uc3m | Universidad Carlos III de Madrid

Master in Statistics for Data Science
2024-2025

Regression models

"Modelling competition report"

Regresión al futuro:

Gabriel Pons Fontoira

Antía Enríquez Yurrebaso

Carlos Fernández Pascual

Raúl Rodríguez García

AI USAGE DECLARISON
AI tools were used throughout the writing process to grammar check some parts of the report and to detect some R-coding errors.

Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento - No Comercial - Sin Obra Derivada**

Contents

Introduction

The real state market is a vital sector of any economy, and its good health constitutes in this day and era one of the main challenges of our society. With prices surging, especially in big cities like Madrid, it is as important as ever to be able to understand as good as possible what affects the price of houses and buildings, best described as the price of square meter.

The aim of this project is to create an effective price prediction model, identify the importante home price atributes and validate the model's prediction accuracy. For that, information for almost 1000 residential properties in Madrid was used as training data (house characteristics and information about neighborhood and district) with a total of 35 predictors, as well as their sale price per square meter.

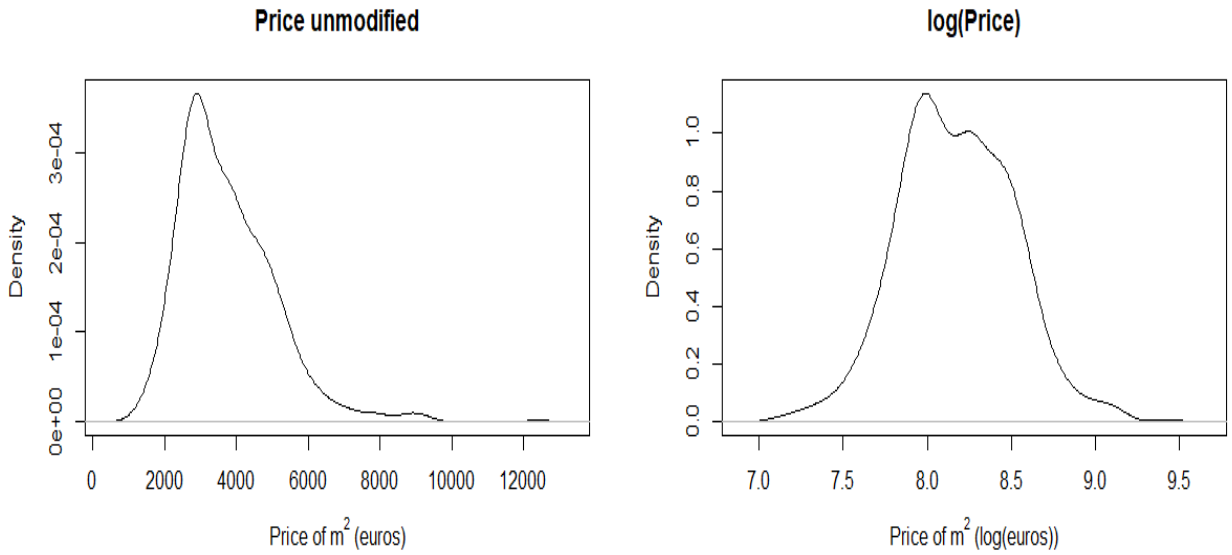
In order to do so, linear regression techniques will be applied to obtain the best possible result.

1. Preprocessing

In order to create a predictive model it is essential to read and transform the data correctly. Different techniques have been applied to the dataset.

1.1 Response variable

The response variable of the dataset is called *Precio.casa.m2* and represents the property sale price in euros per square meter of a house in Madrid. The density plot of the variable is represented in Figure 1.1a. The plot exhibits right-skewness, as most houses have prices concentrated in the lower range of the distribution. Specifically, a large proportion of houses are priced between 2000 and 5000 euros per square meter, while a smaller number of observations have significantly higher prices. This asymmetry is reflected in the summary statistics: the median (3538) is smaller than the mean (3761), and the mode (3000) is smaller than the median. That is: Mode = 3000 $\leq$ Median = 3538 $\leq$ Mean = 3761.



(a) Original distribution of prices per square meter.

(b) Logarithmic transformation of prices.

Figure 1.1: Comparison between distribution of prices with log-transformation applied.

To address this skewness and approximate a symmetric distribution, a log-transformation was applied to the data, represented in Figure 1.1b. Post-transformation, the distribution appears more symmetric and closely resembles a normal distribution. This improvement facilitates more robust statistical analysis and modeling, as many statistical techniques assume normality or symmetry in the data. Therefore, from now on the modeling process will be done using the log-transformed response variable, *Precio.casa.m2_log*.

1.2 Predictor variables

The dataset contains 35 different predictor variables, some of them need to be treated as categorical and others as numeric. The first step is to convert character, binary, multiclass and discrete variables into factors. The variables converted into factors are: *Barrio, Cod_barrio, Distrito, Cod_distrito, Dorm, Banos, Tipo.casa, Inter.exter, Ascensor, Estado, Comercial, Casco.historico* and *M.30*. All

these variables will be considered as categorical variables. This transformation provides a more accurate representation of the data and enhances the interpretability of visualizations such as boxplots, where categories are clearly distinguished.

