# HOUSE PRICE PREDICTION

Meiramov Daniyar, Aipenova Aziza
Suleyman Demirel University meiramovdanik@gmail.com

Suleyman Demirel University *PhD Computer Science,*
aziza.aipenova@sdu.edu.kz

## INTRODUCTION

In statistics, the important thing is to measure, determine and predict values using existing data. In this thesis, using SAS tool, we will first recognize correlation analysis, precisely for what it is used, how can be calculated and see it on practice. We will touch on regression and how to predict price using other parameters, also take a look on multiple regression on practice.

The data we used is 100 observations from krisha.kz with 6 variables as: price, square meter, number of bedrooms, number of bathrooms, type of house and location.

## CORRELATION ANALYSIS

Correlation is a measure of relationship between two variables. Correlation analysis is very useful for researching how much those two variables dependent to each other. For instance, an analyst have to find a relationship between the price of houses and their number of bedrooms. It can be expected that high price of house is directly related to large number of bedrooms. Such relationship shows strong positive correlation between price and number of bedrooms.

It is also possible to search negative correlation. For instance, analyst have to find relationship between size of house (square meter) and its location.

In this case, it can be expected that there is no relationship between this two variables, and most likely it is strong or weak negative correlation depending on data.

Correlation analysis formula is represented as following:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Where,

| Obs | Price | Square Feet | Number of bedrooms | Number of bathrooms | Brick/Non-Brick | Location |
|---|---|---|---|---|---|---|
| 1 | 11,500,000 | 40.00 | 1 | 1 | Brick | Nauryzbai batyr region |
| 2 | 217,637,000 | 310.91 | 4 | 3 | Non-Brick | Turksib region |
| 3 | 147,560,000 | 210.80 | 3 | 2 | Non-Brick | Zhetisu region |
| 4 | 57,500,000 | 90.00 | 2 | 1 | Brick | Bostandyk region |
| 5 | 20,100,000 | 40.00 | 1 | 1 | Brick | Zhetisu region |
| 6 | 131,600,000 | 188.00 | 3 | 2 | Non-Brick | Almaly region |
| 7 | 65,100,000 | 93.00 | 2 | 1 | Brick | Alatau region |
| 8 | 81,900,000 | 117.00 | 2 | 1 | Brick | Zhetisu region |
| 9 | 83,500,000 | 126.00 | 2 | 1 | Brick | Almaly region |
| 10 | 96,600,000 | 138.00 | 3 | 2 | Brick | Auezov region |
| 11 | 12,000,000 | 65.00 | 1 | 1 | Brick | Medeu region |
| 89 | 20,500,000 | 90.00 | 2 | 1 | Non-Brick | Bostandyk region |
| 90 | 33,920,000 | 43.00 | 1 | 1 | Brick | Bostandyk region |
| 91 | 23,350,000 | 80.00 | 2 | 1 | Brick | Bostandyk region |
| 92 | 17,000,000 | 64.00 | 1 | 1 | Brick | Medeu region |
| 93 | 35,000,000 | 63.00 | 1 | 1 | Non-Brick | Turksib region |
| 94 | 32,500,000 | 87.00 | 2 | 1 | Brick | Zhetisu region |
| 95 | 22,500,000 | 120.00 | 3 | 2 | Non-Brick | Auezov region |
| 96 | 40,000,000 | 70.00 | 1 | 1 | Brick | Medeu region |
| 97 | 90,600,000 | 107.00 | 2 | 1 | Non-Brick | Zhetisu region |
| 98 | 18,800,000 | 130.00 | 3 | 1 | Non-Brick | Nauryzbai batyr region |
| 99 | 31,000,000 | 51.00 | 1 | 1 | Brick | Alatau region |
| 100 | 17,000,000 | 80.00 | 2 | 1 | Brick | Alatau region |
| | 3,945,609,749 | 7956.73 | 162 | 114 | 121 | 441 |

**Correlations Scatter Plots with Price**

The CORR Procedure

| | |
|---|---|
| 1 With Variables: | price |
| 5 Variables: | square number_of_bedrooms number_of_bathrooms type location |

**Simple Statistics**

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum | Label |
|---|---|---|---|---|---|---|---|
| price | 100 | 39456097 | 63576549 | 3945609749 | 3000000 | 516800000 | Price |
| square | 100 | 79.56730 | 59.87958 | 7957 | 12.00000 | 420.00000 | Square Feet |
| number_of_bedrooms | 100 | 1.62000 | 0.88512 | 162.00000 | 1.00000 | 5.00000 | Number of bedrooms |
| number_of_bathrooms | 100 | 1.14000 | 0.44992 | 114.00000 | 1.00000 | 4.00000 | Number of bathrooms |
| type | 100 | 1.21000 | 0.40936 | 121.00000 | 1.00000 | 2.00000 | Brick/Non-Brick |
| location | 100 | 4.41000 | 2.19317 | 441.00000 | 1.00000 | 8.00000 | Location |

**Pearson Correlation Coefficients, N = 100**
**Prob > |r| under H0: Rho=0**

| | square | number_of_bedrooms | number_of_bathrooms | type | location |
|---|---|---|---|---|---|
| price<br>Price | 0.84596<br><.0001 | 0.64919<br><.0001 | 0.82546<br><.0001 | 0.30616<br>0.0019 | -0.05488<br>0.5876 |

X – the value of first variable

–

X – the mean value of first variable, sum of all value divided by amount of values(n)

Y – the value of second variable

–

Y – the mean value of second variable

Lets take a real data and see correlation of price with all variables on SAS:

In the first table above we see which variable is used as x and others as y.

The second table give us already calculated information for all variables, including mean and standard deviation.

The last table shows correlation between price and all values.

