

2.3: The Example Data

I obtained the input data used for developing the regression models in the subsequent chapters from the publicly available CPU DB database [2]. This database contains design characteristics and measured performance results for a large collection of commercial processors. The data was collected over many years and is nicely organized using a common format and a standardized set of parameters. The particular version of the database used in this book contains information on 1,525 processors.

Many of the database's parameters (columns) are useful in understanding and comparing the performance of the various processors. Not all of these parameters will be useful as predictors in the regression models, however. For instance, some of the parameters, such as the column labeled Instruction set width, are not available for many of the processors. Others, such as the Processor family, are common among several processors and do not provide useful information for distinguishing among them. As a result, we can eliminate these columns as possible predictors when we develop the regression model.

On the other hand, based on our knowledge of processor design, we know that the clock frequency has a large effect on performance. It also seems likely that the parallelism-related parameters, specifically, the number of threads and cores, could have a significant effect on performance, so we will keep these parameters available for possible inclusion in the regression model.

Technology-related parameters are those that are directly determined by the particular fabrication technology used to build the processor. The number of transistors and the die size are rough indicators of the size and complexity of the processor's logic. The feature size, channel length, and FO4 (fanout-of-four) delay are related to gate delays in the processor's logic. Because these parameters both have a direct effect on how much processing can be done per clock cycle and effect the critical path delays, at least some of these parameters could be important in a regression model that describes performance.

Finally, the memory-related parameters recorded in the database are the separate L1 instruction and data cache sizes, and the unified L2 and L3 cache sizes. Because memory delays are critical to a processor's performance, all of these memory-related parameters have the potential for being important in the regression models.

The reported performance metric is the score obtained from the SPEC CPU integer and floating-point benchmark programs from 1992, 1995, 2000, and 2006 [6–8]. This performance result will be the regression model's output. Note that performance results are not available for every processor running every benchmark. Most of the processors have performance results for only those benchmark sets that were current when the processor was introduced into the market. Thus, although there are more than 1,500 lines in the database representing more than 1,500 unique processor configurations, a much

This page titled [2.3: The Example Data](#) is shared under a [CC BY-NC 4.0](#) license and was authored, remixed, and/or curated by [David Lilja](#) ([University of Minnesota Libraries Publishing](#)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.