively if the model with 6 predictors is used. In fact, the value of $Adjusted\ R^2$ is just less 0.0038957 with respect to the one with 11 predictors, while the value of $C_p\ Mallows$ increases by 3.129738 in the reduced model.

```
summ <- summary(ols_ss)
k <- 10
for (z in 1:3) {
  set.seed(z)
```

```
folds <- sample(1:k, n, replace = TRUE)
cv.matrix <- matrix(NA, k, p, dimnames = list(NULL, paste(1:p)))
for (j in 1:k) {
  best.fit <- regsubsets(
      Sale_Price~Advertising+AvgWalkTransitscore+ViolentCrimeRate+SchoolScore+Zillow_Estimate+Sqft
      +Rent_Estimate +TaxAssessment+YearBuilt+Bathrooms*Bedrooms+PropType+Average_comps+River,
      house.prices[folds != j,], nvmax = p)
  for (i in 1:p) {
    mat <-
      model.matrix(as.formula(best.fit$call[[2]]), house.prices[folds == j,])
    coefi <-  coef(best.fit, id = i)
    xvars <- names(coefi)
    pred <- mat[, xvars] %*% coefi
    cv.matrix[j, i] <-
      mean((pred - house.prices$Sale_Price[folds == j]) ^ 2)}}
m <- colMeans(cv.matrix)
plot(1:p, m, type = "b", xlab = "Number of predictors", ylab = "MSE", main = "Cv error")
lines(x = c(which.min(m), which.min(m)), y = c(min(m),-2000), col = "red", lty = 3, lwd = 2)}
```

