IDS532-Introduction to Operations Management

Submitted By

Dilip Pandey

Aruna Venkatasubramaniam

Satya Devi Julakanti

Prashant Chaudhary

**Forecasting of Home Sales Price for**

**IOWA**

2018

TABLE OF CONTENTS

Introduction………………………………………………………………………………3

Problem Definition………………….…………………………………………………4

Objective………………………………….……………………………………………….4

Assumption……………………………….………………………………………………4

Forecasting of Sales Price……………….………………………………………….4

Forecasting Methods………………………………………………………………….6

Regression Analysis…………………………………………………………………….6

Univariate Analysis…..…………………………………………………………………7

Multivariate Analysis…….….…………………………………………………………9

Final Regression Model……..………………………………..…………………….11

Conclusion………………..……………………………………………….……………….13

Scope of the Project……………………………………………………………………13

References…………………………………………………………………………………...15

**INTRODUCTION**

Understanding the customer needs and predicting customer’s purchase intents form the core of success for any business. In today’s market, with the price of properties increasing, it becomes very essential to study the factors, which affect the price directly or indirectly when a customer decides to buy a house and predict the market trend. In general, for any purchase, a potential customer makes the decision based on the value for their money.

This project proves that apart from total bedrooms or bathrooms, there are number of various factors that influences the price of a property.

We chose this specific problem because it provided us with the opportunity to work on real life problems like prediction of prices for houses in Ames, IOWA.

Also, we wanted to study regression techniques and their implementation. Understanding the theory and implementing it into practice was a challenge for us

**PROBLEM DEFINITION**

The original Ames Home Sales data set contains information about the sale of individual residential property in Ames, Iowa, from 2006 to 2010 including 2,930 observations and 98 explanatory variables involved in assessing home values. In addition, some summary variables including the natural log of the sale price of the homes is created. A subset of the full data set, for homes with normal sales conditions (to avoid analyzing foreclosure or distressed sales) and gross living area of 1,500 square feet or less (to focus on homes of modest size) is obtained. Then, a dataset with 1300+ observations is sampled. We are trying to forecast the sales price of the house for the year 2011.

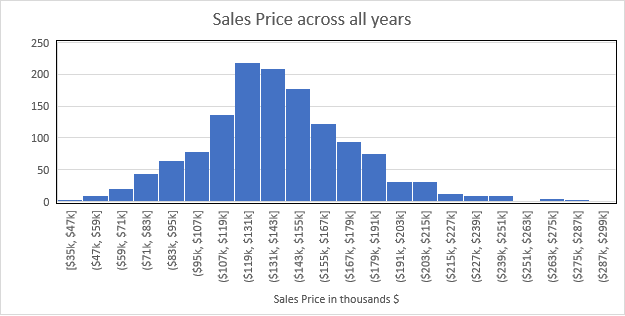
**OBJECTIVE**

Our objective is to know what factors are contributing significantly to the selling price of the houses and come up with a robust model taking all factors into consideration, to predict home prices in the region

**ASSUMPTION**

All statistical methods rely on initial assumption of data being normally distributed and the relationship is indeed linear. Hence before applying any statistical method, it should be confirmed whether the data is normally distributed or not.

Histogram and scatted plots were plotted to determine the normal distribution and linear relationship



**FORECASTING OF SALES PRICE**

What is forecasting?

**Forecasting** is a process of making predictions of the future based on past and present data and most commonly by analysis of trends.

**FORECASTING METHODS**

To predict the sales price of the houses in IOWA, for future years, we used the 2year moving average method

We took the average sales price of homes for the year 2006-2007 from our dataset and forecasted the average selling price of the homes for 2008 to 2010 and compared with the actual price of the homes.

Below is the graph plotted for actual and the forecasted price of the homes for years 2008-1010

The actual sales price with the forecasted sales price is to show how forecasted values match with the actual values.

Graph 1

The main drawback of this technique is it doesn’t consider other factors into account which might affect the sales price of a property.

Therefore, we used regression technique to build a model which would take all factors into consideration and the predict the sales price more accurately.

**REGRESSION ANALYSIS**

In [statistical modeling](https://en.wikipedia.org/wiki/Statistical_model), regression analysis is a set of statistical processes for [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) and one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable) (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

In our case, we are predicting sales price of the houses and factors that would influence the sales price such as age, living area, garage area etc.

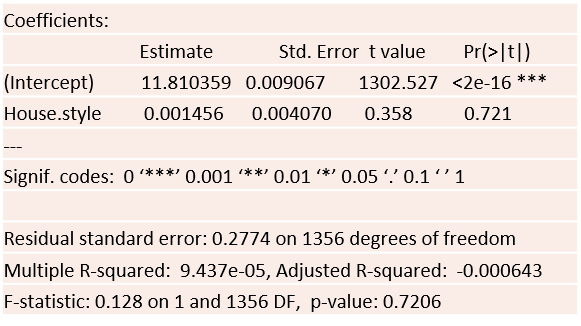
We will start with our univariate analysis and find out the statistical significance, which means how the price gets affected when we take only 1 factor into consideration

**UNIVARIATE ANALYSIS**

We have taken a risk factor of 5% that is there is a 5% chance of making a mistake if one factor contributes to the sales price.

1. Log price vs House Style

Below is the output after running the regression technique to determine if house style is a significant factor in influencing the sales price.