Furthermore, a data-cleaning process was conducted in order to detect anomalies in the data. During the process, observation with *Train_indices* = 302 was identified as illogic since the useful surface area exceeded the built surface area. This error could be due to a data entry error. Therefore, this observation was excluded from the dataset.

1.2.1 Transformation of categorical variables

When including a categorical variable in a linear regression model, it is important to ensure that it does not have more than four or five categories and that each category is represented at least by 5% of the data. This is because having too many categories can lead to overfitting, making the model too complex and less generalizable. Additionally, each category is represented by a dummy variable and having too many dummy variables can cause multicollinearity issues, which negatively impact the model's performance and interpretability. Therefore, the categorical variables that had more than five categories or categories with less than 5% of the data were merged into others or new categories were created.

Seven different categorical variables were transformed. The variable *Dorm*, which indicates the number of bedrooms, was reduced from eight to four categories. The new categories are 0-1, 2, 3 and 4 $\leq$. Similarly, the variable *Banos*, which indicates the number of bathrooms, was reduced from seven categories to three. The new categories are 1, 2 and 3 $\leq$. The category *Tipo.casa*, which represents the type of property, originally had six categories. This variable has been transformed into three categories: atico (remains the same), big (which includes chalet and duplex) and piso (which includes the categories Otros, piso and estudio).

Moreover, the category *Estado*, which represents the condition of the property, has been reduced to two categories: good (which includes excelente, buen_estado, reformado, nuevo-semin and segunda_mano) and bad (which includes reg,-mal and a_reformar). The binary variables *Comercial*, *Casco.historico*, and *M.30* have been recoded so that category 1 corresponds to si and 0 to no.

Variables *Barrio* and *Cod_barrio* had 116 different categories and *Distrito* and *Cod_distrito* had 20 different categories. Various options were considered with the aim of merging the categories, but none of them were satisfactory, thus, these variables remain the same.

1.2.2 Fitting simple linear models

After correctly reading the data and transforming the neccesary variables, a simple linear model was adjusted with each of the variables. The results indicated that the variables *Nox* and *NO2* were not significant according to the Wald test with a p-value > 0.4 and presented a $R^2 < 0.01$. Additionally, the variable *Inter.exter* presents nonsensical results, as the interior category predicts higher property values than the exterior category.

1.3 Preliminary selection of predictor variables

An essential step in this process is refining the dataset by addressing the large number of variables it contains, some of which are not useful. Since a key objective of the model is to simplify and identify the most predictive variables, it is important to exclude less relevant ones early in the analysis.

The variables *Barrio* and *Distrito* were excluded because they contained the same information as *Cod_barrio* and *Cod_distrito*, respectively. Furthermore, *Cod_barrio* and *Cod_distrito* were excluded because they had a large number of categories and some of them with less than 5% of the data.

Nox and *NO2* were excluded due to their lack of significance when fitting a simple linear model and *Inter.exter* was excluded because the results obtained from fitting the simple linear model were illogical.

1.3.1 Correlation study

Studying the correlation between predictor variables when fitting a linear model is essential to address multicollinearity, which can distort coefficient estimates and inflate standard errors, making statistical tests and predictions less reliable. Understanding these correlations helps create a more stable and interpretable model by reducing redundancy and noise among predictors.

Therefore, the correlation between numeric variables and the association of categorical variables was analyzed in order to avoid multicollinearity and achieve a simpler and more robust model with improved predictive performance.

Numeric variables

Correlation is a statistical measure that indicates the strength and direction of a linear relationship between two variables. It can be computed as follows It can be computed as follows:

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

where X and Y are two numeric variables. Recall that $r \in [-1, 1]$ and that the closer $|r|$ to 1, the stronger the relationship between the variables.

The correlation between all the numeric variables was calculated. The variable pairs that presented $r > |0.6|$ are shown in Table 1.1.

Variables	Correlation	R^2
Sup. const vs Sup. util	0.989	0.01649 vs 0.01759
PM10 vs CO	0.800	0.0789 vs 0.1205
Poca limpieza vs Delincuencia	0.767	0.06649 vs 0.07784
Nox vs NO2	0.748	0.0006 vs 0.0011
Delincuencia vs Inmigrantes	0.668	0.07784 vs 0.04434
SO2 vs Pob. <14	0.623	0.2366 vs 0.2137
SO2 vs ref.hip	0.59	0.2366 vs 0.4587

Table 1.1: Correlation of numerical variables.

For each pair of variables, those highlighted in blue were removed due to their lower R^2 when selected as the sole regressor in the model. A specific case involves the pair *Nox* and *NO2*. Both variables were previously removed because, when taken as the sole regressor, their p-value exceeded 0.4, indicating a lack of statistical significance. Another particular case happens with *SO2*, which explains slightly more variability than *Poblj14*, but much less than *ref.hip*.

