Introduction to Regression Models

Johanna Norell

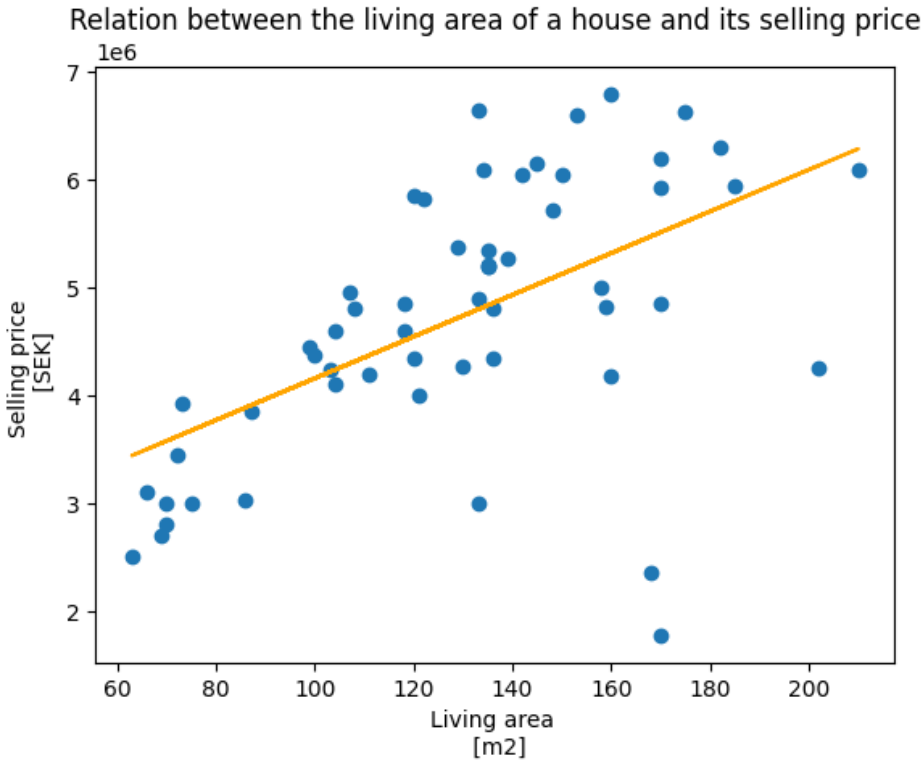
March 1, 2024

Basic Regression Modeling

1. Find a linear regression model that relates the living area to the selling price. If in doing so, you performed any data cleaning step(s), describe what you did and explain why.

The data imported from Hemnets website was very clean and no data wrangling had to be done in order to perform a linear regression.

We can see in the scatter plot with the regression line added to it that there are at least some correlation. It will be easier to analyse it with the residual plot that is further down in the document.



2. What are the values of the slope and intercept of the regression line?

Table 1: Regression Coefficients

Slope	Intercept
19370.14	2220603.24

The slope represents how much the price increases with increasing living area. According to our linear regression, the price increases with around 20 kSEK per additional m2. The intercept represent how much a house with the least living area, in this case 60m2, would cost. According to our linear regression, a house of 60m2 would cost around 2.2 mSEK.

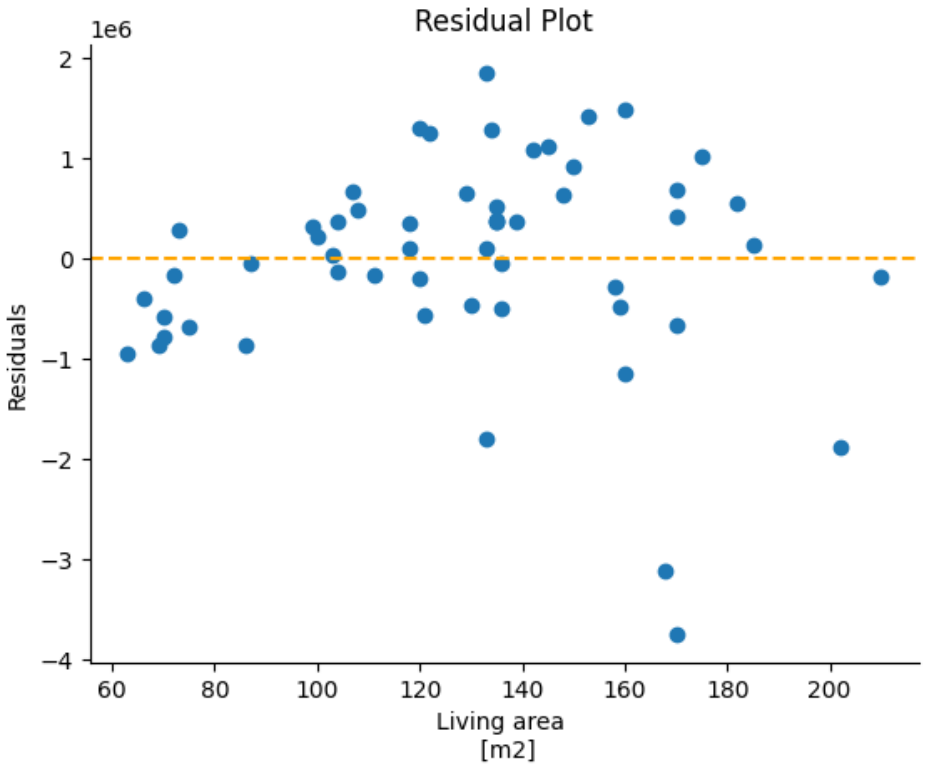
3. Use this model to predict the selling prices of houses which have living area 100 m2, 150 m2, and 200 m2.