The p-value as we can see in the output is 0.7206

Since it is greater than our risk factor of 0.05, therefore we conclude that the house style is not a significant factor when it comes to determining the sales price of the houses in IOWA

Similarly, the significance is checked for other variables.

Below is the table to show the variables and their significance to the model.

|  |  |
| --- | --- |
| **Variables** | **Significance** |
| House Style | No |
| Central Air | Yes |
| No of Bedrooms | Yes (negative slope) |
| Fire Place | Yes |
| Living area | Yes |
| Garage Area | Yes |
| Basement Area | Yes |
| Total Bathrooms | Yes |
| Overall Quality | Yes |
| Lot Area | Yes |
| Age when sold | Yes (negative slope) |
| Year sold | No |

Although the year when the houses were sold didn’t come out as a significant factor in determining the selling price, but it is practically one of the most important factors to get an idea on sales price for a particular year keeping all other factors as constant

**MULTIVARIATE ANALYSIS**

After determining all the significant factors which influenced the sales price of the house individually, we built a model by taking all those factors in combination to predict the sales price

|  |
| --- |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 1.066e+01 2.964e-02 359.652 < 2e-16 \*\*\* |
| Central.Air 1.511e-01 1.325e-02 11.406 < 2e-16 \*\*\* |
| Bedrooms -6.071e-03 5.712e-03 -1.063 0.288 |
| Fireplaces 2.773e-02 6.246e-03 4.439 9.78e-06 \*\*\* |
| Lot\_Area 1.097e-05 9.328e-07 11.760 < 2e-16 \*\*\* |
| Gr\_Liv\_Area 2.636e-04 1.995e-05 13.213 < 2e-16 \*\*\* |
| Garage\_Area 1.941e-04 2.090e-05 9.287 < 2e-16 \*\*\* |
| Basement\_Area 1.422e-04 1.206e-05 11.788 < 2e-16 \*\*\* |
| Total\_Bathroom 3.893e-02 6.324e-03 6.157 9.79e-10 \*\*\* |
| Overall\_Qual 7.986e-02 3.972e-03 20.103 < 2e-16 \*\*\* |
| Year.sold 1.065e-03 2.446e-03 0.435 0.663 |
| Age\_Sold -2.090e-03 1.606e-04 -13.018 < 2e-16 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 0.1176 on 1346 degrees of freedom |
| Multiple R-squared: 0.8216, Adjusted R-squared: 0.8202 |
| F-statistic: 563.7 on 11 and 1346 DF, p-value: < 2.2e-16 |

In the above output we see, as the number of bedrooms increases, the price decreases, which is not practical, and the p value is also greater than 0.05 Therefore it is not a significant factor when all the other factors are taken into account. So, we need to build a new model to predict the price accurately

**FINAL REGRESSION MODEL**

|  |
| --- |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 1.065e+01 2.885e-02 369.284 < 2e-16 \*\*\* |
| Central.Air 1.491e-01 1.312e-02 11.369 < 2e-16 \*\*\* |
| Fireplaces 2.888e-02 6.151e-03 4.696 2.93e-06 \*\*\* |
| Lot\_Area 1.082e-05 9.225e-07 11.731 < 2e-16 \*\*\* |
| Gr\_Liv\_Area 2.540e-04 1.779e-05 14.278 < 2e-16 \*\*\* |
| Garage\_Area 1.968e-04 2.074e-05 9.492 < 2e-16 \*\*\* |
| Basement\_Area 1.417e-04 1.205e-05 11.755 < 2e-16 \*\*\* |
| Total\_Bathroom 3.959e-02 6.294e-03 6.291 4.27e-10 \*\*\* |
| Overall\_Qual 8.050e-02 3.927e-03 20.500 < 2e-16 \*\*\* |
| Year.sold 1.058e-03 2.446e-03 0.432 0.665 |
| Age\_Sold -2.089e-03 1.606e-04 -13.009 < 2e-16 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 0.1176 on 1347 degrees of freedom |
| Multiple R-squared: 0.8215, Adjusted R-squared: 0.8202 |
| F-statistic: 619.9 on 10 and 1347 DF, p-value: < 2.2e-16 |

Looking at the final regression model we derived an equation to predict the sales price of the houses in Ames, IOWA

The equation is :

Log(Sales Price) = 10.65 + 0.001058\* YearSold

- 0.002089 \* AgeSold

+0.1491 \* CentralAir

+0.02888 \* fireplace

+0.00001082 \* LotArea

+0.0002540 \* LivingArea

+0.0001968 \* GarageArea

+0.0001417 \* BasementArea

+0.03959 \* Bathrooms

+0.0805 \* OverallQuality

**CONCLUSION**

* In the final regression model we see that the adjusted R-squared value is 0.82 which means 82% of total variability in price is explained by all these variables
* We have used the equation above to calculate the sales price for all the years. The MPSE came to be 0.69%.
* The MPSE now is lower than that of the moving average method (0.7% as compared to 0.89%), indicating that **Regression is the more accurate method** to predict home sales of IOWA.

**SCOPE OF THE PROJECT**

On performing the forecasting and the regression technique to predict the sales price of houses in IOWA, we came to a conclusion that in future we can also include other variables like income, no of children, age of the family group and other regions in Iowa to explain variability in sales price and to predict price more efficiently.

**REFERENCES**

<https://cran.r-project.org/doc/contrib/Short-refcard.pdf>

<https://swcarpentry.github.io/r-novice-inflammation/reference>

<https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique>

<https://en.wikipedia.org/wiki/Forecasting>