Categorical variables

To measure the association between categorical variables, Cramer's V has been computed for each pair of categorical variable, which can be obtained using the following formula:

$$V = \sqrt{\frac{\chi^2}{n \times (k-1)}}$$

where χ^2 is the Chi-squared statistic, n is the total number of observations and k is the minimum of the number of categories in the two variables, i.e., $k = \min(r, c)$, where r is the number of rows and c

is the number of columns in the contingency table. $V \in [0, 1]$ and that the closer V to 1, the stronger the relationship between the variables.

The association between all the categorical variables was calculated and it was found that the association between *Casco.historico* and *Comercial* is 0.85. *Comercial* was eliminated following the same criterion: it explains less variability as a single regressor.

1.3.2 Multicollinearity detection

Multicollinearity detection is a crucial step in the process of building a reliable linear model. Multicollinearity occurs when predictor variables are highly correlated with each other. This is why it was decided to remove some of the highly correlated variables. This section outlines the methods used to identify multicollinearity. The study was performed with the variables that had not been previously eliminated.

Variance Inflation Factor

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model. Specifically, VIF quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity among the predictor variables.

The VIF for a predictor X_j is calculated by regressing X_j on all the other predictors and determining the R_j^2. for that regression. So if R_j^2 is close to 1, it means the j-th predictor is highly correlated with the other predictors, leading to a high VIF value. The VIF can be obtained as follows:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

A VIF value greater than 10 is often considered indicative of severe multicollinearity. In simpler terms, the higher the VIF, the more problematic the multicollinearity. The VIF values of the remaining variables are shown in Table 1.2. None of the values are higher than 10, thus, there are no indices of severe multicollinearity.

Longitud	*Latitud*	*Sup.util*	*Ref.hip.zona*	*Antig*	*Ruidos_ext*	*Mal_olor*	*Malas_comunic*	*Delincuencia*	*CO*	*O3*
1.667449	2.545400	1.116517	2.053073	1.096460	2.440324	3.437244	1.486283	3.225588	1.908513	1.229370

Table 1.2: VIF values for each variable.

Condition number

The condition number (C.N.) is another measure to calculate the multicollinearity between numerical regressors. A high condition number indicates that the matrix is nearly singular, which can lead to unstable and unreliable regression estimates. The condition number of the selected variables is obtained as follows:

$$C.N. = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = 122564.4$$

As the obtained C.N. is large, a 'for' loop was coded where the C.N. is obtained by removing the $j-th$ variable from the matrix. This insight retrieves high C.N. values for every variable, concluding the same as the general C.N.

Contradictory results were obtained. Initially, it will be assumed that multicollinearity does not exist because the VIF is used for both categorical and numerical variables, while the condition number (C.N) applies only to continuous variables. Later, a section will be dedicated to Ridge regression in case multicollinearity does indeed exist.

2. Variable selection

After selecting the dataset variables, taking into account the correlations and associations that exist between features, the best subset of predictors must be chosen. To do so, the selected approach is stepwise regression, considering two different criteria: AIC and BIC.

2.1 Stepwise Regression

The stepwise regression combines both forward selection and backward elimination. The process chosen begins with a full model. At each step, a predictor can be added to the model if it significantly improves the fit based on the chosen criterion, which will be either AIC or BIC. Simultaneously, predictors already in the model are evaluated, and any that no longer contribute significantly are removed. This combination of adding and removing predictors continues iteratively until no further improvements can be made, ensuring the model is neither overfitted with unnecessary variables nor underfitted by missing important predictors. This bidirectional approach provides a balanced and flexible way to refine the model, but it can be sensitive to multicollinearity and may not always produce the optimal model due to its reliance on local improvements at each step. AIC and BIC can be obtained using the following formulas:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

where k is the number of parameters in the model, n is the number of observations, and $\hat{L}$ is the maximum likelihood of the model.

Stepwise regression was applied to fit two different models: one using AIC as the criterion and the other using BIC. The models obtained have the following characteristics.

Criteria	$\mathbf{R^2_{adj}}$	**Number of coefficients**	**Non-Significant coefficients**
AIC	0.6091	18	2
BIC	0.6002	13	0

Table 2.1: Comparison of AIC and BIC Models.

Regarding these two models, as both have almost the same adjusted R^2 (R^2_{adj}), the best option is the BIC model. This is because it is a simpler model and all variables are significant fixing $\alpha = 0.05$. In particular, the two non-significant variables in the AIC model hold 0.114 and 0.059 p-values.

2.1.1 Models with interactions

Regarding the possibility of introducing interactions in the model, stepwise regression was repeated including all possible interactions. The model obtained using AIC criterion includes 72 coefficients and only improves the adjusted R^2 by 6.32% (to 0.6634), indicating that it is very complex with only a modest improvement. Furthermore, it contains 27 non-significant coefficients.

On the other hand, the model obtained using BIC criterion presents 21 coefficients, only three of them are non-significant and the adjusted $R^2 = 0.622$. That is, an improvement in the explained variability of 2.18% in exchange of adding eight coefficients. The non-significant coefficients and their respective p-values are shown in Table 2.2.

Coefficient	p-value
Sup.util	0.454233
Banos2	0.618384
Sup.util:Banos2	0.604653

Table 2.2: Non-significant coefficients.