Using the slope and intersection we can create the linear function the regression line is (y = k * x + m). We simply plug in the values (100, 150, 200) to the equation in order to get the predicted price.

Table 2: Predicted Selling Prices for Different Living Areas

Living Area (m^2)	Predicted Selling Price (SEK)
100	4157617.1
150	5126124.0
200	6094631.0

4. Draw a residual plot. Discuss some potential strategies for improving the model.

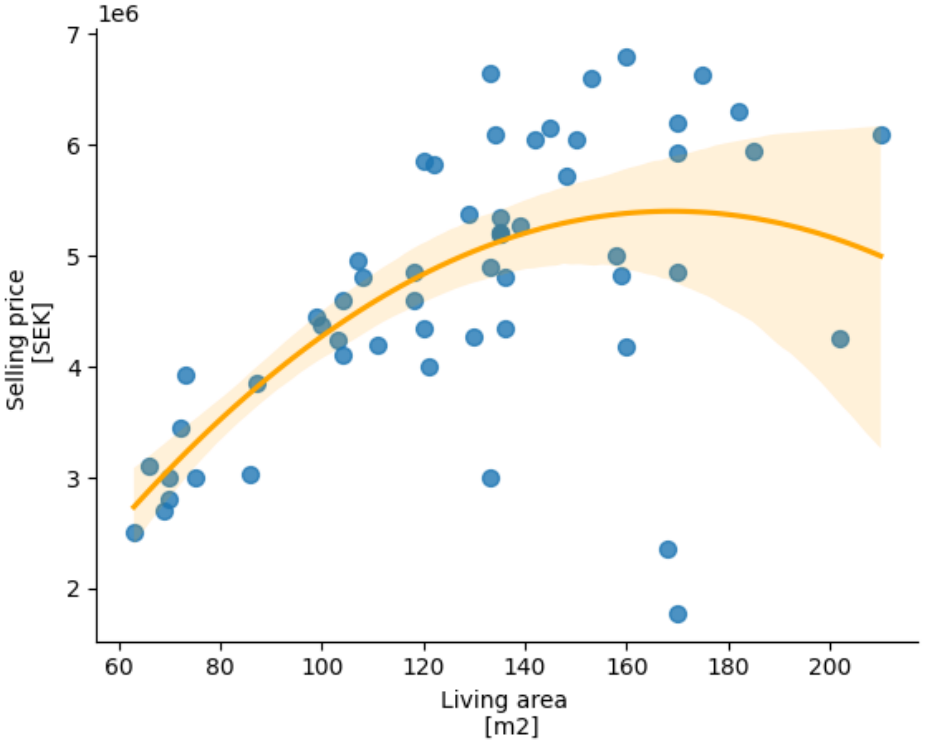


As previously stated, the residual plot makes it easier to analyse the regression plot. From the residual plot we can see that it is somewhat unbalanced residuals. The points under 100 m2 are below the line, in the middle most of the points are above the line and above around 160 m2 the points once again are under the line.

From this we can conclude that a linear regression might not be the best strategy in order to predict housing prices based on living area. Perhaps a non-linear regression would have been better to use to account for the lower priced small and big houses while also accounting for a higher price for the houses in the middle.

Since the dataset was not huge (only 55 datapoints) the outliers might just be noise. Using a nonlinear regression might be over fitting the data. The first step towards improving the prediction would be to add more data points.

