# Maximize your Project History with Linear Regression

By Tony Colburn

www.eosgroup.com info@eosgroup.com 602.429.0200

A historical project repository is one of the most important assets a company can possess. Although successful use of historical cost data requires rigor across multiple fronts, the application of some basic data validation principles opens the door to a treasure trove of information that is essential for future project planning and reducing the systemic risks associated with budget and schedule overruns.

In fact, there may not be a more impactful topic in construction technology than big data coupled with artificial intelligence and machine learning. For those of you working to refine your data strategy, it can be daunting as you collect, analyze, and identify process improvements surrounding usable data.

However, if you start with a handful of trusted project characteristic attributes, you can use basic statistics to learn more about past, current, and future project costs than ever before. Using statistics is nothing new, though with nearly infinite storage and processing capabilities, today's technology solutions make the use of data easier than ever.

**Linear Regression is one of the most effective and straight forward statistical methods**

While common in the capital project industry, parametric models for linear regression are especially useful for the vertical building industry. Linear regression is typically the preferred method for parametric models because it is widely accepted as a sound, practical method for using validated project history to achieve its maximum benefit as a feed-forward capability in new endeavors.

This method uses the concept of independent variables that correlate strongly to dependent variables. Typically, independent variables are explanatory variables, such as *Quantity*. *Quantity* can be expressed in commonly understood terms like area, number of floors in a building, stalls in a parking structure, hours per CY of production, and so on. The dependent variables are usually a scalar response to the independent variable. We most often see the dependent variable represented as *Cost*.

So a typical regression model might be used to predict the total cost of a new warehouse based on the square foot of production area. Or it might be used to predict the total cost of a new 100 bed hospital where a quantity variable of square feet and a number of beds variable are the independent variables and cost is the scalar dependent variable.

The application of regression is limited only by the statistical relationship between variables. The independent variable must be a statistically reliable predictor of dependent outcomes of the model, such as cost in our example. The R-squared ($R^2$) value is the measurement of the closeness of this relationship.

R-squared is a statistical measure of how close the data are to the fitted regression line (values closer to 1 show most significance). This value is also known as the coefficient of determination and is commonly used as a proxy for how well the algorithm predicts the calculated costs.