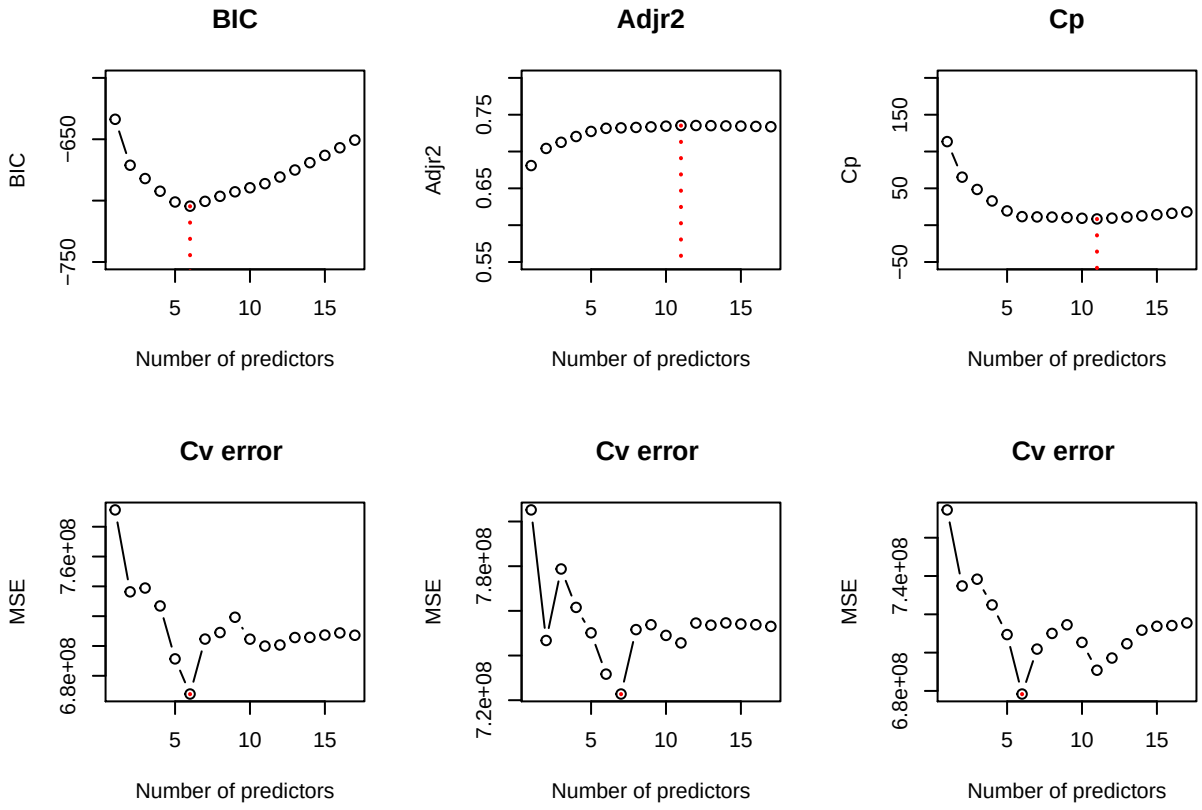


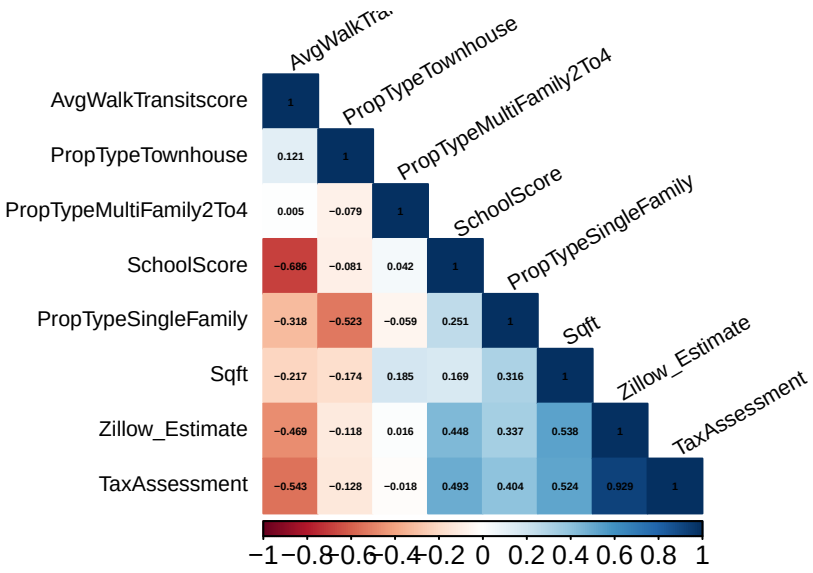

Figure 4: Best model choice

```
ols <- lm(Sale_Price~AvgWalkTransitscore+SchoolScore+Zillow_Estimate+Sqft
        +TaxAssessment+PropType, house.prices)
summar <- summary(ols)
```

## 6. Collinearity Issues

In Figure 1 were shown the relationship among predictors and looking at the plot it is likely to assume that there could be some collinearity between the predictors `Zillow_Estimate` and `TaxAssessment`. Considering the Figure 4 it is possible to notice that there is a correlation of 0.92948290 among these predictors that is a really high correlation. Anyway, considering the Variance Inflation Factor displayed below it is always lower than 10 therefore it is possible to assume that

there are not multicollinearity issues among the predictors. Therefore, even if `Zillow_Estimate` and `TaxAssessment` are strongly correlated they explain different part of the response variable, therefore they will be still included in the model.




| Predictor | Vif |
|---|---|
| AvgWalkTransitscore | 2.166633 |
| SchoolScore | 2.001726 |
| Zillow_Estimate | 7.809048 |
| TaxAssessment | 8.855483 |
| Sqft | 1.567030 |
| PropTypeMultiFamily2To4 | 1.091138 |
| PropTypeSingleFamily | 1.736648 |
| PropTypeTownhouse | 1.422222 |

## 7. Diagnostics

### 7a. Constant variance

One of the main assumption to build a linear model is constant variance that can be detected from the residual plot:

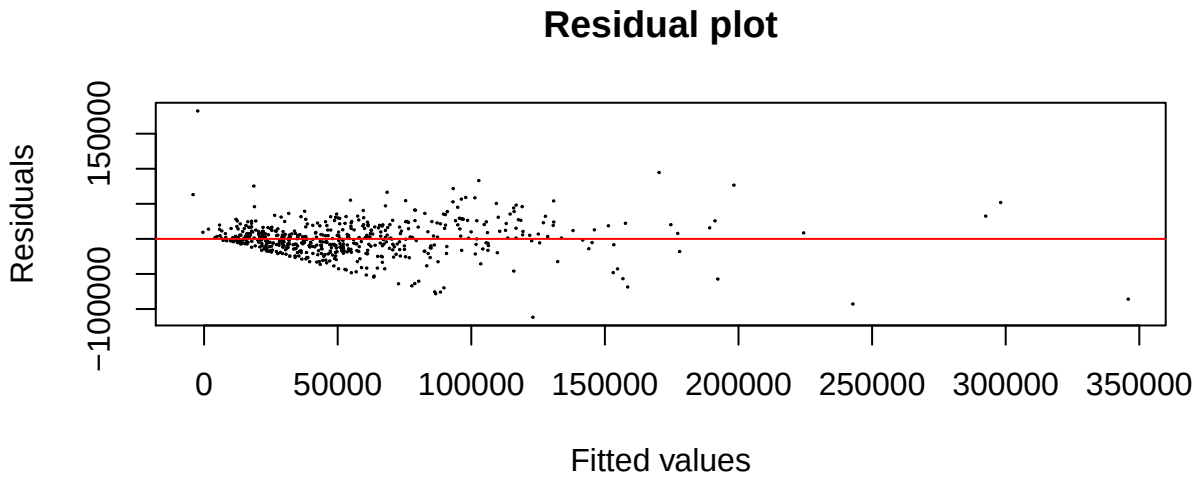



Figure 5: Residual plot

Figure 5, that shows the residuals vs the fitted values of the model, shows that the residuals are distributed as a right-opening microphone, in fact they are more concentrated and close to the horizontal line for low values and increase in

absolute value as the fitted values increase. Therefore, it is possible to conclude that the variance of the residuals is not constant, so there is heteroscedasticity that is in contrast with the assumption of linear models.

### 7b. Relationship between the predictors and the response.

Another important assumption is the linearity of the residuals that can be examined in the residual plot in Figure 5. From the graph there is no evidence of a particular pattern among the residuals, therefore it is possible to assume the linearity of the residuals.

### 7c. Normality assumption

Normality assumption can be tested with QQ-plot (Figure 6) and the Shapiro-Wilk Normality Test that test the following hypothesis:

$$\begin{cases} H_0 : \text{errors are normally distributed} \\ H_1 : \text{errors are NOT normally distributed} \end{cases}$$