The CORR Procedure

| 6 Variables: | price square number_of_bedrooms number_of_bathrooms type location |
|---|---|

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

| | price | square | number_of_bedrooms | number_of_bathrooms | type | location |
|---|---|---|---|---|---|---|
| **price** Price | 1.00000 | 0.84596 <.0001 | 0.64919 <.0001 | 0.82546 <.0001 | 0.30616 0.0019 | -0.05488 0.5876 |
| **square** Square Feet | 0.84596 <.0001 | 1.00000 | 0.88402 <.0001 | 0.82927 <.0001 | 0.50694 <.0001 | -0.08458 0.4028 |
| **number_of_bedrooms** Number of bedrooms | 0.64919 <.0001 | 0.88402 <.0001 | 1.00000 | 0.69297 <.0001 | 0.44549 <.0001 | -0.11666 0.2477 |
| **number_of_bathrooms** Number of bathrooms | 0.82546 <.0001 | 0.82927 <.0001 | 0.69297 <.0001 | 1.00000 | 0.38720 <.0001 | -0.12018 0.2337 |
| **type** Brick/Non-Brick | 0.30616 0.0019 | 0.50694 <.0001 | 0.44549 <.0001 | 0.38720 <.0001 | 1.00000 | -0.06312 0.5327 |
| **location** Location | -0.05488 0.5876 | -0.08458 0.4028 | -0.11666 0.2477 | -0.12018 0.2337 | -0.06312 0.5327 | 1.00000 |

This table represents all correlations between variables.

We can see that price has **strong positive correlation** with square, number of bedrooms and number of bathrooms (noticed by green) because it has more than 0.5 correlation coefficient. On the other hand, price and type of house have **weak positive** 0.30616 (yellow), and price with location **weak negative correlation** -0.05488 (red).

## REGRESSION ANALYSIS

Regression is a technique that used for define and predict values of one or more variables. It is useful when analyst using one value have to determine another one through dependence of those two variables. For instance, analyst have to find possible price of house where the size of house should be 180 square meter. In order to predict it, the only and best way is to use *simple regression analysis*.

It is also possible to predict price of house using more than one input values, such as number of bedrooms, type of house, location and size. Such more difficult operation called *multiple regression analysis*.

Regression analysis formula is represented as following:

1. $Y = a + bX$ - simple regression
2. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$ - multiple regression

Where,

Y – the dependent variable, in our case price

a – the intercept, also can be represented as b0, has value of Y when X = 0

b – angular coefficient or gradient of the estimated line; it is the value by which Y increases on average, if we increase x by one unit.

X – the independent variable, in our case square meter

X1, X2, Xn – used for multiple regression and in our case it can be square meter, number of bedrooms, number of bathrooms, type, location, etc.

The *a* and *b* can be found by following formulas:

$$A = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$B = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Lets take a look a real example of multiple regression using data from web source:

As we may notice, SAS already calculated all equations for us☺.

| | | | |
|---|---|---|---|
| Root MSE | 28128284 | R-Square | 0.8141 |
| Dependent Mean | 39456097 | Adj R-Sq | 0.8043 |
| Coeff Var | 71.29008 | | |

**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -26033055 | 14945011 | -1.74 | 0.0848 |
| square | Square Feet | 1 | 1053328 | 137018 | 7.69 | <.0001 |
| number_of_bedrooms | Number of bedrooms | 1 | -27978541 | 6954948 | -4.02 | 0.0001 |
| number_of_bathrooms | Number of bathrooms | 1 | 47047648 | 11473428 | 4.10 | <.0001 |
| type | Brick/Non-Brick | 1 | -23492713 | 8036057 | -2.92 | 0.0043 |
| location | Location | 1 | 407214 | 1306837 | 0.31 | 0.7560 |

R-Squared (blue) – to measure accuracy from 1 to 100%, it tells us how much of the difference in outcome is explained by the model. The good regression should be not less than 70% of R-Squared. In our case is total 81%.

Intercept – is our a, and it is -26033055.

The slope of the line – or also it is b1, b2, b3, b4, b5, b6.

**predicted price**

| Obs | square | number_of_bedrooms | number_of_bathrooms | type | location | Price |
|---|---|---|---|---|---|---|
| 1 | 180 | 1 | 1 | 2 | 5 | 137,623,454.86 |
| 2 | 65 | 4 | 2 | 2 | 8 | -66,223,208.63 |
| 3 | 134 | 3 | 1 | 1 | 7 | 34,027,727.75 |

Here can be seen input values X1, X2, X3, X4, X5, X6 (green) and predicted value Y (red).

Equation proven by us represented as follows:

Y = -26033055 + (1053328*180) + (-27978541*1) + (47047648*1) +

+ (-23492713*2) + (407214*5) = 137,685,736

137,623,454.86 – the predicted price for specific input values, calculated by SAS.

137,685,736 – the predicted price calculated by myself.

62,281.14 – some measurement error during analytics.

## CONCLUSION

In this thesis, we conclude that using correlation and regression analysis, we can analyze relationship between values, are they strong or weak, positive or negative, or there is no relation at all. Correlation analysis is necessary to use if it is difficult to expect relationship of two variables and for searching the most precise information of relationship in numerical form, from -1 to 1. It is useful statistical formula that measures the strength between two continuous variables. Regression analyses can help business analysts build models to predict trends, make decisions, and model the real world for decision-making support. These models can be used to predict the value of one or more variables from a knowledge of the value of other variables using real data set.

During the analysis, we saw on real data set, how the correlation and multiple regression were calculated, using 100 observations examine relationship of their values. Find out why price of house and its square meter size has strong positive dependence and type with location has vice versa; also understand how SAS statistical tool facilitates the work for data analysts.