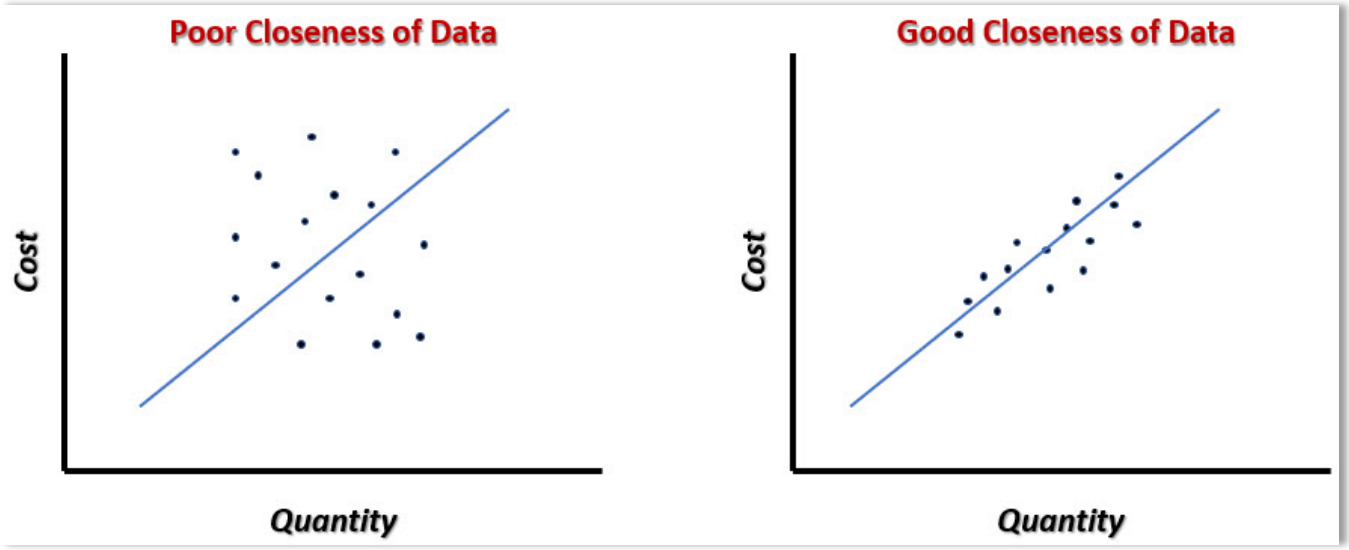


Figure 1. Illustration of Closeness of Data

**Planning for statistical modeling**

The first step in developing a parametric model is to establish its scope. This includes defining the end use of the model, the physical characteristics of the model, the cost basis of the model, and the critical components and cost drivers.

Typically, the end use of the model is to prepare conceptual estimates for a building or facility. However, benchmarking new projects against the model is also useful. You should determine the type of process the model will cover, the type of costs the model will estimate, and the intended accuracy range of the model as part of the end use definition.

**Building the statistical model**

In this example, we illustrate the data analysis process with the goal of producing a predicable statistical model for use in both developing a cost estimating capability for an office building, and benchmarking new projects against the model. The model will use the project size (quantity of gross square feet of building area) as an independent variable.

*Historical trends*

This example assumes a company wants to analyze their historical data to test the possibility of creating a building capacity-based model. An analysis of normalized costs for all office building projects produces a linear trendline demonstrating the relationship between the building gross area and the construction cost (Figure 2).

While not a strong correlation ($R^2$ value 0.49), this tells us that there are several possible reasons to explore further. Often projects included in the sample data suffer from systemic process issues, design and execution inefficiencies, and/or project-specific considerations which cause them to deviate from the larger data sample.

Data normalization is a legitimate, accepted practice for correcting the overall data sample to account for other factors skewing the results. The most important consideration is that any future predictive use of the model will be bounded by the range of the original data selected for use.

| Project title | Sub-project type | Project size | Capital construction (MM) | Square error |
|---|---|---|---|---|
| Owasco Office | Multi-story office | 70,000 SF | $ 77 | $ 5,973 |
| Canandaigua Tower B | Multi-story office | 40,000 SF | $ 168 | $ 28,231 |
| Canandaigua Tower A | Multi-story office | 50,000 SF | $ 131 | $ 17,233 |
| Empire Bay Office Zone 205 | Multi-story office | 35,000 SF | $ 125 | $ 15,653 |
| Empire Bay Office Zone 326 | Multi-story office | 60,000 SF | $ 233 | $ 54,385 |
| Ellison Office Retrofit | Multi-story office | 50,000 SF | $ 34 | $ 1,142 |
| Ellison Office Expansion 2010 | Multi-story office | 20,000 SF | $ 51 | $ 2,574 |
| Cayuga Office Zone 5A | Multi-story office | 53,500 SF | $ 230 | $ 52,683 |
| Cayuga Office Zones 4S & 5A | Multi-story office | 88,500 SF | $ 353 | $ 124,414 |
| Honeoye Office Area 3 | Multi-story office | 60,000 SF | $ 261 | $ 68,225 |
| Honeoye Office Area 2 | Multi-story office | 80,000 SF | $ 189 | $ 35,565 |
| Monroe Office | Multi-story office | 40,000 SF | $ 163 | $ 26,554 |
| Cayuga Office Zone 4S | Multi-story office | 35,000 SF | $ 123 | $ 15,177 |
| Honeoye Office Area 1 | Multi-story office | 90,000 SF | $ 267 | $ 71,328 |
| Monroe Office | Multi-story office | 50,000 SF | $ 142 | $ 20,030 |
| Lakeside Product A | Multi-story office | 52,000 SF | $ 135 | $ 18,276 |
| Lakeside Product B | Multi-story office | 41,800 SF | $ 172 | $ 29,656 |
| Skaneateles Office | Multi-story office | 35,000 SF | $ 53 | $ 2,816 |
| Skaneateles Office | Multi-story office | 43,000 SF | $ 63 | $ 3,969 |
| Keuka Office Expansion | Multi-story office | 19,000 SF | $ 41 | $ 1,685 |
| Keuka Office Reconfiguration | Multi-story office | 34,000 SF | $ 32 | $ 1,009 |
| Hamlin Office Reconfiguration | Multi-story office | 40,000 SF | $ 40 | $ 1,621 |
| Hamlin Office Expansion | Multi-story office | 20,000 SF | $ 55 | $ 3,072 |
| Seneca Office Area 1 | Multi-story office | 50,000 SF | $ 170 | $ 29,036 |
| Seneca Office Area 2 | Multi-story office | 40,000 SF | $ 207 | $ 42,994 |
| Owasco Office Retrofit | Multi-story office | 50,000 SF | $ 31 | $ 967 |
| Owasco Office Expansion 2010 | Multi-story office | 20,000 SF | $ 46 | $ 2,134 |
| Durand Eastman Office Section A | Multi-story office | 45,000 SF | $ 270 | $ 72,866 |
| Durand Eastman Office Section B | Multi-story office | 60,000 SF | $ 273 | $ 74,491 |
| Durand Eastman Office Section C | Multi-story office | 50,000 SF | $ 156 | $ 24,258 |
| Fairhaven Office Area 1 | Multi-story office | 50,000 SF | $ 178 | $ 31,701 |
| Otisco Office Section C | Multi-story office | 60,000 SF | $ 143 | $ 20,376 |
| Otisco Office Section B | Multi-story office | 75,000 SF | $ 431 | $ 185,982 |
| Otisco Office Section A | Multi-story office | 55,000 SF | $ 266 | $ 70,985 |
| Sodus Point Office Area 1 | Multi-story office | 90,000 SF | $ 315 | $ 99,290 |
| Sodus Point Office Area 3 | Multi-story office | 60,000 SF | $ 264 | $ 69,607 |
| Sodus Point Office Area 2 | Multi-story office | 80,000 SF | $ 215 | $ 46,217 |