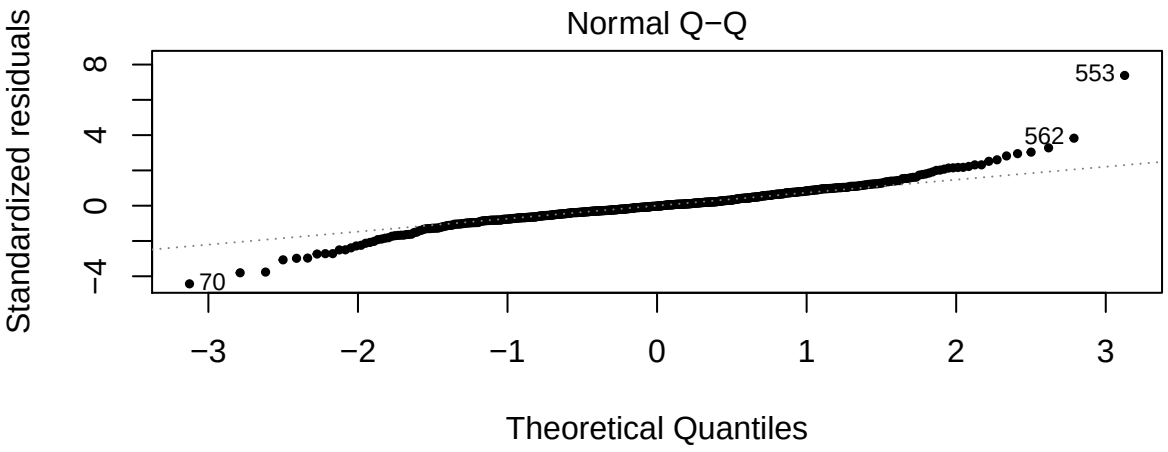



Figure 6: QQ plot

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ols)
## W = 0.93714, p-value = 1.07e-14
```

From the output of the `shapiro.test` we get a p-value equal to 1.07e-14, therefore there are evidence against the null hypothesis therefore we can't assume the residuals to be normal. This result is also confirmed from the Figure 6 that shows a long tails distribution of the quantiles. Especially, the observation 553 seems to be particularly far from the theoretical quantiles.

### 7d. Large leverage points.

High leverage points are values that have unusual values of the predictors. These values can change completely inferential conclusions. To identify high leverage points their leverage statistics are calculated:

$$h_{ii} = \mathbf{x_i}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_i}$$

