Regression Analysis on Sale Price and Construction Cost of Iran Residential Buildings

Team 5: Xing Hong, Yiyi Xu, Yuru Zhou, Shuyang Zhang

Background

We collected our data from the UCI Machine Learning Repository.[1] The data set consists of 372 samples and 105 variables corresponding to real estate single-family residential apartments in Tehran, Iran. The variables contain the essential indicators of housing price including 8 physical and financial variables, 19 economic variables from 5 time-lags, and two output variables: actual construction cost and sale price. This study aims to investigate that whether the sale price and actual construction cost can be accurately predicted based on the chosen physical and economic factors:

- Total floor area of the building [v2]
- Price of the unit at the beginning of the project per m^2 [v8]
- CPI of housing, water, fuel & power in the base hear(economic) [v26]
- Population of the city(economic) [v28])

as well as their combination. We have implemented a brief correlation check for the datasets we choose in the first place. The result shows that the datasets v2, v8, v26 and v28 are acceptably correlated to the sale price and the actual construction cost.

Methods

With the cleaned data set, we first performed descriptive data analysis to identify the general pattern of data points across dimensions. Then we calculated and visualized the Pearson Correlation Matrix to examine the correlation between the variables as well as how the variables correlate with our target outcome. Next, we utilized cross-validation method to split the data into training data (80%) and testing data(20%) for the preparation of univariate and multivariate linear regression. We first trained the univariate linear regression model for each of chosen variable with training data. We then use the trained linear model on testing data to see whether they can well predict the actual sale price and actual construction cost respectively. Mean squared error (MSE), a criteria we use to evaluate whether the model has a good prediction without the issue of overfitting, is calculated for each of the univariate linear regression models. In addition, we reported the training MSE and testing MSE with a bar-plot on Price and Cost for the purpose of comparing the error of prediction across all models. Overfitting is suspected when there is a remarkable difference between training MSE and testing MSE, in particular the model accuracy was high with the training data and dropped significantly with respect to the testing data.

Subsequently, we performed a multivariate linear regression model to measure how multiple variables interact together to influence the prediction. We imported the sklearn linear model package to help us construct multivariate linear regression models. We built models for 2 variables, 3 variables, and 4 variables and computed Mean Squared Error (MSE), R-squared, Adjusted R-squared, k-Fold Cross Validation score.

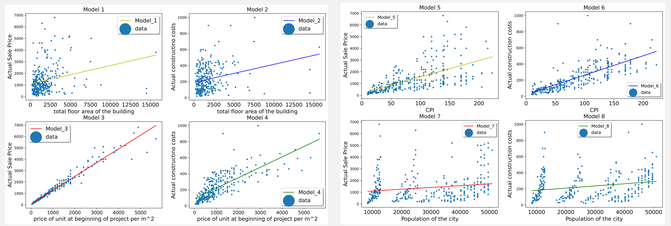
Furthermore, the main idea of linear models is to fit a straight line to our data, considering that the data might have a quadratic distribution, we should choose a quadratic function and applying a polynomial transformation may give us better results as a supplement to our experiment. Therefore, we conducted Polynomial regression as an additional analysis to see whether it has higher accuracy than univariate and multivariate linear regression models.

Results

Univariate Regression Models:

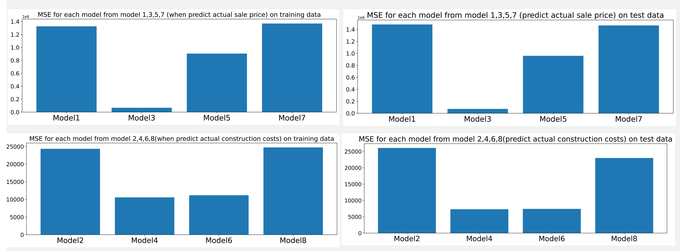
1. Model1: Actual Sale Price = 1043.548416 + 0.212883 * V2
2. Model2: Actual construction costs = 183.782995 + 0.027523 * V2
3. Model3: Actual Sale Price = 100.046507 + 1.186078 * V8
4. Model4: Actual construction costs = 94.975321 + 0.123372 * V8
5. Model5: Actual Sale Price = 126.129641 + 14.770122 * V26
6. Model6: Actual construction costs = 16.997770 + 2.462560 * V26
7. Model7: Actual Sale Price = 833.073955 + 0.020715 * V28
8. Model8: Actual construction costs = 138.757408 + 0.003315 * V28

Graphs of Model 1 – 8: As we can see from the graph, only Model 3, Model 4 and Model 6 have relatively good linear relationship.



MSE and Cross-Validation of Univariate Regression:

The MSE graphs showed that M3 is a good model for the MSE values of both training and testing datasets are the smallest.



Multivariate Linear Regression:

Models on Sale Price:

1.MM1_P: Actual Sale Price = 504.2266053 + 0.2065703* V2 + 0.0196603* V28

2.MM2_P: Actual Sale Price = 123.5540659 + 1.2000097* V8 + -0.4468449 * V26

3.MM3_P: Actual Sale Price = 92.2894127 + 1.1989932* V8 + -0.5591759 * V26 + 0.0015069 * V28

4.MM4_P: Actual Sale Price = 46.6237633 + 0.0196092* V2 + 1.1731741 * V8 + 0.0012035 * V28

5.MM5_P: Actual Sale Price = 66.4102942 + 0.0190059* V2 + 1.1886041 * V8 + -0.5175850 * V26 + 0.0015348 * V28