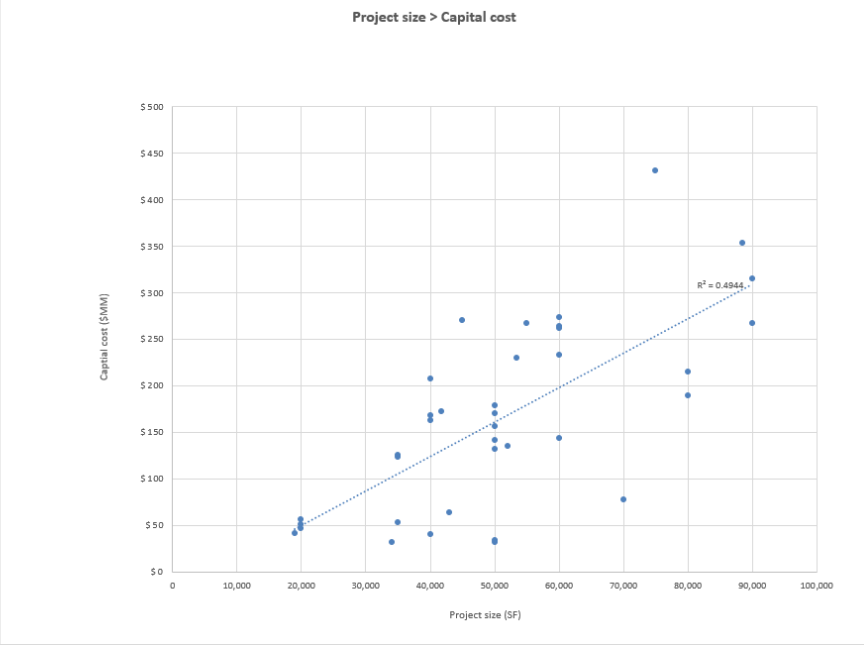


Figure 2. A selected group of project sample data and corresponding statistical grouping