As the coefficient *banos1* is still significant, this means that the factor leveling can be reshaped adding the properties with 2 bathrooms to the next category (3 or more bathrooms) in order to keep it in the model with the desired signification level. Now, the variable *Banos* indicates weather if the property has 1 bathroom or more than one. When the model is recomputed with this changes, the obtained Adjusted $R^2 = 0.6247$ with 21 coefficients. Recalculating the model has driven to another change. *Pocas_zonas* is not significant anymore with a p-value of 0.149234 and *Sup.util* now it is significant.

When *Pocas_zonas* and its corresponding interactions are deleted the obtained model presents 19 coefficients, none of them non-significant and $R^2_{adj} = 0.6212$.

Therefore, the final chosen model is the latest model proposed. The variables taken into account for this model are: *Latitude*, *Sup.util*, *Ref.hip.zona*, *Antig*, *O3*, *Pobl.0_14_div_Poblac.Total*, *PoblJubilada_div_Poblac.Total* (continuous variables), *Banos*, *Tipo.casa*, *Ascensor*, *Estado*, *Comercial* and *M30*. As for the interactions between the variables, only interactions between categorical and continuous variableswere considered. The only interactions considered were of *Sup.util* with *Banos*, of *PoblJubilada_div_Poblac.Total* with *Tipo.casa*, and of *O3* with *Ascensor*.

2.2 Ridge regression

A different variable selection approach that was taken into account is the Ridge regression. Ridge regression is a technique used to address the problem discussed during section 1.3.2. This technique deals with multicollinearity in linear regression by adding an L2 regularization penalty to the loss function. In situations where predictor variables are highly correlated, Ridge regression helps to stabilize the estimates of regression coefficients by shrinking them towards zero, though not exactly zero. The objective function for ridge regression minimizes the residual sum of squares while adding a penalty proportional to the square of the magnitude of the coefficients. Mathematically, it can be expressed as:

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where y_i is the actual value of the response variable, $\hat{y}_i$ is the predicted value, β_j are the regression coefficients, λ is the regularization parameter that controls shrinkage, and p is the number of predictors.

To apply this regression, all the predictor variables where used with the exception of *Barrio, Distrito, Cod_barrio* and *Cod_distrito* due to their high number of categories, *Nox* and *NO2* because there are not significant variables when they are modeled individually to predict the prices, *Sup.const* because it presents a correlation of 0.98 with *Sup.util* and *PM10* because it has a correlation of 0.8 with *CO2*.

Once this is done, with the fitted Ridge regression model, a cross-validation is performed to improve model performance by controlling the magnitude of the regression coefficients and identifying the optimal value of the regularization parameter λ, which can be used to make predictions.

However, it is important to note that Ridge regression has a drawback: it does not allow inference about the individual predictors. This means that while Ridge regression can improve prediction accuracy, it does not provide insights into the significance of individual variables.

3. Best model selection

This section focuses on selecting the best model for predicting the price per square meter in Madrid. There exist different measures to calculate the predictive power of a model, i.e., Cross-validation (CV) or Mean Square Error (MSE). However, in this project the predictive power of the models is calculated using the Root Mean Square Error (RMSE), which measures the differences between the predicted values and the real values. The RMSE can be obtained using the following formula.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents the observed values and $\hat{y}_i$ represents the predicted values.

In practice, the dataset is divided into two parts: a training set and a test set. Then, the model is fitted using the training set and evaluated on the test set. Nonetheless, in this case the test set does not include the response variable. Therefore, a slightly different method has been applied in order to obtain the predictive power of the models explained in section 2.

Five models are studied in this section. The first model, model ω_1, includes all the variables selected in section 1, the second model, model ω_2, includes all the variables of the previous model with all the possible interactions. The third model, model ω_3, is obtained using stepwise regression starting from the first model and using BIC as the criteria. The fourth model, model ω_4, is obtained in a similar way to the previous model but the stepwise regression starts with the second model, i.e., with the model with all possible interactions. The last model, model ω_5, is fitted with Ridge regression as explained in section 2.

Once the models were decided, the dataset containing the response variable (*data_train*) was divided into two parts: new training and test sets, using the same proportion as in the original data. Specifically, 74.87% of the individuals belong to the training set and the remaining 25.13% belong to the test set. The sampling was done randomly. The selected models were then fitted using the new training set and evaluated based on their RMSE on the new test set. This process was repeated 500 times, changing the partition in each iteration. Note that the same 500 partitions were used for testing all the proposed models.

The RMSE is going to be calculated in two different ways; 1) in logarithmic scale, that is, with the values obtained in the predictions and 2) in the original scale, transforming the predictions with the exponential function.