```
infl <- influence(ols)
hat <- infl$hat
levpoint <- sum(hat > 2*(p+1)/n)
```

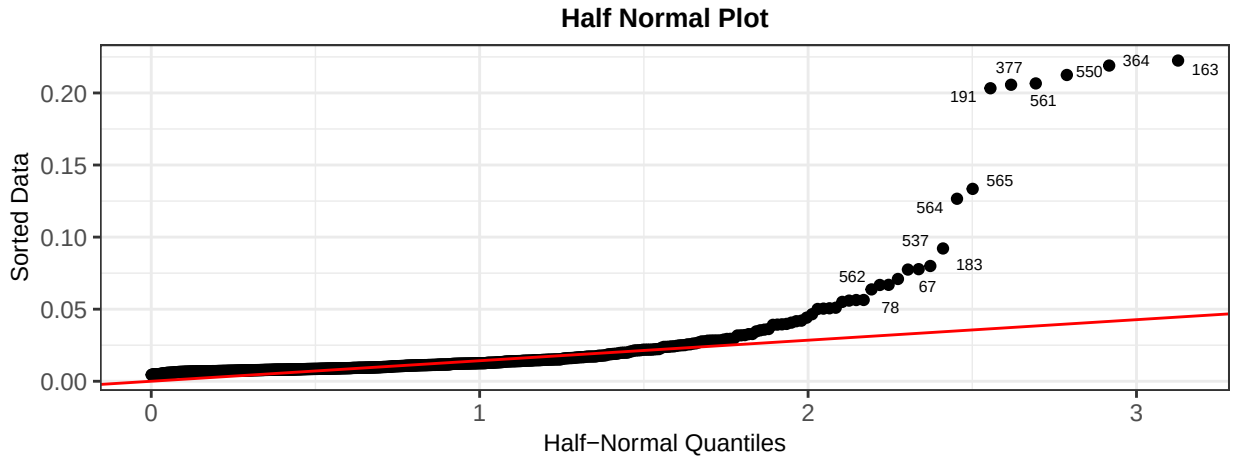



Figure 7: Leverages points

Figure 7 shows the leverage of the observations. Considering the rule of thumb, a point can be considered a high leverage point if its leverage is higher than: $h_{ii} > \frac{2*(p+1)}{n}$ that in this case is 0.06371681. Therefore, there are 15 observations (indicated in the plot) that have a high leverage statistics. Therefore, these points can be considered as potentially problematic and need to be further checked.

### 7e. Outliers

An outlier is a point that does not follow the same model as the rest of the data. Potential outliers can be found analysing the values of the standardized residuals $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$. If the absolute value of a standardized residual is higher than 3, the corresponding observation might be an outlier.

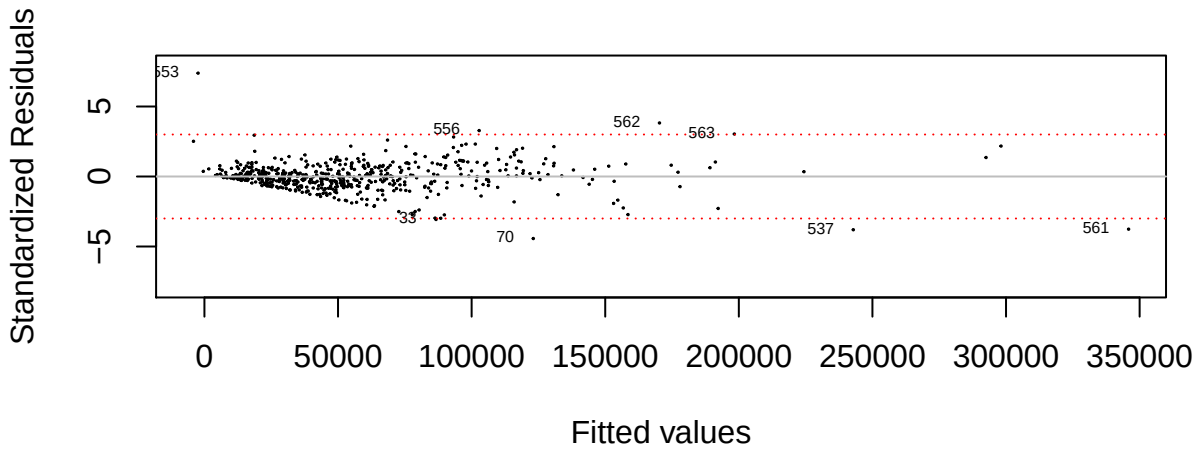



Figure 8: Standardized Residuals plot to identify outliers

### 7f. Influential points.

Influential points are points that have a huge impact on the model, therefore excluding them will change the model a lot. These points can be detected calculating the Cook's distance that combines standardized residuals and leverages.

Figures 9 provides information regarding the Cook's distance of the observations. Figure 9 on the left shows that the observations 553, 561 and 537 have a high Cook's distance compared with the others observations. These points are also the ones that in the Figure 9 on the right are nearer to the dotted lines. Considering the rule of thumb $D_{ii} < 1$ none of the highlighted point can be considered as an influential point.

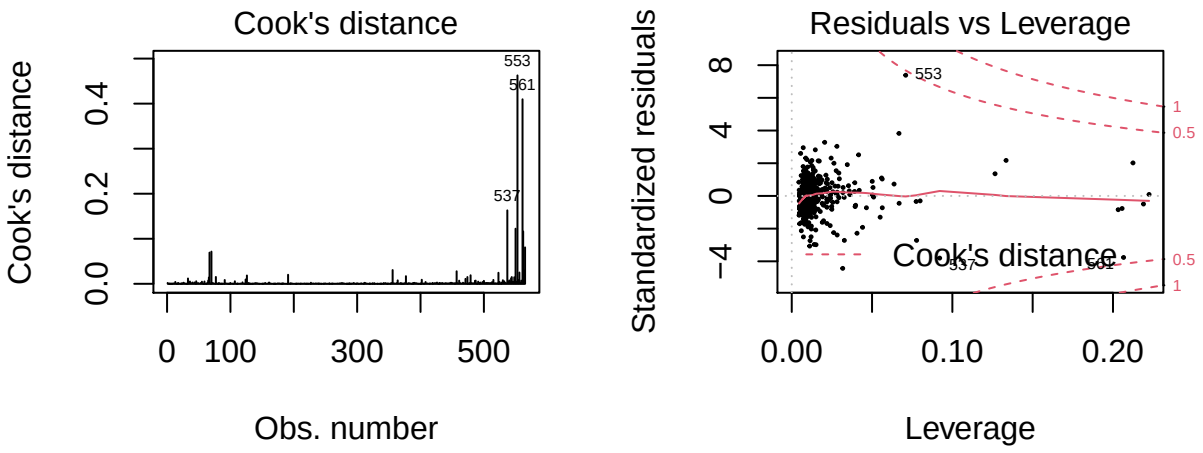



Figure 9: Influencial Points

## 8. Improvement of the model.

The issues of the model seems to be the normality assumption and the heteroscedasticity. Therefore, to improve the model 3 new fitting have been created:
1) OLS fitting with the same predictors as before but removing the most influential points even if they didn't exceed the rule of thumb. These points are the observations 553, 561 and 537 that were also identified as potential outliers and potential high leverage points.
2) OLS fitting with the same predictors and n observation but with a transformation of the response variable that it is considered as $\sqrt{Sale\_Price}$.
3) Another option that is analyzed below is OLS fitting with the same predictors and n observation but with a transformation of the response variable that it is considered as $\ln(Sale\_Price)$. Also others transformations of the response variable have been performed but the best ones seem to be the one with the square root of `Sale_price` and the log.

Shown below there are the fitting of the four different models: the first one is the model analyzed without any modification, the second one is the model excluding the most influential points, the third and the fourth model are obtained through transformation of the response variable trying to solve the issue of non-constant variance. Comparing the four models it is possible to notice that the second one is the best model, in fact it has a higher *adjusted* $R^2$ compared to the others models and has also a lower residual standard error of the model with the square root. On the other hand, using the models fitted with $\sqrt{Sale\_Price}$ and $ln(Sale\_Price)$ the variable `AvgWalkTransitscore` becomes not significant. Moreover, the model with $ln(Sale\_Price)$ has the lowest residual standard error but also the lowest $R^2$.

```
ols_out <- lm(Sale_Price~AvgWalkTransitscore+SchoolScore+Zillow_Estimate+Sqft
        +TaxAssessment+PropType,
             data = house.prices, subset = cooks.distance(ols)<0.16)
ols_sqrt <-  lm(sqrt(Sale_Price) ~ AvgWalkTransitscore+SchoolScore+Zillow_Estimate+Sqft
        +TaxAssessment+PropType,
             data = house.prices)
ols_log <-  lm(log(Sale_Price) ~ AvgWalkTransitscore+SchoolScore+Zillow_Estimate+Sqft
        +TaxAssessment+PropType,
             data = house.prices)
stargazer(ols,ols_out, ols_sqrt, ols_log, report="vcp*", type = "text", summary.stat= "sd",
          notes.append = FALSE, align=TRUE, no.space=TRUE, omit.stat=c("LL","f"), digits = 3)
```

```
##
## ================================================================================================
##                                                Dependent variable:
##                          ------------------------------------------------------------------------
##                                        Sale_Price                    sqrt(Sale_Price)  log(Sale_Price)
##                                (1)                    (2)                   (3)              (4)
## ------------------------------------------------------------------------------------------------
## AvgWalkTransitscore          718.270                801.549               0.464            -0.001
##                          p = 0.0002***          p = 0.00001***        p = 0.249        p = 0.780
## SchoolScore                  919.572                819.622               1.543            0.011
```