When we look at data normalization (understanding the variation from the mean within a standard deviation) we can better focus on outliers and a range where the organization tends to have more predictability. Perhaps projects clustered around a specific range, and within a standard deviation, can be counted on to help make decisions. A range of projects between 35,000 and 90,000 SF is highlighted (Figure 3).

| Project title | Sub-project type | Project size | Capital construction (MM) | Square error |
|---|---|---|---|---|
| Owasco Office | Multi-story office | 70,000 SF | $ 77 | $ 5,973 |
| Canandaigua Tower B | Multi-story office | 40,000 SF | $ 168 | $ 28,231 |
| Canandaigua Tower A | Multi-story office | 50,000 SF | $ 131 | $ 17,233 |
| Empire Bay Office Zone 205 | Multi-story office | 35,000 SF | $ 125 | $ 15,653 |
| Empire Bay Office Zone 326 | Multi-story office | 60,000 SF | $ 233 | $ 54,385 |
| Ellison Office Retrofit | Multi-story office | 50,000 SF | $ 34 | $ 1,142 |
| Ellison Office Expansion 2010 | Multi-story office | 20,000 SF | $ 51 | $ 2,574 |
| Cayuga Office Zone 5A | Multi-story office | 53,500 SF | $ 230 | $ 52,683 |
| Cayuga Office Zones 4S & 5A | Multi-story office | 88,500 SF | $ 353 | $ 124,414 |
| Honeoye Office Area 3 | Multi-story office | 60,000 SF | $ 261 | $ 68,225 |
| Honeoye Office Area 2 | Multi-story office | 80,000 SF | $ 189 | $ 35,565 |
| Monroe Office | Multi-story office | 40,000 SF | $ 163 | $ 26,554 |
| Cayuga Office Zone 4S | Multi-story office | 35,000 SF | $ 123 | $ 15,177 |
| Honeoye Office Area 1 | Multi-story office | 90,000 SF | $ 267 | $ 71,328 |
| Monroe Office | Multi-story office | 50,000 SF | $ 142 | $ 20,030 |
| Lakeside Product A | Multi-story office | 52,000 SF | $ 135 | $ 18,276 |
| Lakeside Product B | Multi-story office | 41,800 SF | $ 172 | $ 29,656 |
| Skaneateles Office | Multi-story office | 35,000 SF | $ 53 | $ 2,816 |
| Skaneateles Office | Multi-story office | 43,000 SF | $ 63 | $ 3,969 |
| Keuka Office Expansion | Multi-story office | 19,000 SF | $ 41 | $ 1,685 |
| Keuka Office Reconfiguration | Multi-story office | 34,000 SF | $ 32 | $ 1,009 |
| Hamlin Office Reconfiguration | Multi-story office | 40,000 SF | $ 40 | $ 1,621 |
| Hamlin Office Expansion | Multi-story office | 20,000 SF | $ 55 | $ 3,072 |
| Seneca Office Area 1 | Multi-story office | 50,000 SF | $ 170 | $ 29,036 |
| Seneca Office Area 2 | Multi-story office | 40,000 SF | $ 207 | $ 42,994 |
| Owasco Office Retrofit | Multi-story office | 50,000 SF | $ 31 | $ 967 |
| Owasco Office Expansion 2010 | Multi-story office | 20,000 SF | $ 46 | $ 2,134 |
| Durand Eastman Office Section A | Multi-story office | 45,000 SF | $ 270 | $ 72,866 |
| Durand Eastman Office Section B | Multi-story office | 60,000 SF | $ 273 | $ 74,491 |
| Durand Eastman Office Section C | Multi-story office | 50,000 SF | $ 156 | $ 24,258 |
| Fairhaven Office Area 1 | Multi-story office | 50,000 SF | $ 178 | $ 31,701 |
| Otisco Office Section C | Multi-story office | 60,000 SF | $ 143 | $ 20,376 |
| Otisco Office Section B | Multi-story office | 75,000 SF | $ 431 | $ 185,982 |
| Otisco Office Section A | Multi-story office | 55,000 SF | $ 266 | $ 70,985 |
| Sodus Point Office Area 1 | Multi-story office | 90,000 SF | $ 315 | $ 99,290 |
| Sodus Point Office Area 3 | Multi-story office | 60,000 SF | $ 264 | $ 69,607 |
| Sodus Point Office Area 2 | Multi-story office | 80,000 SF | $ 215 | $ 46,217 |

