Machine Learning for the Masses: Regression Analysis

Machine learning and artificial intelligence have become very popular topics in the pricing community. However, many perceive these tools as overly complex and not practical to implement quickly into their existing processes. In this article, the author presents a machine learning technique called regression analysis and demonstrates simple applications that can be readily employed by pricers in any industry. *Nathan L. Phipps is a Consultant at Wiglaf Pricing. He is responsible for the training preparation and conjoint analysis that supports pricing projects for Wiglaf clients. He can be reached at nphipps@wiglafpricing.com.*

by Nathan L. Phipps

Machine learning has become a very popular topic in recent years. However, I fear that most people hear "machine learning" and assume that it must be a very difficult topic to understand. After all, machine learning is a subfield of artificial intelligence, so machine learning *must* be complex. On the contrary, some machine learning techniques are easy to understand, and they can be easily implemented in data analysis.

Today, I am going to discuss a machine learning technique known as regression analysis, and I aim to show how it is accessible to the masses.

There are many types of regression analysis. To keep things simple, I am going to review an example of simple linear regression. Simple linear regression looks for the relationship between a single independent variable and a single dependent variable.

Combining Exploratory Data Analysis and Regression Analysis

A friend recently asked me what exactly it is that I do for work. Wishing to spare her the technical details of a typical pricing workday, I summarized, "I look for patterns in data." I realized that my response is actually a good summary of exploratory data analysis: looking for patterns in your data and teasing out relationships. Exploratory data analysis is a way to glean insight from your dataset, usually using visualizations.

A typical pricing project begins with exploratory data analysis – with visualizing the data. An important data visualization tool is the scatterplot. Typically, a scatterplot visualizes two variables at a time, one on the x-axis and one on the y-axis.

A scatterplot can quickly show a relationship (or lack of relationship) between two separate variables in your dataset. Likewise, regression analysis allows you to model the relationship between different variables in your dataset. I'm going to review aspects of both exploratory data analysis and regression analysis for this article.

Height vs. Weight

Determining the relationship of height and weight is a very common simple linear regression exercise, so I will use that for my example. If you wish to follow along, you can access the dataset I used from Kaggle.

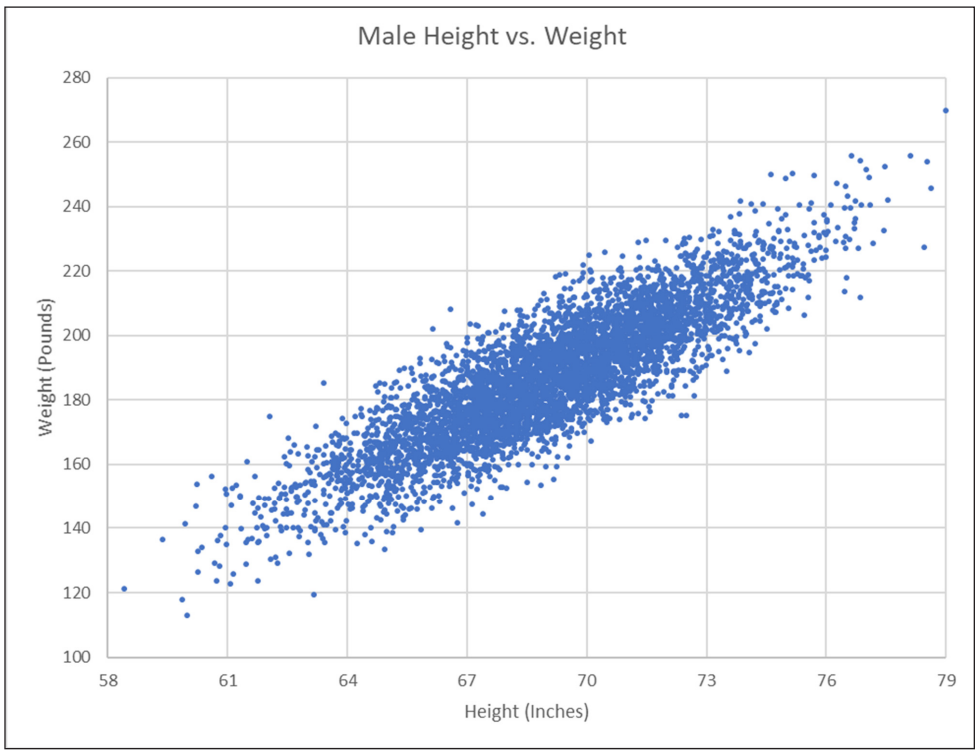
First, I filtered the list to only males so that I can focus on the relationship between height and weight for a single sex. Second, I graphed the data in a scatterplot, with height on the x-axis and weight on the y-axis. After shrinking the size of my markers and adjusting the axis ranges, I have the scatterplot shown in Figure 1.