```
##                                p = 0.000***          p = 0.000***       p = 0.00000***      p = 0.001***
## Zillow_Estimate                   0.313                 0.391               0.001              0.00000
##                                p = 0.000***          p = 0.000***       p = 0.000***      p = 0.00005***
## Sqft                             -11.917               -15.806             -0.024             -0.0002
##                                p = 0.0005***        p = 0.00001***      p = 0.001***       p = 0.015**
## TaxAssessment                     0.324                 0.301               0.001              0.00001
##                                p = 0.000***          p = 0.000***       p = 0.00000***     p = 0.00001***
## PropTypeMultiFamily2To4        13,314.430            15,017.780            30.263             0.299
##                                 p = 0.269             p = 0.179          p = 0.246          p = 0.303
## PropTypeSingleFamily           -8,324.719            -9,396.561            -23.374            -0.322
##                                p = 0.009***          p = 0.002***       p = 0.001***      p = 0.00003***
## PropTypeTownhouse               -374.956              -791.921             -4.959             -0.096
##                                 p = 0.887             p = 0.744          p = 0.382          p = 0.131
## Constant                       59,367.270            59,896.460            227.392            10.713
##                                p = 0.000***          p = 0.000***       p = 0.000***       p = 0.000***
## ------------------------------------------------------------------------------------------------------
## Observations                        565                   562                 565                565
## R2                                0.735                 0.764               0.664              0.531
## Adjusted R2                       0.731                 0.761               0.659              0.524
## Residual Std. Error     25,622.750 (df = 556) 23,652.490 (df = 553) 55.437 (df = 556) 0.618 (df = 556)
## ======================================================================================================
## Note:                                                                  *p<0.1; **p<0.05; ***p<0.01
```

Considering the assumptions, the model without the most influential points has a short-tailed distribution of the quantiles that is better than the long-tailed distribution, but, on the other hand doesn't solve the problem of non-constant variance, while the linearity assumption is still valid. The models with $\sqrt{Sale\_Price}$ and $\ln(Sale\_Price)$ have a variance that seems more constant, but always similar to a right opening megaphone shape. However these models loose the linearity (Figure 10). Therefore, the model that will be considered is the one without the most influential point and with the response variable not transformed.

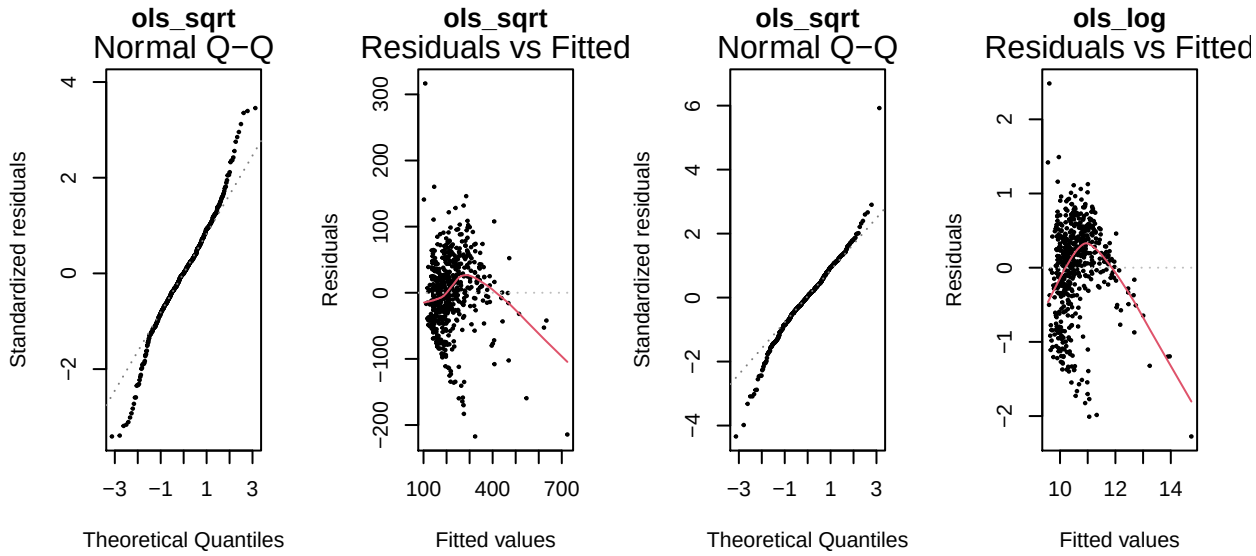



Figure 10: First plot from the left: QQplot model without influencial points. Second plot: Residual plot of the model with sqrt(SalePrice). Third plot: QQplot of the model with sqrt(SalePrice). Fourth plot: Residual plot of the model with ln(SalePrice)

As explained before, the assumption of constant variance will be not valid, this fact has some consequences. Least squares estimators are still unbiased, but, due to the fact that the Gauss-Markov Theorem depends on the assumption that variance is constant, the estimator will not be BLUE (best linear unbiased estimator) and therefore there would be another estimator with a lower variance. For that reason all the following analysis will be quite unreliable and the OLS estimator will not be efficient. Another consequence is that tests and confidence intervals are not reliable.

## 9. Parameters and Uncertainties