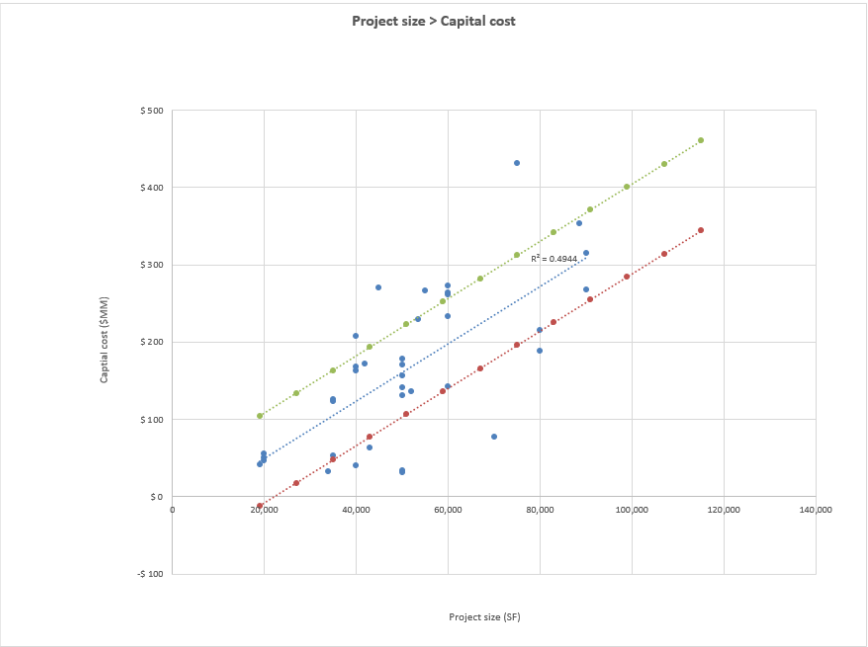


Figure 3. A normalized group of project sample data and corresponding statistical grouping

During our data normalization process, we removed project data outliers. Removing the unnecessary projects produces a simplified model that we could call *Office buildings - 35K-90K SF* with a much stronger $R^2$ value (Figure 4).

