Applied Probability and Statistics DECISION 518Q

Section C • Team 55 •
Jamelia Gordon
Yu Jun Kim
Vivian Yang
Dylan Liu
Xinyu Jiang

lm1<-lm(formula = sqrt(price) ~ factor(`City Zone`) * Type + Rooms * `m^2` + Bathrooms/Rooms + Rooms * Kitchen + `"Atico"`/`m^2` + Terrasse/`m^2` + Parking/`m^2`, data = train_data)

1. **Price ~.**

We noticed that the p values of Kitchen, Type and Yard were significantly high. Looking at the Price ~ Yard the p-value was 0.6, also the correlation between Yard and m^2 was high, and as such we believed the information provided by this variable could be captured elsewhere, so we decided to drop this variable. Similarly with Price ~ Elevator, the p-value was 0.15 and the R^2 was significantly low, and as such we believed it would not significantly explain the variability of Price

2. **Price ~ factor(`City Zone`)* Type**

Based on the City Zone we believed house owners might have different sentiments on owning an apartment or house. We therefore scaled the City Zone by Type.

3. **Price ~ Rooms * `m^2`**

We multiplied Rooms with m^2 because we believe that the price will increase when the house has more rooms and larger m^2. We first thought of using m^2/Rooms as the independent variable, however, the value of having one more room and the amount of increase in m^2 has a large difference, so having one more room actually drops the value of m^2/Rooms. Thus, we decided to multiply Rooms and m^2 because the more rooms they have, they will have more m^2 and this could highlight the effect of rooms and m^2 on the price.

4. **Price ~ Bathrooms/Rooms**

The relationship between Rooms and Bathrooms can be described by the linear relationship Rooms = 1.18 + 0.90 Bathrooms. Intuitively, the larger the number of rooms the larger the expected number of bathrooms. We therefore included the Bathrooms/Rooms to account for the incremental change in the number of Bathrooms, contingent on the number of rooms in the household.

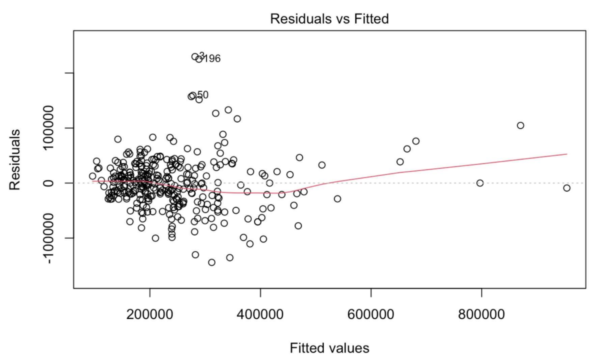
5. **Price ~ Rooms * Kitchen**

The independent variable Kitchen has a high correlation with m^2, Rooms and Bathrooms. Intuitively, the more rooms and the higher the square metric of a house, the higher the preference for an open kitchen space. We therefore introduced the variable Kitchen * Rooms to account for this relationship.

6. **Price ~ `"Atico"`/`m^2` + Terrasse/`m^2` + Parking/`m^2**

We noticed that the m^2 was highly correlated with most of the independent variables and as such we scaled a majority of the dummy variables by m^2.

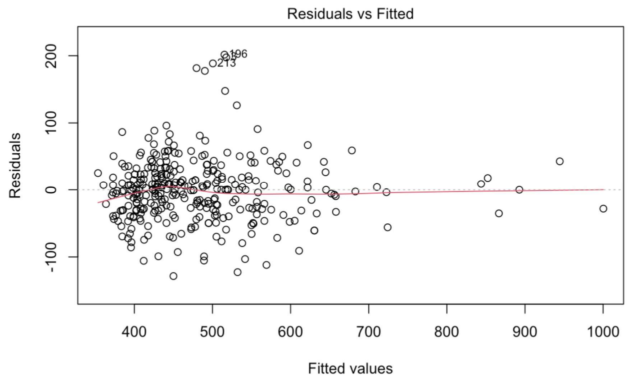
7. **Price Without Square Root** lm1<-lm(formula = price ~ factor(`City Zone`) * Type + Rooms * `m^2` + Bathrooms/Rooms + Rooms * Kitchen + `"Atico"`/`m^2` + Terrasse/`m^2` + Parking/`m^2`, data = train_data)



-> We first plotted a model without square root, we noticed a slightly quadratic relationship between the residuals and fitted model. Thus, we thought of using the sqrt(price) to make a more linear relationship. Therefore we decided to use sqrt(price) as the dependent variable for a

better implication. (We used predicted value ^ 2 to compare with the price.)

8. **Price With the Square Root** lm1<-lm(formula = sqrt(price) ~ factor(`City Zone`) * Type + Rooms *`m^2` + Bathrooms/Rooms + Rooms * Kitchen + `"Atico"`/`m^2` + Terrasse/`m^2` + Parking/`m^2`, data = train_data)



-> After square rooting the price, the trend between Residuals and Fitted reflected a more linear trend, Therefore, we decided to use sqrt(price) as our dependent variable for our most fitted linear regression.