```
##
## Call:
## lm(formula = Sale_Price ~ AvgWalkTransitscore + SchoolScore +
##     Zillow_Estimate + Sqft + TaxAssessment + PropType, data = house.prices,
##     subset = cooks.distance(ols) < 0.16)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120381  -12694     237   12813   79631
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.990e+04  1.884e+03  31.787  < 2e-16 ***
## AvgWalkTransitscore       8.015e+02  1.717e+02   4.669 3.80e-06 ***
## SchoolScore               8.196e+02  1.279e+02   6.410 3.13e-10 ***
## Zillow_Estimate           3.911e-01  3.777e-02  10.354  < 2e-16 ***
## Sqft                     -1.581e+01  3.289e+00  -4.806 1.99e-06 ***
## TaxAssessment             3.007e-01  4.951e-02   6.073 2.34e-09 ***
## PropTypeMultiFamily2To4   1.502e+04  1.114e+04   1.348  0.17823
## PropTypeSingleFamily     -9.397e+03  2.917e+03  -3.221  0.00135 **
## PropTypeTownhouse        -7.919e+02  2.414e+03  -0.328  0.74302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23650 on 553 degrees of freedom
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.7608
## F-statistic:   224 on 8 and 553 DF,  p-value: < 2.2e-16
```

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5} + \hat{\beta}_6 x_{i6} + \hat{\beta}_7 x_{i7} + \hat{\beta}_8 x_{i8} + \epsilon_i$$

- $\hat{\beta}_0$ is $\$59,896.46$, therefore on average the estimated sale price for an apartment in a condominium with the features of a regular house (described at page 1) would be $\$59,896.46$. The standard error associated to this intercept is $1,884.33$, therefore the value of $\beta_0$ varies in a range equal to $\$59,896.46 \pm 1,884.33$.
- $\hat{\beta}_1$ is \$801.55, so, the estimated sale price for an apartment in a condominium with the features of a regular house increases of a value within the interval $\$801.55 \pm 171.68$ (considering the associated standard error) as the average of the walk and transit score increases of 1 point.
- $\hat{\beta}_2$ is \$819.62, therefore, on average, the price of an apartment in a condominium (with the features of a regular house) should increase of a value within the range $\$819.62 \pm 127.87$ as the school score increases of 1 point. This result is coherent with what we expected because a good rate of the schools is a factor that increase the distinction of that area and therefore the prices.
- $\hat{\beta}_3 = 0.39$, therefore, on average (with all the others predictors fixed as indicated in page 1), the price of an apartment in a condominium should increase of a value within the range $\$0.39 \pm 0.04$ as the estimate done by Zillow on the price of a house increases of 1\$. In the Figure 11 on the left it is possible to observe the effect of `Zillow_Estimate` on the `Sale_Price` when all the others predictors are kept constant. The blue area is the 95% confidence interval and it is quite thin compared for example to the one represented in the graph on the right.
- $\hat{\beta}_4 = (\frac{\$}{sqft}) - 15.81$, therefore, on average (with all the others predictors fixed as above), the price of an apartment in a condominium should decrease of \$15.81 as the dimension of the house increases of 1 sqft. This result seems to be quite unreasonable because it is unusual that a cheap house is a large house. The associated uncertainty is 3.29, therefore considering all the value of the interval $\$15.81 \pm 3.29$ the estimated $\hat{\beta}_4$ is always negative. In Figure 11 on the right is represented the effect plot of this predictor. It shows that houses with higher dimension are observed to be cheaper than smaller houses. It is essential also to notice that the 95% confidence interval in this case is quite large for high values of `Sqft` therefore there could be some inaccuracies.
- $\hat{\beta}_5 = 0.30$, therefore, on average the price of a regular apartment in a condominium should increase of \$0.30 as the Tax Assessment increases of \$1. This value can vary within the range $\$0.30 \pm 0.04$

- $\hat{\beta}_6$ is \$15,017.78, so, the estimated sale price for a multi-family house is \$15,017.78 more than a (regular) apartment in a condominium. This estimate can vary within the interval: $\$15,017.78 \pm 11,141.33$ that is quite a large error. Therefore, on average the price of a multi-family house should be about \$74,914.24.
- $\hat{\beta}_7$ is $(\$) - 9,396.56$, so, the estimated sale price for a single family house is \$9,396.56 less than the price of an apartment in a condominium. This result has to be considered within the interval $\$ - 9,396.56 \pm 2,917.45$. Therefore, on average a regular single family house cost less than an apartment in a condominium and less than a multi-family regular house.
- $\hat{\beta}_8$ is $(\$) - 791.92$, so, the estimated sale price for a regular townhouse is \$791.92 less than a regular apartment in a condominium. The standard error in this case is equal to 2,414.23 therefore, it is not certain that the fact that a regular house is a townhouse leads to a decrease in the sale price, in fact the estimates varies within the interval $[-3,206.155; 1,622.313]$

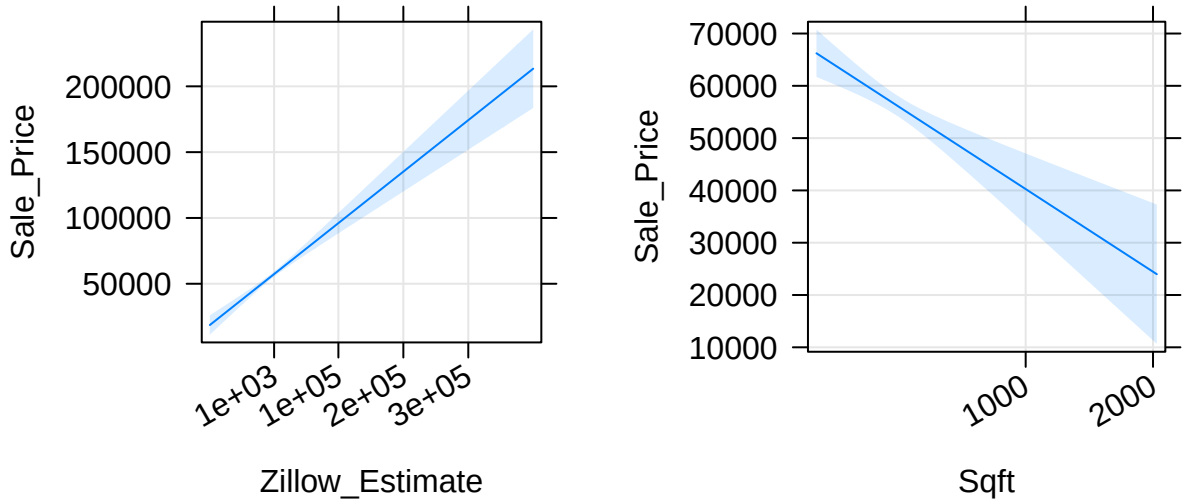



Figure 11: Predictor effect plot. On the left: SalePrice vs ZillowEstimate. On the right: SalePrice vs Sqft.

## 10. Residual standard error and Coefficient of determination

