**Fishers Multilinear Predictive Model**

Nathan Rollins

**Abstract**

Our goal has been the creation of a multilinear regression model that predicts at least somewhat accurately the sale price of a housing unit in the Fishers Indiana current real estate market. This project undertook three large processes: first, the collection of Fishers real estate data; second, the creation of a multilinear model for this data; and third, the evaluation and refinement of this model. The result of this work was mostly successful, with a new and key dataset being made and, despite its quirky faults, a fairly accurate multilinear model.

**1. Introduction**

A multilinear model for the Fishers area does not exist, nor does there exist a dataset like it. There have been many different, classic regression models proposed for real estate, all using interesting and varying datasets, but none tackle this problem with both the focus and type of data as our model required. Typically, these regression models focus on very large areas, (perhaps so that their specific models are unaffected by local market trends that are difficult to detect). This is not to say that models do not exist for localized areas, but for these their focus is different from ours in various ways. The closest we could find was a paper in analyzing the real estate prices of King County in the state of Washington (Agarwal) which was very instructive but was too different from a Fishers dataset since Fishers is mostly homogenous in housing types and is very rapidly developing. These deficiencies are what drove us to collect hundreds of samples from the Fishers area, (even working with specific realtors), in order that we could create this specialized model for the Fishers area.

**2. Method**

After the appropriate data was collected, (and it was very challenging as most realtors and real estate companies try to limit the proliferation of their data), we set about creating our model. A linear regressive model is one which models a real relationship between some predictor (called x here) and a resulting, predicted value (called y). In other words, the equation y = mx + b, (again, where y is our predicted value, x is our predictor value, m is the slope or relationship between x and y, and b is the intercept of the line), is the model.

This would be enough for a simple model where the relationship between the predicting value and predictor value was also simple. However, in order for our real estate model to work we must identify multiple predictor values, their slopes, and combine them into a single, multilinear regression model. Thus the model we used to predict the sale price, (our y), of a Fishers home is: y = b + ∑m1x1 + m2x2 + … mnxn where n is the number of predictor values. The specific predictor values used in the finished model were: square footage, number of bedrooms, number of bathrooms, the ownership of a garage, the ownership of a basement, and the year the unit was built.

To find the slope values for each x value, (store in **B** and **X**, respectively), we first centralized our variables by the equations ynew = yold – the mean of yold, and xnew = xold – the mean of x. The results were stored in and thus redefined matrices **Y** and **X**, (all done in the MATLAB application). This centralization was done to remove the effect of the slope-intercept. Technically, MATLAB’s regress function was used to find the slopes of each value, but the expression B = (XTX)-1XTY was used to check the results of the MATLAB regression function (with no noticeably significant errors found).

Finally, with the full equation, (**1**) **Y** = **BX**, we created our predicted values in order to evaluate the accuracy of our B matrix. This resulted in a matrix of nx1 dimensions called **P** where each value in **P** is a predicted sale price value for a specific house. As a side note, Eighty percent of our data was used in the training and construction of this model, (or in other words, the finding of the **B** matrix), and the remaining twenty percent was used to evaluate and test the model by being run through equation **1** and fulfilling the **P** matrix.

**3. Results**

Our results were mildly successful. Overall, the error was higher than desired, (with error defined as the actual sale price – the predicted sale price squared), but to say the model was inaccurate would simply be untrue. It was extremely accurate for certain areas of Fishers and for certain homes. We ran multiple variations of our model (with different data subsets) to see what was the cause of our inaccuracies for certain houses and to see which predictor values had the most impact on the sale price. Of the values we used, it seems the square footage, bedroom number, and bathroom number were the most impactful while the garage ownership, basement ownership, and year built values had little to no influence on the sale price.

This means that for every house whose value, consequently its sale price, can be summed by their square footage, bathroom number, and bedroom number, our model was dead-on accurate. It was for houses where these predictor values were not the only (and perhaps not even the most significant) attributes impacting the house’s value that our model was inaccurate for.

From our research into this, it seems there are two significant factors that we did not account: Location and the market’s buyers. Our mistake was collecting data from all over Fishers, instead of one single area. Our working hypothesis is that for the more secluded areas of Fishers our model is be more accurate, but for the more desirous and more developing areas (places closer to the town hall for example) the value of the homes were driven higher. Also, what we did not consider is the particular shape of the market’s buyers—specifically their ages. After speaking with a local realtor about the values of homes, we learned that a large number of recent buyers fit into the 50-70 year old range. This means that single story houses, despite not having basements or having smaller square footage and less bathrooms and bedrooms, could sell higher than what their typically worth to these particular buyers. In other words, a large facet of the market is skewing the sale prices higher for homes of lesser technical value in favor of homes without stairs.

**4. Conclusion**

This has been an incredible, eye-opening project. In the American educational system, mathematics is not taught: mathematical operations are. This means that to a common lay person, mathematics is simply arithmetic and algebra. But this is wrong: mathematics is the way we express the world in formal notation, and this misunderstanding was the greatest challenge I faced. To accomplish this project’s specific mathematical operations was not difficult, even if we were to not use applications like MATLAB. It was far harder to understand regression—to understand why we did what we did and to understand why it went right and wrong—than to simply multiply. The great takeaway from my time working this project is not how to do regression—that is, how to make linear graphs or perform operations to find slopes—but rather how to think in such a way as to capture the real world and its messy relationships into formal language. This is what separates an Information Scientist from a mindless programmer; this very ability to not merely see what-is but see why-it-is the way-it-is is our greatest achievement here. There are plenty of wonderful and valuable regression models out there—yes, even better than ours. Of course, there are plenty of improvements to be made: more data, more analysis of the location of our data, and analysis of the market incorporated into the model, but the point of this project was not simply to create a model, but to understand how to create a good one in order to prove that we are good Computer and Information Scientists. Certainly this means our work focused on creating this model, but everything presented here is not merely for that end. The greatest result of this project was not that a model has been built, but that the man who built the model who knew nothing of regression at this semester’s beginning is now ready and equipped not only to create and improve his own model but also, perhaps, challenge the work of those far more established.

Works Cited:

Gan, Victor, Vaishali Agarwal, and Ben Kim. "Data Mining Analysis and Predictions of Real Estate Prices." Issues in Information Systems, Volume 16, Issue IV. Pp. 30-36, 2015.