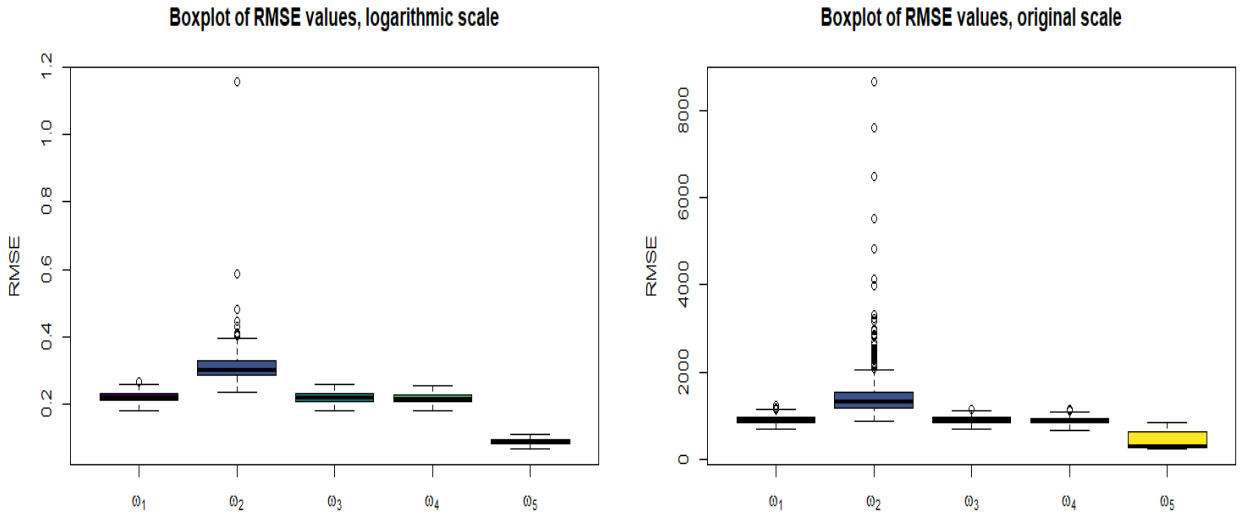
The RMSE values for each model are shown in Tables 3.1 and 3.2. These values indicate how well each model predicts the price per square meter on the test set. Recall that lower RMSE values indicate better predicting power. Their respective boxplots are represented in Figure 3.1.

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ω_1	0.1808	0.2098	0.2206	0.2208	0.2314	0.2655
ω_2	0.2336	0.2849	0.3021	0.3115	0.3301	1.1565
ω_3	0.1788	0.2081	0.2193	0.2190	0.2290	0.2563
ω_4	0.1807	0.2062	0.2158	0.2158	0.2251	0.2527
ω_5	0.0646	0.0805	0.0863	0.0865	0.0923	0.1083

Table 3.1: Summary of RMSE values for different models using the logarithmic scale.

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ω_1	682.1	846.5	906.5	908.4	965.1	1222.1
ω_2	876.3	1188.4	1318.4	1495.1	1542.1	8656.2
ω_3	682.4	837.9	901.1	897.0	952.6	1149.2
ω_4	667.9	831.2	892.0	890.7	944.0	1157.2
ω_5	225.2	269.3	297.3	394.1	621.1	854.5

Table 3.2: Summary of RMSE values for different models using the original scale.



(a) (b)

Figure 3.1: RMSE values.

As expected, the RMSE values using the logarithmic scale are lower than those using the original scale. Overall, the models behave similarly compared to each other with both scales.

The results indicate that model ω_5, fitted with Ridge regression, consistently outperformed the other models on both scales. Specifically, model ω_5 achieved the lowest RMSE values, with a median RMSE of 0.08632 on the logarithmic scale and 297.3 on the original scale. This demonstrates its superior predictive accuracy and generalization capability. However, this model is based on the perturbation of the original dataset to avoid collinearity, and in this way it is less interpretable than multiple linear models without penalty. In this sense, and even though the Ridge model showed greater predictive power, the next best model was chosen, in order to understand better the model and the conclusions. The chosen model was a model with interactions obtained through BIC minimization, ω_4.

4. Results of the model

4.1 Interpretation of the results of the model

As it had been seen previously from the analysis of the goodness of fit of the different models through their adjusted R^2 and the RMSE obtained through the cross-validation method, the model chosen was a linear model with interactions obtained through a stepwise method centered around minimizing the BIC of the models.

The model obtained a R^2_{adj} of 0.6212, that is, it was able to explain a 62.12 % of the variability of the data. In addition to this, the RMSE obtained in the cross validation is of **0.2158** for the predictions of the logarithm of the price of the square meter and of **890.7** for the prediction for the price of the square meter itself.

The variables taken into account for this model are: *Latitude*, *Sup.util*, *Ref.hip.zona*, *Antig*, *O3*, *Pobl.0_14_div_Poblac.Total*, *PoblJubilada_div_Poblac.Total* (continuous variables), *Banos*, *Tipo.casa*, *Ascensor*, *Estado*, *Comercial* and *M30*. As for the interactions between the variables, only interactions between categorical and continuous variableswere considered. The only interactions considered were of *Sup.util* with *Banos*, of *PoblJubilada_div_Poblac.Total* with *Tipo.casa*, and of *O3* with *Ascensor*. For a linear model with categorical and continuous variables, the interpretation of the coefficients are as follows.