```
s <- summary(ols_out)
s$sigma
s$r.squared
```

On the whole, the model has a quite large $R^2 = 0.7642$ and an *adjusted* $R^2 = 0.7608$ that is larger than the one obtained in the first model considered with all the predictors that was (0.7337) and it is also higher than the *adjusted* $R^2$ of the model before removing outliers (0.7311). Therefore, this model explains approximately $76,42\%$ of the variability of the response variable `Sale_Price`. To conclude, the standard deviation of the model is \$23,650 therefore every prediction on the price of a house in Philadelphia should be considered in the range $\hat{y} \pm \$23,650$ that is quite large uncertainty considering that is a price in dollar to buy a house.

## 11. Test each of the predictors

Each predictor is tested with the following test of hypothesis:

$$\begin{cases} H_0 : \beta_j = 0 & \text{all the others coefficients arbitrary} \\ H_1 : \beta_j \neq 0 & \text{all the others coefficients arbitrary} \end{cases}$$

The t test for each predictor and the associated p-values are:

```
(s$coefficients)[, c(3, 4)]
```

```
##                             t value      Pr(>|t|)
## (Intercept)              31.7865452 6.746078e-127
## AvgWalkTransitscore       4.6689358  3.804283e-06
## SchoolScore               6.4097449  3.125374e-10
## Zillow_Estimate          10.3539288  4.402373e-23
## Sqft                     -4.8059160  1.987450e-06
## TaxAssessment             6.0728979  2.337007e-09
## PropTypeMultiFamily2To4   1.3479334  1.782317e-01
## PropTypeSingleFamily     -3.2208037  1.353424e-03
## PropTypeTownhouse        -0.3280216  7.430195e-01
```

The confidence intervals with a confidence level of 99% for each parameter are:

```
confint(ols_out, level = 0.99)
```

```
##                                 0.5 %        99.5 %
## (Intercept)              5.502593e+04 64766.9951680
## AvgWalkTransitscore      3.578072e+02  1245.2917092
## SchoolScore              4.891069e+02  1150.1368985
## Zillow_Estimate          2.934702e-01     0.4887410
## Sqft                    -2.430660e+01    -7.3050275
## TaxAssessment            1.726911e-01     0.4286223
## PropTypeMultiFamily2To4 -1.377978e+04 43815.3424470
## PropTypeSingleFamily    -1.693746e+04 -1855.6633513
## PropTypeTownhouse       -7.032109e+03  5448.2674840
```

The two previous outputs provide evidence against the null hypothesis for the parameters associated to the intercept $\beta_0$ and to the predictors "AvgWalkTransitscore" ($\beta_1$) ,"SchoolScore" ($\beta_2$), "Zillow_Estimate" ($\beta_3$), "Sqft" ($\beta_4$) and "TaxAssessment" ($\beta_5$). Therefore, these predictors are extremely significant at levels of 99% and and shouldn't be removed from the model. This conclusion is done due to the fact that the p-values of these predictors are really small ($p-value < 99\%$) and their confidence intervals with a significance level of 99% don't include the 0. The p-value of `PropTypeSingleFamily` is 0.00135, therefore this predictor is very significant and this is confirmed by the associated 99% confidence interval that doesn't include 0. Therefore, also $H_0 : \beta_6 = 0$ is rejected. Considering the predictors `PropTypeMultiFamily2To4` and `PropTypeTownhouse`it is possible to notice that their p-values are higher than 0.01, therefore they are not significant at a significance level of 99% and the null hypothesis can't be rejected. The same information is provided by their 99% confidence intervals that include the 0. This result was quite expected because in the best subset selection performed at point 4 it was included only the level `PropType = SingleFamily` in the best model with 6 predictors. This result suggests that the type of the property is significance in defining the price of a house only if it is a `SingleFamily` house.

## 12. Test groups of predictors

At this point all the regressors are tested jointly with the following hypothesis test:

$$\begin{cases} H_0 : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8 = 0 \\ H_1 : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8 \neq 0 \end{cases}$$

```
ols_nul <- lm(Sale_Price~1,
              data = house.prices, subset = cooks.distance(ols)<0.16)
anova(ols_nul, ols_out)
```

```
## Analysis of Variance Table
##
## Model 1: Sale_Price ~ 1
## Model 2: Sale_Price ~ AvgWalkTransitscore + SchoolScore + Zillow_Estimate +
```

```
##     Sqft + TaxAssessment + PropType
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    561 1.3118e+12
## 2    553 3.0937e+11  8 1.0024e+12 223.99 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that the p-value is less than $2.2 \cdot 10^{-16}$ therefore there it is possible to reject the null hypothesis.