	Model	Details	Mean Squared Error (MSE)	R-squared (training)	Adjusted R-squared (training)	R-squared (test)	Adjusted R-squared (test)	5-Fold Cross Validation
0	Multiple Regression-1P	2 features on Price	1347872.830	0.136	0.133	-0.161	-0.177	0.112
1	Multiple Regression-2P	2 features on Price	69582.114	0.955	0.955	0.943	0.942	0.950
2	Multiple Regression-3P	3 features on Price	69187.614	0.956	0.955	0.941	0.939	0.950
3	Multiple Regression-4P	3 features on Price	68574.098	0.956	0.956	0.941	0.940	0.950
4	Multiple Regression-5P	All features on Price	68173.398	0.956	0.956	0.942	0.939	0.950

Models on Construction Cost:

1.MM1_C: Actual Construction Costs = 16.8323763 + 0.0770600 * V8 + 1.4853850 * V26

2.MM2_C: Actual Construction Costs = 4.6706821 + 2.4131911 *V26 + 0.0005941* V28

3.MM3_C: Actual Construction Costs = 8.4620408 + 0.0767878 * V8 + 1.4553112 * V26 + 1.4553112 * V28

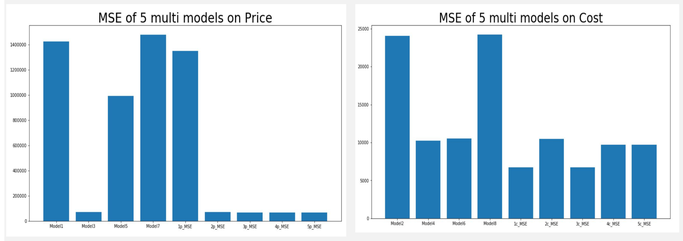
4.MM4_C: Actual Construction Costs = 51.53900332 + 0.0081044 * V2 + 0.1154430 * V8 + 0.1154430 * V28

5.MM5_C: Actual Construction Costs = -4.91746260 + 0.0098260 * V2 + 0.0714167 * V8 + 1.4768137 * V26 + 0.0004179 * V28

	Model	Details	Mean Squared Error (MSE)	R-squared (training)	Adjusted R-squared (training)	R-squared (test)	Adjusted R-squared (test)	5-Fold Cross Validation
0	Multiple Regression-1C	2 features on Cost	6748.842	0.744	0.743	0.779	0.776	0.743
1	Multiple Regression-2C	2 features on Cost	10491.562	0.602	0.600	0.626	0.620	0.612
2	Multiple Regression-3C	selected features	6720.566	0.745	0.743	0.781	0.775	0.742
3	Multiple Regression-4C	3 features on Price	9711.646	0.631	0.629	0.675	0.666	0.596
4	Multiple Regression-5C	All features on Price	9711.646	0.755	0.754	0.785	0.779	0.745

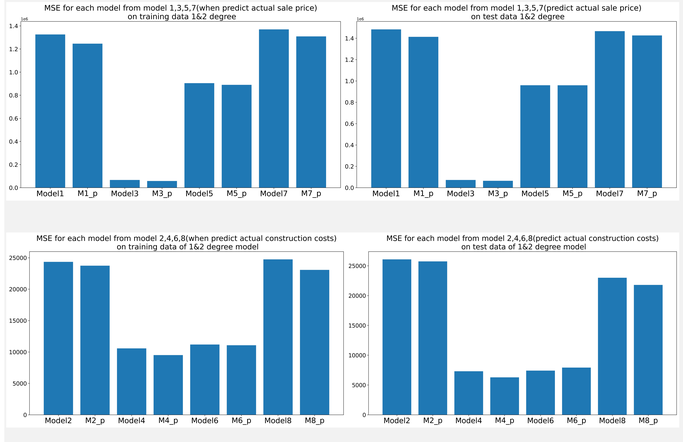
Comparison between Univariate Regression and Multivariate Regression:

Multivariate regression Models have much lower MSE values than the univariate models. They are better choices.



Polynomial Regression of One Single Variable VS. Univariate Linear Regression:

The graphs showed that Polynomial Regression of a single variable have a similar pattern as the univariate linear regression across the variables, since the MSE values of one variable in two models are very close to each other.



Discussion & Conclusion

The models the research adopted here well answered the research questions. Price of the unit at the beginning of the project per m^2 affected the sale price and actual construction price the most. Model 3 and Model 4 (price of Unit at Beginning of Project per m ^2(V8) VS. sale price, and V8 VS. construction cost) had very low MSE value in both testing and training sets. Moreover, the combination of the variables showed stronger linear relationship with the price and construction cost than single variables. Although some single variables, such as the Price of Unit at Beginning of Project per m ^2(V8), could well predict the sale price alone, not all studied variables in the dataframe were able to predict the price and cost by themselves using regression analysis.

The further result concluded from the experiment was that the multivariate regression was more reliable when predicting both housing sale price and construction cost in general, even though some univariate model(V8) showed high reliability in predicting the price.

MSE bar graphs on both price and cost demonstrated that: compared to the univariate linear regression models, the multivariate linear regression models on price and cost were significantly better than the models using single variables, except for the model 3, which was almost as good as the multivariate ones.

The polynomial regression on one variable was adopted here to extensively explore the relationship between the independent variable and the dependent one. The resulting MSE graphs of polynomial regression of degree 2 showed a very similar pattern as the univariate regression model of degree one.

For further improvement, the students could try to explore the best multivariate regression model they could get for the actual sale price and construction cost. Another possible addition could be to consider the time change of cpi and population change during the construction of housing, which required more complex models, since the original dataset had data for five different time ran

References

1. http://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set