Generally, you want the independent variable on the x-axis and the dependent variable on the y-axis. But how do you know which variable is your independent and which is your dependent?

One way to think about it is to determine which variable changes in response to the other. Your independent variable is the in-

➢ CONTINUED, next page ➢

Figure 1



put that you manipulate, and your dependent variable is what changes as a result.

For height and weight, you can see that adjusting someone's height would result in a change in weight. However, the converse is not necessarily true: adjusting someone's weight would not result in a change in height. Thus, the independent variable is height, and the dependent variable is weight. Weight changes in response to height, but not the other way around.

Naturally, pricing professionals will more than likely want to see a price on the y-axis and some other marketing variable on the x-axis. For instance, you may want to see what happens to price as the revenue of a transaction increases, or perhaps you are curious whether the price is impacted by the annual revenue of the customer purchasing.

Regression Analysis

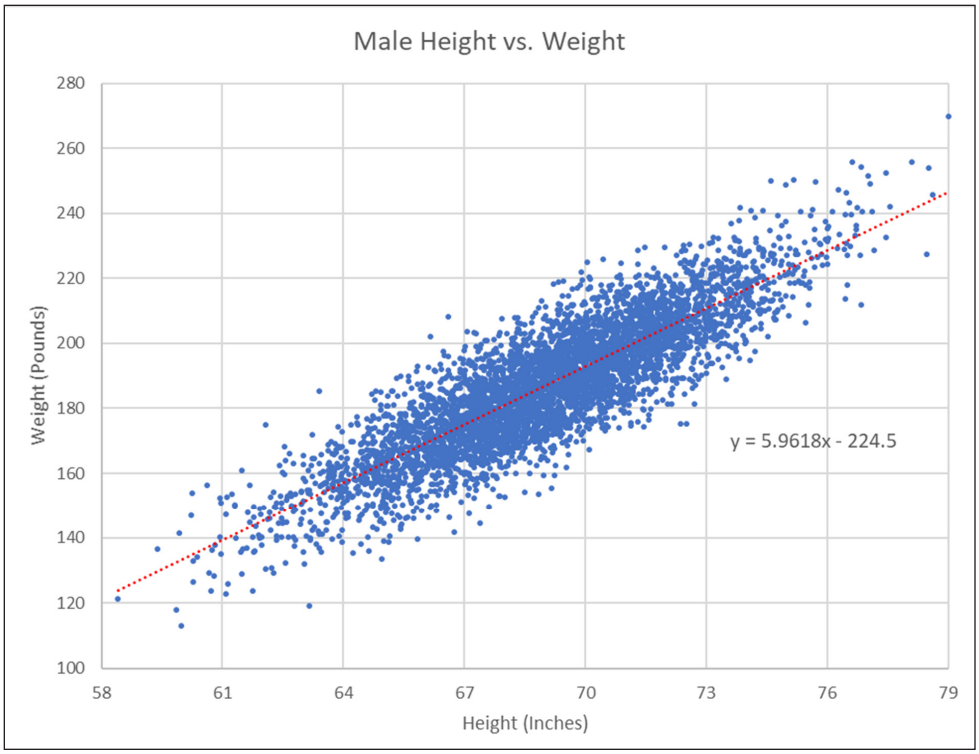
Scatterplots help to give you an intuitive feel for your data and any relationships that may exist. It is very clear from the scatterplot above that there is a strong relationship between height and weight. It appears that an increase in height results in an increase in weight, on average.

But by how many pounds does weight increase for each inch that height increases? Unfortunately, a scatterplot alone does not provide you with that level of detail. However, a regression analysis will.

Regression analysis is a common technique that is easily accessed in many statistical packages. The basic idea is that a regression analysis finds the line of best fit for the dataset. This regression line shows how your two variables are related, on average.

In Microsoft Excel (which I used to create these charts), adding the regression line for a simple linear regression is as simple as selecting your scatterplot, going to "add chart element", and selecting "add trendline." You can also choose to add the formula for the regression line, which I have.

Figure 2



(Microsoft Excel also allows you to complete a regression analysis using the Analysis ToolPak. This will give you additional statistics to help you determine how strong the fit is, in addition to the regression line above. Detailing that option is beyond the scope of this article. However, you can find detailed instructions in the book.)

The red line in Figure 2 on the scatterplot is the regression line. This shows the relationship of weight to height, on average.

The formula for the regression line provides us with the slope and y-intercept. We can see that each 1-inch change in height is associated with a 6-pound change in weight.

Parting Thoughts

I should remind the reader that for this analysis, I only reviewed linear regression. You should note that this method will not work if you are dealing with a non-linear curve (i.e., if the regression line is not best described by a straight line).

However, as you can see, simple linear regression is a machine learning technique that is very easy to implement, even if you are using Microsoft Excel. Combine it with a simple scatterplot, and you have a simple and yet powerful tool for data analysis.

So, don't fear machine learning. Embrace it! ❖