Data Driven Estimating Part 1: How Data Feeds The Estimation Modeling Paradigm at Galorath / SEER

Summary

- Galorath constantly collects data from many sources, both public and private.
- Data must be processed to be useful, using methods that Galorath routinely and openly discusses.
- In addition to data, numerous other sources play an important role in maintaining the accuracy of SEER for Software estimates.
- As important as data is, innovation is equally so. The core SEER model has been improved and extended over time, and new features have been added, to make estimating ever easier, more applicable, and more accurate.
- "Data Driven Estimating Part 2: Data Driven Estimating Features in SEER for Software" is available upon request.

What Does It Mean To Be *Data Driven*?

These days software estimation vendors are competing to have the largest repositories of completed software projects, and the customer is encouraging this competition, which is fundamentally good. However, there is more to insuring the accuracy of an estimation model than just having a lot of data points sitting on the proverbial shelf.

Where Data Comes From

The first question asked of a vendor is, where does your data on completed software projects come from? Early on, much of it came from Government agencies, who in turn collected from contractors. Over time, public sources have emerged that contain voluntarily submitted information from private companies worldwide; the prime example of this being the International Software Benchmark Standards Group (ISBSG). Galorath has obtained software project data over the years through numerous private and public sources. The data comprises many thousands of total observations that have passed data quality tests. Most observations contain size and effort information, thousands more do not contain all the desired fields.

What is Done with Data

While plain-vanilla data can reveal a lot, it has its limitations. For this reason, Galorath maintains extensive surveillance of industry trends, including third-party analyses. These can reveal insight into changes in modern practices such as Agile development, the

productivity gained by the latest IDEs, and many other ongoing evolutions. The company is a member of numerous industry consortiums - in part to obtain access to the latest research available.

At Galorath, once data is acquired, it is processed into a form that is usable for analysis. This involves *normalization* so that the data points are comparable, i.e., include the same activities from early requirements through testing and the same types of labor, including programmers, testers, management, etc.. We also try to find and understand "outliers" – those projects that are so different that they are not useful. At numerous conferences and in webinars, we have described our normalization process and compared our results against other methods.

Using the collected data we update *SEER for Software* in several ways. A key method is to run our model against various *stratifications* (specific subsets such as "Business" and "Client-Server") that are defined by *SEER for Software's* knowledge bases. Simply put, we compare the model's estimates to observed outcomes. Based on these results, knowledge bases are re-calibrated when necessary. In fact, data sets are not uniform in terms of the information observed. Some completed project records may include peak staff, development activities, and software language used, while others don't. We account for this by performing a separate analysis of various factors: language productivity, development proportions, productivity variation by application or development method, and so on. These analyses are done first, and the model is adjusted, before gross analysis is begun.

SEER for Software's core model is configured to a particular circumstance by a set of knowledge bases, and it's these knowledge bases that are calibrated based on new industry information and trends. Each knowledge base is defined in terms of a set of parameters, some visible to users, and others normally hidden. When a knowledge base is updated the visible parameters, such as *Modern Development Practices*, may be modified and some underlying calibration factors may be adjusted. These knowledge base adjustments occur every few years as evidence warrants.

Innovative Features Support the Data Driven Approach

The overall evolution of *SEER for Software* is best called "innovation-driven" as well as "data-driven". While data analysis is a very important part of how we maintain *SEER for Software*, we also continually enhance the model's ability to estimate real world projects. These enhancements have often been industry firsts: flexible project staffing, off-the-shelf (COTS) integration modeling, translation of estimates into detailed project plans having intricate interdependencies, extended schedule and small schedule estimating, cloud computing solutions, to name only a few. All these innovations, alongside data-driven updates, serve an important role in insuring the model's precision.

Note "Data Driven Estimating Part 2: Data Driven Estimating Features in SEER for Software" is available upon request via info@galorath.com Additional information may be found at www.galorath.com