Due to the fact that there are some predictors that are not really significant it would be interesting to test them. In particular the non-significant predictors are `PropTypeMultiFamily2To4` and `PropTypeTownhouse` that were not included in the best model selected at the point 4, however the best subset selection included `PropTypeSingleFamily` and for that reason all the `PropType` predictor has been included. For this reason at this point it is interesting to test these predictors with the following hypothesis test:

$$\begin{cases} H_0 : \beta_6 = \beta_7 = \beta_8 = 0 \quad \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \text{ arbitrary} \\ H_1 : \beta_6 = \beta_7 = \beta_8 \neq 0 \quad \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \text{ arbitrary} \end{cases}$$

```
ols_nest <- lm(Sale_Price~AvgWalkTransitscore+SchoolScore+Zillow_Estimate+Sqft
        +TaxAssessment,
             data = house.prices, subset = cooks.distance(ols)<0.16)
anova(ols_nest, ols_out)
```

```
## Analysis of Variance Table
##
## Model 1: Sale_Price ~ AvgWalkTransitscore + SchoolScore + Zillow_Estimate +
##     Sqft + TaxAssessment
## Model 2: Sale_Price ~ AvgWalkTransitscore + SchoolScore + Zillow_Estimate +
##     Sqft + TaxAssessment + PropType
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1    556 3.1820e+11
## 2    553 3.0937e+11  3 8833557513 5.2633 0.001381 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test it possible to affirm that there is evidence against the null hypothesis, therefore the qualitative predictor is significant.

## 13. Prediction

Considering a new observation with the following features (`AvgWalkTransitscore` = 71.16, `SchoolScore` = 23.57, `Zillow_Estimate` = 778,525.1, `TaxAssessment` = 132,400.1, `Sqft` = 1,290.84, `PropType` = "SingleFamily") it is possible to predict the price of the house with the corresponding prediction interval

```
newdata <- data.frame(AvgWalkTransitscore = 10, SchoolScore = 10, Zillow_Estimate = 660123,
                      TaxAssessment = 28827,Sqft = 1, PropType = "SingleFamily")
predict(object = ols_out, newdata = newdata, interval = "predict")
```

```
##        fit      lwr      upr
## 1 333540.7 267514.2 399567.2
```

```
vec_new <- c("Intercept" = 1, "AvgWalkTransitscore" = 10, "SchoolScore" = 10,
             "Zillow_Estimate" = 660123, "TaxAssessment" = 28827,"Sqft" = 1,
             "PropTypeMultiFamily2To4" = 0, "PropTypeSingleFamily" = 1, "PropTypeTownhouse" = 0)
xtx_1 <- s$cov.unscaled
pr.se <- s$sigma* sqrt(1+ t(vec_new)%*% xtx_1 %*% vec_new)
cat(333540.7 - pr.se,333540.7 + pr.se)
```

```
## 235296.2 431785.2
```

The result is that a `SingleFamily` house with the features specified above has a predicted price of $\$333,540.7$. The associated 95% confidence interval is [267,514.2; 399,567.2]. On average, the house defined above has an average variability equal to 98,244.5, therefore on average, the predicted price of the house will be within the interval [235,296.2; 431,785.2].

## 14. Simulation assuming the estimated parameters as the true parameters

It is now possible to simulate the values for response based on the estimated parameters used as true parameters:

```
r <- coef(ols_out)%*%t(model.matrix(ols_out))+ + rnorm(n = n, mean = 0, sd = s$sigma)
y <- house.prices[cooks.distance(ols)<0.16,]
plot(r, y$Sale_Price, cex = 0.15, pch =20, xlab = "Simulated values", ylab = "Observed response")
lines(c(-10,3000000), c(-10, 3000000), col = "red", lwd = 1, lty = 4)
```

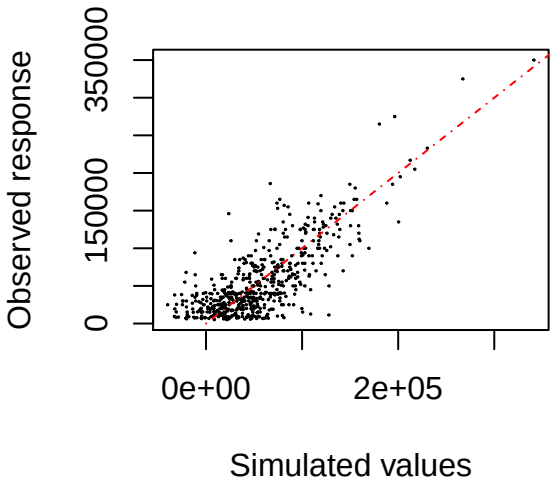



Figure 12: Simulated response and observed response

In Figure 10 are plotted the simulated response calculated with an error $\epsilon$ ($\epsilon \sim N(0, \hat{\sigma}^2)$ and the observed responses. From the graph it is possible to notice that just a few points lay on the diagonal and some of them are quite far from the diagonal. This result means that the fitted model is not really precise because just a few of the simulated response are correct.

All this analysis has provided information regarding the association between the response variable `Sale_Price` and the predictors. Some results are quite unexpected, for instance, the fact that the number of rooms in the house or the Violent Crime Rate are not considered good predictors to explain the response. The most strange result is the fact that the coefficient associated with the dimension of the house is negative, anyway this result can be caused by the fact that probably there are some cheap big house in the suburb of Philadelphia that influence the results. Moreover, as explained before the analysis is not really reliable due to the fact that the constant variance assumption is not respected.