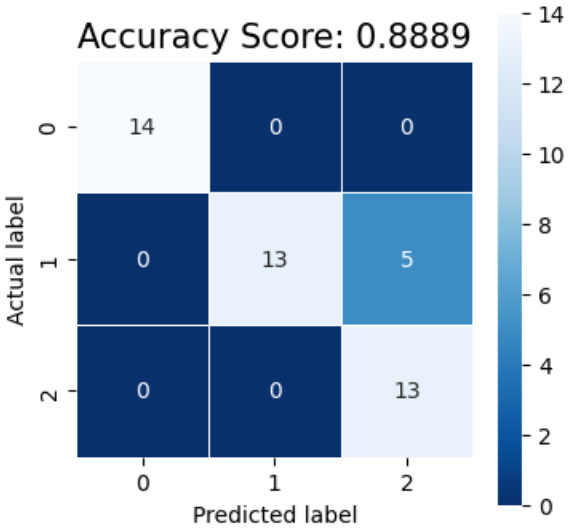
Here is a non-linear regression which somewhat fits the data better, but still not perfect. As said previously more data is needed in order to conclude which is noise and which is trends.



Evaluation

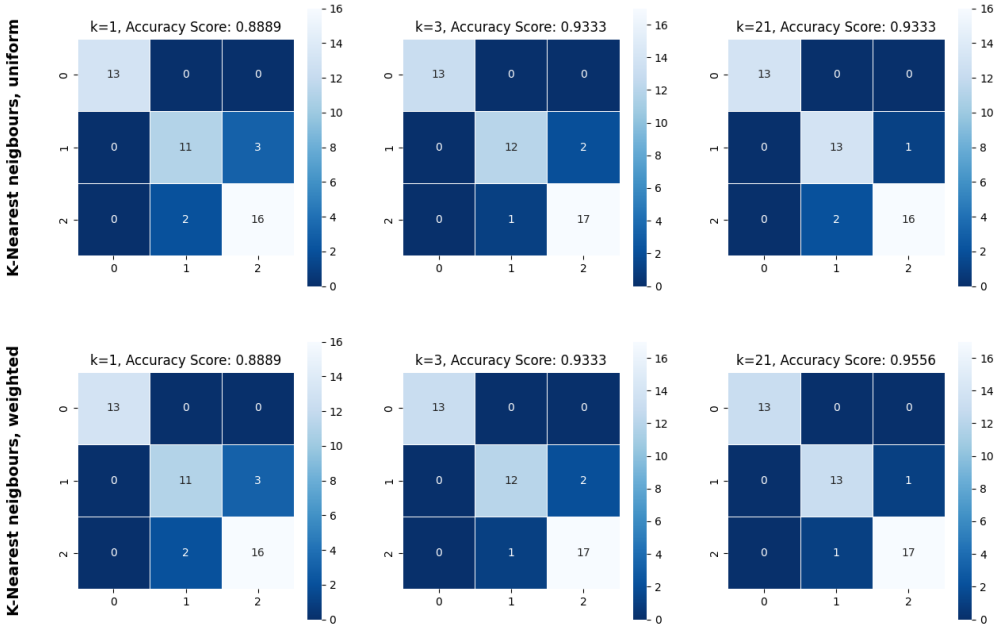
1. Use a confusion matrix to evaluate the use of logistic regression to classify the Iris data set.

The use of logistic regression gives quite a good prediction but we can see that 5 Irises gets wrongly classified. It gives us a score of 88.89%.



2. Use k-nearest neighbors to classify the Iris data set with some different values for k, and with uniform and distance-based weights. What will happen when k grows larger for the different cases? Why?

For this we used k = 1, 3 and 21 respectively.



When we increase the k-value, more data points gets taken into account when labeling the data. It will have smoothing effect where the boundaries will be smoother and can help with reducing over fitting and the impact of noise. Too high of a k-value can however lead to under fitting the data and thus smoothing the decision boundaries too greatly. The effect of under fitting the data is reduced when using the distance based weighting since the closest data points is still weighted higher.

This can be seen as the higher k-values when using distance based weights leads to an increase in accuracy score. When using uniformly weights the accuracy score does not increase from k=3 to k=21.

3. Compare the classification models for the Iris data set that are generated by k-nearest neighbors (for the different settings from question 2) and by logistic regression. Calculate confusion matrices for these models and discuss the performance of the various models.

The confusion matrices can be seen above. Using k-nearest neighbour gives a better accuracy score than the use of logistic regression. All k-values produces the same or a better accuracy score. Thus it could be said that for this data set the use of k-nearest neighbour might be better to use.

However, this might be only due to noise in the data. The difference is not that radical between the two different types of classification and thus nothing can be said for certain. Perhaps a bigger data-set would be needed to draw further conclusions.

References

[1] Hemnet Dataset. Retrieved October 18, 2020, from `https://www.hemnet.se/`

[2] Iris Dataset. Retrieved September 03, 2023, from `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html`