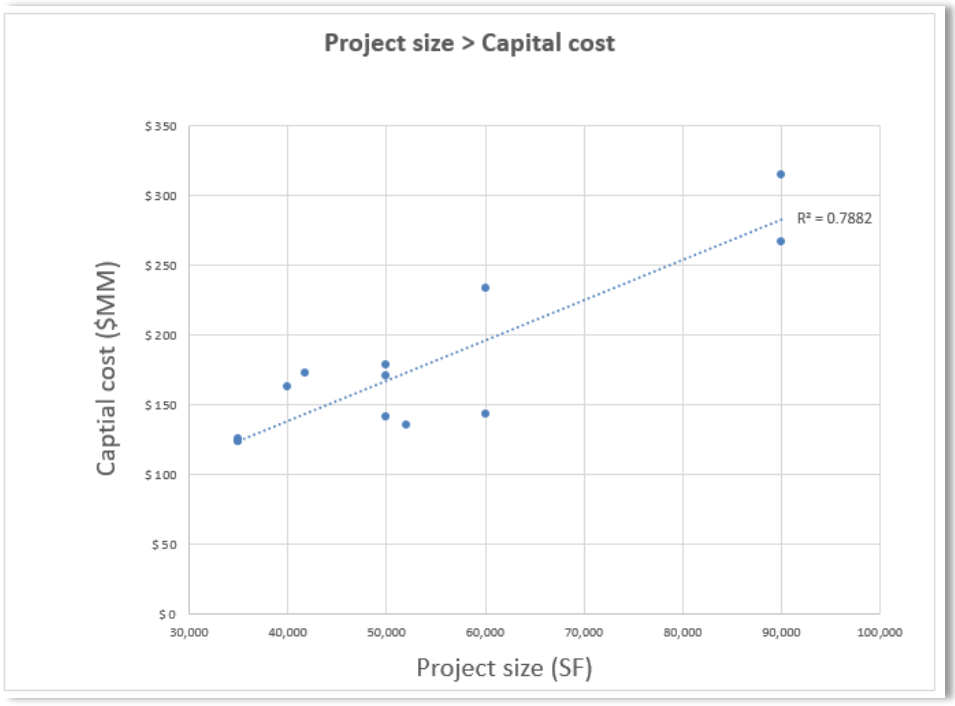


Figure 4. Normalized model Office Buildings

Adding standard deviations (annotated by the green and red lines) to the chart provides a level of confidence that we can use when exercising the model. It also gives us the ability to calculate the high and low ranges of cost associated with a new office building (Figure 5).

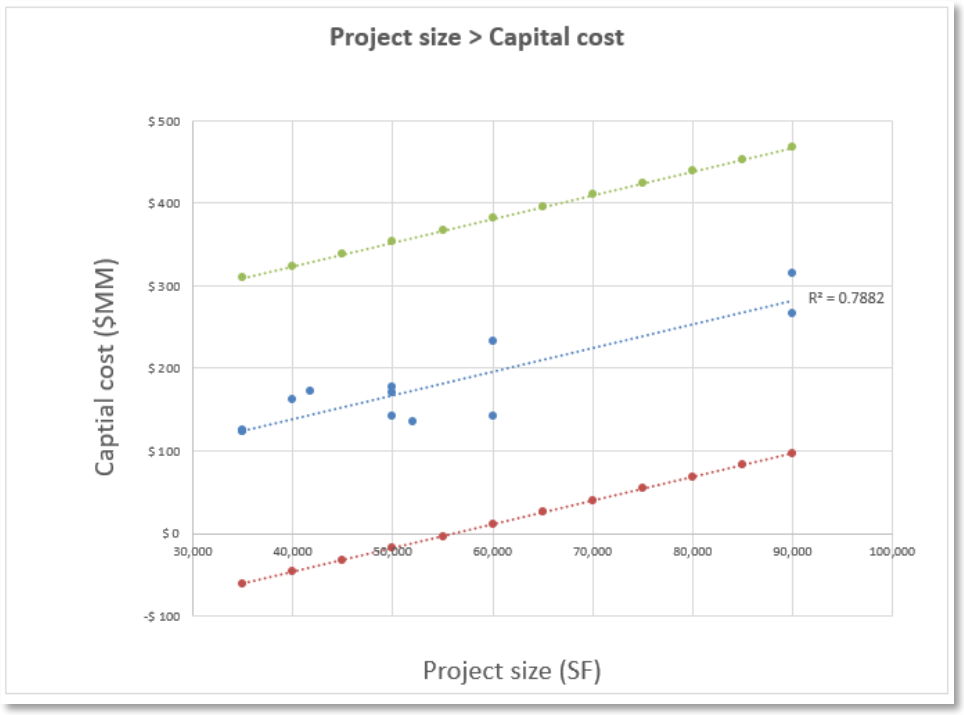


Figure 5. Normalized model Office Buildings (including Standard Deviation)

**Using the statistical model**

Now that we've built the model, we can quickly and easily use it for prediction and comparison by adding new data to the model.

*Benchmarking*

Using historical data to benchmark current data tells us how a project compares to a standard or comparator set (Figure 6).

Because we established the model based on trusted historical data, we know where we want projects to fall within the model: within a statistically significant range (or if not, to explore why).