For only continuous variables, the coefficients of the variables represent the expected change in the mean of the response when all other predictors are held constant and the variable increases by one unit (the slope of a one dimensional linear regression, as only one variable changes). When categorical factors are involved, the coefficients corresponding to each of the categories of a variable represent the difference in mean of the response with respect to the reference category (included in the intercept). When interactions between continuous and categorical variables are involved, the coefficients of the continuous variables are the expected difference in the mean of the response when the variable changes by one unit *in the reference level* when all other covariates are held constant. The interaction components are how much that difference in the expected mean of the response changes between groups of a variable under the same circumstances (the difference in slopes between the groups of a one dimensional regression is considered with only the two variables of the interaction as predictors). The coefficients associated to the categories of the categorical variables correspond then to the difference between the expected value of the response when all continuous variables are set at 0 and all other categorical variables are held constant.

Knowing this, we can give an interpretation of the parameters of the model. First of all, it is vital that the reference category is clear in order to be able to give an interpretation of the coefficients. This reference category corresponds to houses with more than one bathroom, belonging to the *Big* category (Chalets and Duplex), without an elevator, in good state, not in a commercial district and outside the M30.

The interpretation of the continuous variables without an interaction is pretty straight forward, it gives the impact of an increase of said variable when everything else is held constant. Under equal conditions, then, an increase in latitude or mortgage reference zone means an increase in the expected price. This leads us to believe that, the norther the property is, the more expensive the square meter is, in concordance to the general belief and data that in general the north of Madrid is more expensive (however, it is expected that latitude has a non linear relationship also having to do with the distance

to the center). Also, the mortgage reference zone seems to indicate that higher values correspond to higher prices (the interpretation of the variable is confusing).

In return, an increase in the age of the property or in the amount of infant population leads to a decrease of the price. It makes sense that older houses are less expensive in principle under the same conditions, for then problems due to wearing of the property.

For all of these variables, it is found that, keeping all other covariates constant, the price for atics is higher than the one for flats and this one is higher than for big properties; the price for houses with an elevator is higher than the for the ones without it; that houses in a bad state have cheaper prices of square meter than the ones in a good state; that houses in commercial districts are more expensive than the ones not in them; and that houses inside the M30 have more expensive prices than the ones outside. All of these coefficients are in accordance to what is expected.

Finally, for the coefficients of the interaction of continuous with categorical variables, the expected change of the response due to an increase of said variable for each group is not the same. For that, we find that, when the rest of the covariates are held constant, an increase in the useful surface leads to an increase of the price of houses with more than 1 bathroom, but to a decrease of the price for houses with one bathroom; an increase in the percentage of retired population leads to an increase of the price for flats and big properties, but to a decrease of the price for atics; and an increase in O3 leads to an increase in the price for houses without elevator, but to a decrease for houses with an elevator.

With this, a profile for the properties with the cheapest and most expensive prices of square meter can be drawn. The properties with the highest price per square meter correspond to newer in a good state houses to the north, with high values for the mortgage reference zone, in areas with smaller percentage of children, and located in commercial areas inside M30. Analogously, the properties with the lowest price per square meter correspond to older houses in a bad state to the south, with low values of the mortgage reference zone, in areas with higher percentage of children, and located outside of commercial areas and outside the M30. As for the rest of the variables that are not mentioned, the profile depends on the specific values that the interacting values take, so a profile cannot be drawn for a general case when taking them into account.

In the study of this problem, it was found that the variable that contributed the most to explain the variability in price was the variable of the neighborhood code, with a R^2_{adj} of 0.5136 in a simple linear regression. However, this variable is a categorical one with 116 different categories, and as such it cannot be taken into account, for they are too many to obtain reliable and interpretable results. One idea that arose during the project was to try and merge groups, but the codes did not seem to have a clear relationship with the price (the ordering of the codes was not something known). One could be prone to think that the codes could be grouped into levels according to the prices associated to them in the data, but that would create great overfitting, and thus this idea was rejected. The variable that explained the most percentage of variability after the neighborhood code was the variable *Ref.hip.zona*, with a R^2_{adj} of **0.458** in a simple linear regression. Indeed, this variable appeared in all the models fitted. The interpretation of the variable (from the information provided, that was nevertheless not very clear) is that it makes reference to similar properties that have approximately the same value. In this way, it contains local information of the neighborhood of the points taken into account, somehow "smoothing" the information. Taken into account this understanding of the local nature of the variable *ref.hip.zona*, a GAM model was explored (also one based on interactions between latitude and longitude to create a spatial model), but the results did not prove to be an improvement.

As for the variables that explained the least amount of the variability, they are *NO2* and *Nox*. The R^2 of their correspondent linear regressions are of the order of 10^{-4}, so they explain around 0.1 % of

the total variability. Due to this, they were discarded for the creation of the models.

4.2 Diagnosis of the model

Before predicting values with model, the diagnostics of it was performed, in order to check if the basic hypothesis of the linear model were satisfied, and thus, whether inference could be performed on the model, in order to be able to provide significative confidence intervals for the predictions.

