Chapter Project

Home Sweet Home: Using Multiple Regression to Analyze and Predict Home Prices

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Mount Pleasant Real Estate Data.**

An important problem in real estate is determining how to price homes to be sold. There are so many factors—size, age, and style of the home; number of bedrooms and bathrooms; size of the lot; and so on—which makes setting a price a challenging task. In this project, we will try to help realtors in this task by determining how different characteristics of homes relate to home prices, identifying the key variables in pricing, and building multiple-variable regression models to predict prices based on property characteristics. Our analysis will be based on the Mount Pleasant Real Estate Data (available on stat.hawkeslearning.com). This data set includes information about 245 properties for sale in three communities in the suburban town of Mount Pleasant, South Carolina, in 2017.

Phase 1: Data Preparation.

1. Download the Mount Pleasant Real Estate Data from stat.hawkeslearning.com and open it with Microsoft Excel.

2. To ensure the data contains comparable properties, eliminate duplexes and properties whose prices are outliers. What limitations does this impose on our analysis? Consider the following variables associated with each property.

x_1 = number of bedrooms	x_5 = has pool?
x_2 = number of bathrooms	x_6 = has dock?
x_3 = number of stories	x_7 = fenced yard?
x_4 = square footage	x_8 = golf course?

3. Are any of the variables qualitative? Adjust this data in a reasonable quantitative way for use in a regression analysis.

Phase 2: Constructing Predictive Models

Enable the Analysis ToolPak add-in to Excel. The regression tool will be used.

The idea of linear regression can be easily extended to the case where there are multiple independent variables that are used to predict the dependent variable. The linear regression equation will look like

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m$$

where each b_i is a coefficient and each x_i is an independent variable. In the context of real estate pricing, $\hat{y}$ = predicted home price. Excel can calculate regression models with multiple variables via the same regression tool that it does for single-variable regression models by simply using more columns of data for the X inputs. Intuitively, this should be more realistic for real estate pricing as there may be several variables that contribute to property values.

4. Construct the multiple regression equation with input variables x_1, x_2, ..., x_8.

5. What is the adjusted coefficient of determination, R_a^2, of the regression model? Explain the meaning of this value and how it differs from R^2.

6. Perform a hypothesis test to determine if the model is useful for predicting home values at a significance level of $\alpha = 0.05$. State the *P*-value and interpret its meaning.

7. Are any variables not useful predictors of home price at a significance level of $\alpha = 0.05$? State the *P*-values of these variables. Intuitively, what does this mean with respect to pricing properties?

8. Construct the multiple regression model with only the input variables whose coefficients are significant in the eight-variable regression?

9. How does the adjusted coefficient of determination of the new model in Problem 8 relate to the adjusted coefficient of determination from Problem 5? What conclusion can you draw from this?

Phase 3: Applying and Interpreting the Model

10. Suppose you own a 2000 ft^2 2-story house in one of the communities in the data set with 3 bedrooms, 2.5 baths, a pool, and it is located on a golf course, but has no dock or fenced yard. What does the model from Problem 4 predict the price of your house to be?

11. A common term in real estate is "comparables," or "comps" for short, which are properties that have similar characteristics. It is common for realtors to look up "comps" for a certain property to get an idea of how to price it. Locate the "comps" for your home in the data set. Create a box plot of the "comps" and estimate a price range for your house on this basis.

12. What advantages and disadvantages does this approach have to the multiple regression model above?