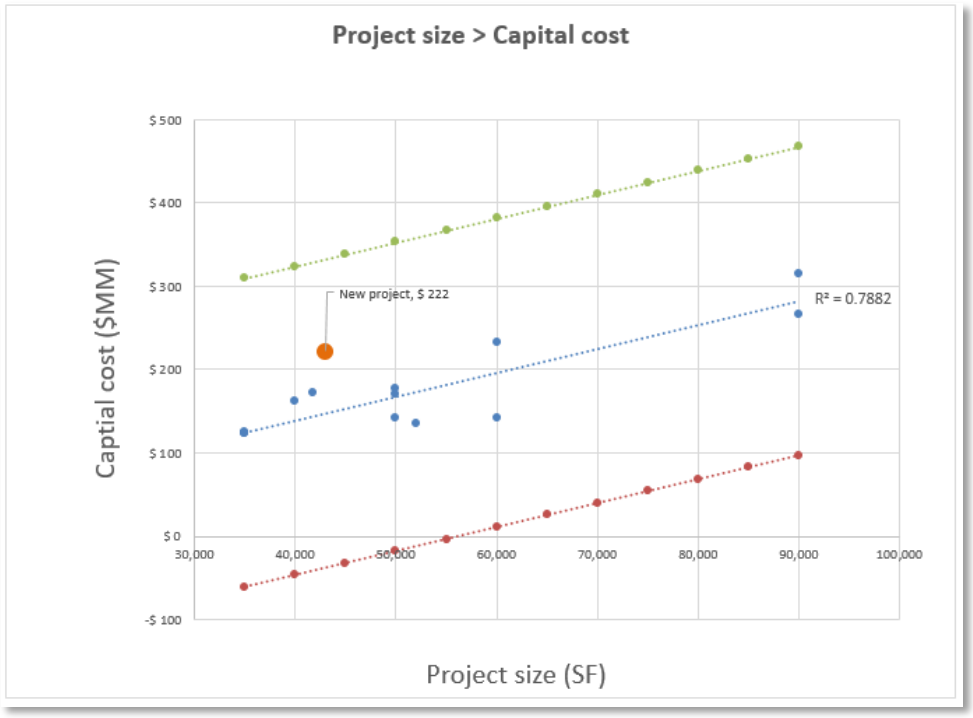


Figure 6. Benchmark project in the model

*Estimating*

Providing defensible, accurate estimate information quickly is also a high value use of historical data. As we've illustrated, estimates using parametric statistical means are defensible and accurate because they are based on statistically reliable historical project data. This is important to note because estimating best practices require this method be proven using the statistical means described in this article. In particular, using important indicators of data quality (i.e., R-squared)

Estimating cost or other data can be obtained using basic calculations readily available in Excel. Let's assume we need to provide a quick order of magnitude cost based on the gross building area of a new office tower. Using the same information, you can forecast cost using the model's linear regression function (Figure 7).

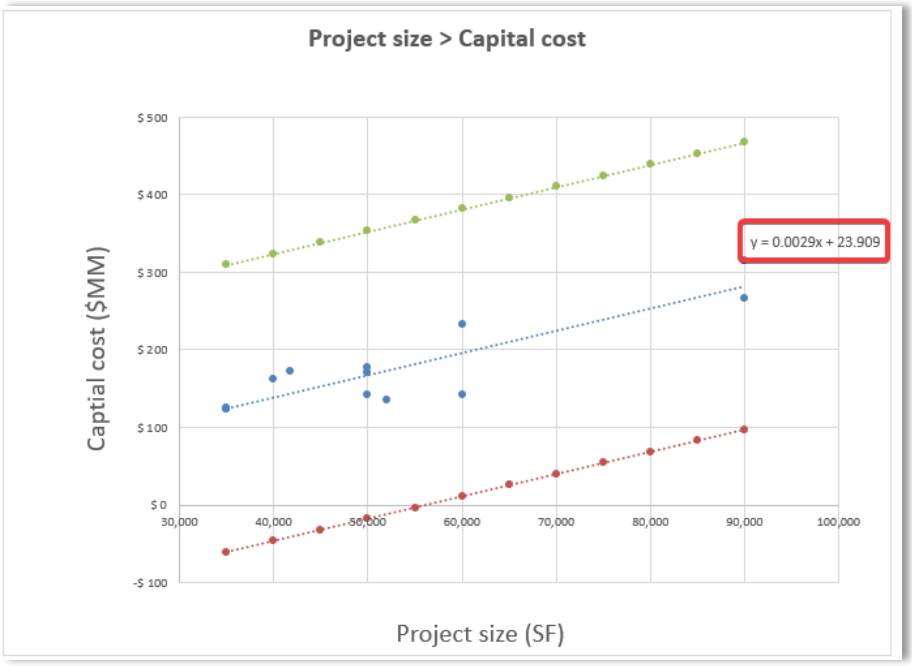


Figure 7. Linear regression function

As mentioned earlier, the application of regression is limited only by the statistical relationship between variables, where the independent variable must be a statistically reliable predictor of dependent outcomes of the model, such as *cost* in our example. The linear regression function provides a way to estimate cost.