4.2.1 Normality

The normality can be checked through a QQ plot of the residuals of the model, comparing them to a normal distribution. The chosen residuals were the standarized residuals. Figure 4.1 shows said QQ plot, in which the residuals seem to follow a normal distribution.

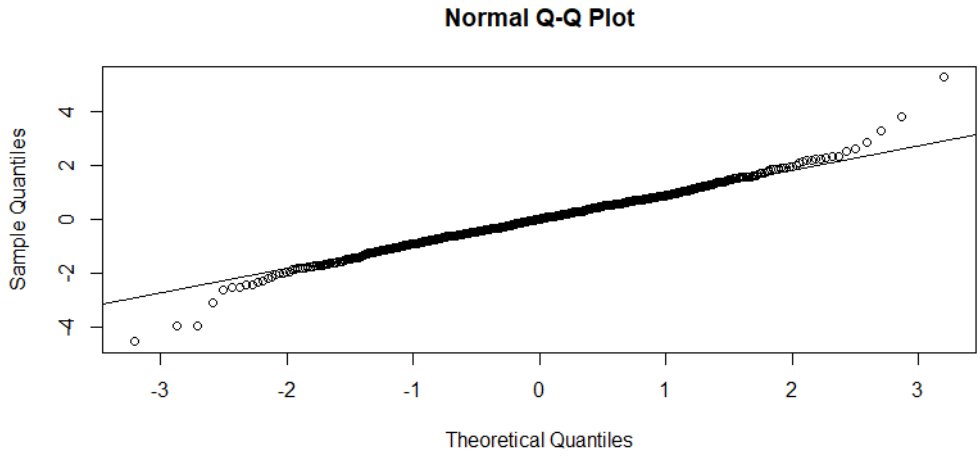


Figure 4.1: QQplot of the residuals

In addition to this, a normality test was performed on the residuals, the Lilliefors test (a correction of the Kolmogorov-Smirnov test), with null hypothesis that the data belongs to a normal distribution, and alternative hypothesis that it doesn't. The p-value obtained by the test was 0.0843, and so, as it is larger than 0.05, we do not have evidence to reject the null hypothesis and assume that the residuals are normal.

4.2.2 Homocedasticity and linearity

To check for homocedasticity and linearity, the standarized residuals were plotted against the fitted values of the model, as shown in Figure 4.2.

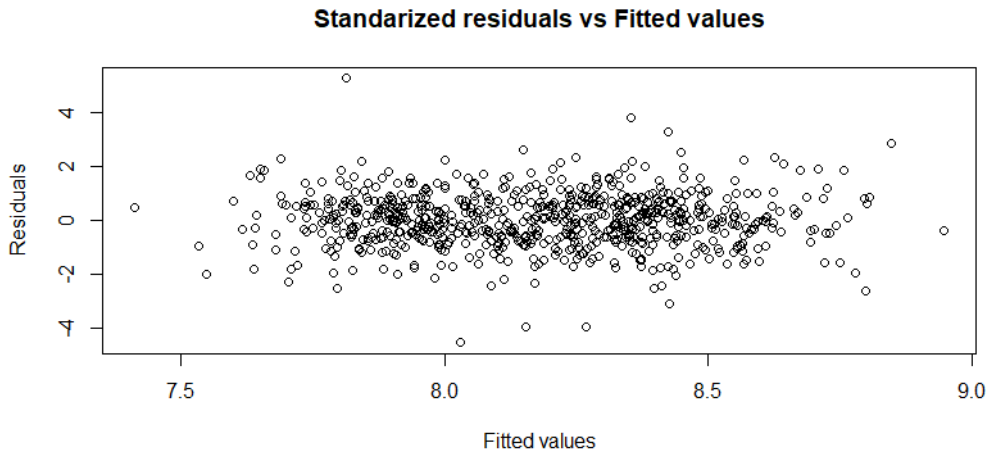


Figure 4.2: Standarized residuals versus fitted values

As there does not seem to be any trend in the residuals, it is concluded that the linearity of the model holds. Also, the residuals are distributed along a band of constant width, so the hypothesis of homocedasticity is assumed to be satisfied.

The fact that the hypothesis of the models are satisfied means that the distributions of the predicted mean and the predicted distributions are as shown by Equations 4.1 and 4.2, which allows for the creation of confidence intervals based on the quantiles of a normal distribution, as well as for the calculation of probabilities for the values taken by the predicted instances.

$$\overline{\hat{Y}_h} \sim N(X_h\beta, \sigma^2 X_h(X'X)^{-1}X_h') \tag{4.1}$$

$$\hat{Y}_h \sim N(X_h\beta, \sigma^2(1 + X_h(X'X)^{-1}X_h')) \tag{4.2}$$

As all hypothesis of the linear model are satisfied, it is possible to perform and produce significative confidence intervals. In addition to this, as the hypothesis hold, it was decided not to go along with the suppression of possible outliers from the data, as this would reduce the effective predictive range of the model, which frontally collides with the objective of the project, to predict new prices.