The function calculates the "y" value, or *cost* in this case, which will fall directly on the regression line:

$$y = mx + b$$

Where "m" (slope) uses the difference of at least two cost values divided by the difference of at least two project size values. Leaving "b" (y intercept), which is where the trendline would fall on the "y" axis (Figure 8):

- New building size (x): 43,000 SF
- Find the cost (y):

$$y = 0.0029 \times 43{,}000 + 23.209$$

$$y = \$147{,}909{,}000$$

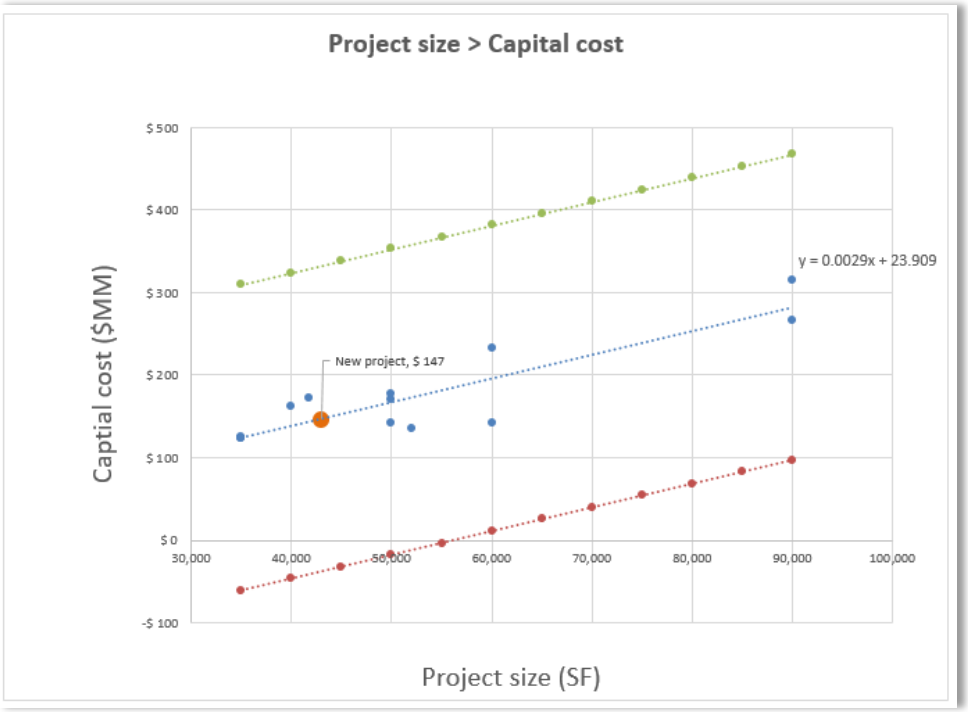


Figure 8. Estimated cost based on the linear regression function

Using spreadsheet software like Microsoft Excel to prototype statistical models makes it easy to chart and calculate data using linear regression. As stated earlier, a lot of rigor is needed to consistently represent, collect, and arrange your project data. Because of this time investment, Excel may not be the best long-term choice for housing parametric models.

Moreover, most corporate IT and intellectual property policies require storing sensitive estimating data in a more secure environment. It can be challenging to ensure data quality and storage practices using spreadsheets and documents. Check your company's policies because most companies prefer to use tools like Eos Cortex to protect their intellectual capital and sensitive competitive information.

More importantly, Eos Group's latest release of Cortex Project History provides simple, easy-to-use scatter charts showing data points, linear regression, and basic statistical information to help construction and engineering companies make the most of their data.

Simply follow the steps in this article to define your project storage strategy, publish project data, and leverage your cost history like never before.