4.3 Prediction of new values

To end the project, starting from a new dataset, the final objective was to predict the prices of square meter for the new individuals. Figures 4.3 and 4.4 shows the predicted values along with 95 % confidence intervals for both the prediction of the mean and the prediction of the individuals. The actual predictions for both values are the same, but the confidence intervals are bigger in the case for the prediction of individuals.

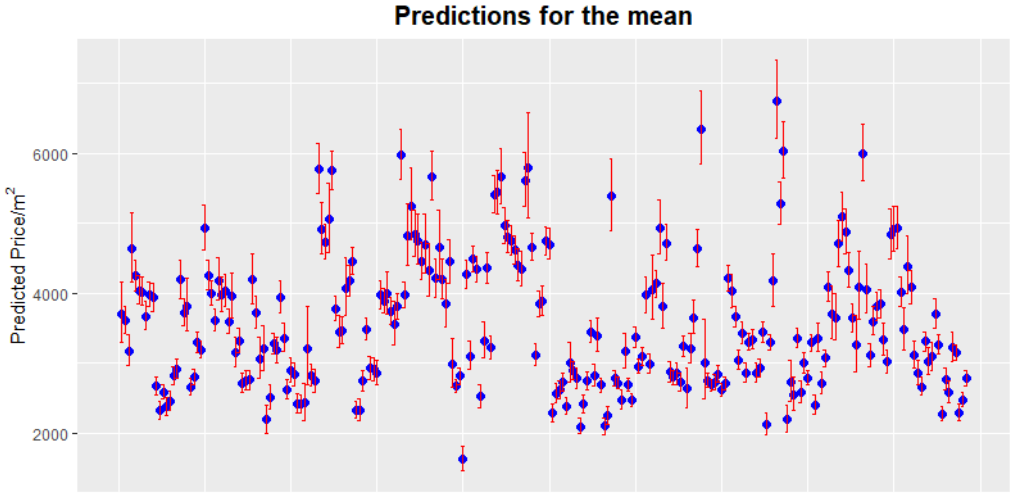


Figure 4.3: Predictions for the mean

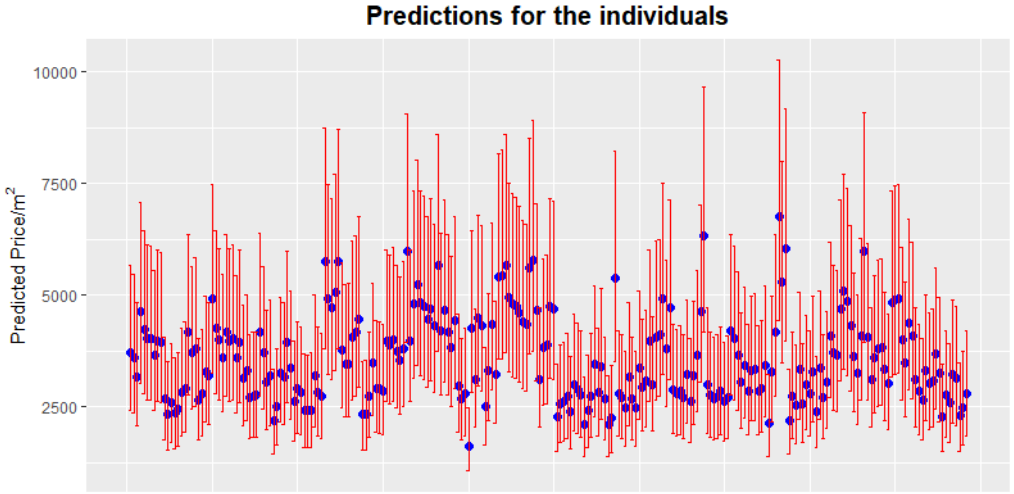


Figure 4.4: Predictions for the individuals

5. Conclusion

In this project, the objective was to create an effective price prediction model, identify the importante home price atributes and validate the model's prediction accuracy.

First of all, a preprocessing of the data was performed, as a preliminary analysis and an early selection of variables, as well as some transformations for the better understanding of the data and to improve the posterior performance of the model. The variables that were more highly correlated to other ones were dismissed, and measures of collinearity of the remaining ones were calculated.

Afterwards, in order to select between those variables, and trying to eliminate those of lesser importance to the explanation of the information contained in the data, an stepwise method based on the minimization of the BIC of the models was performed. AIC was also tried, but as its criterion is less severe, more variables were considered, and thus the models were bigger and harder to work with. The final model considered, the one that best explained the variability of the data, that is, the one with the highest R^2_{adj}, (0.6212), was a model with continuous and categorical variables and their interactions obtained through BIC minimization. Also, a ridge regression model was considered, as the lack of collinearity in the dataset was not clear, in case the results were better.

After the model was in principle selected, its predictive power was measured through cross validation on the training data through the RMSE obtained, which was also obtained for the ridge model and other models. It was found that the ridge model had the lowest RMSE, but still the model with interactions obtained through BIC was chosen, as its interpretability was greater and the important attributes were more easily interpretable. Moreover, it was found that the model satisfied the hypothesis for multiple linear regression, and thus inference could be performed on the predictions to provide reliable confidence intervals.