2 | Understand Your Data

GOOD data is the basis of any sort of regression model, because we use this data to actually construct the model. If the data is flawed, the model will be flawed. It is the old maxim of *garbage in, garbage out.* Thus, the first step in regression modeling is to ensure that your data is reliable. There is no universal approach to verifying the quality of your data, unfortunately. If you collect it yourself, you at least have the advantage of knowing its provenance. If you obtain your data from somewhere else, though, you depend on the source to ensure data quality. Your job then becomes verifying your source's reliability and correctness as much as possible.

2.1 || Missing Values

Any large collection of data is probably incomplete. That is, it is likely that there will be cells without values in your data table. These missing values may be the result of an error, such as the experimenter simply for- getting to fill in a particular entry. They also could be missing because that particular system configuration did not have that parameter available. For example, not every processor tested in our example data had an L2 cache. Fortunately, R is designed to gracefully handle missing values. R uses the notation *NA* to indicate that the corresponding value is not available. Most of the functions in R have been written to appropriately ignore *NA* values and still compute the desired result. Sometimes, however, you must explicitly tell the function to ignore the NA values. For example, calling the `mean()` function with an input vector that contains NA values causes it to return *NA* as the result. To compute the mean of the input vector while ignoring the *NA* values, you must explicitly tell the function to remove the *NA* values using `mean(x, na.rm=TRUE)` .

2.2 || Sanity Checking and Data Cleaning

Regardless of where you obtain your data, it is important to do some *sanity checks* to ensure that nothing is drastically flawed. For instance, you can check the minimum and maximum values of key input parameters (i.e., columns) of your data to see if anything looks obviously wrong. One of the exercises in Chapter 8 encourages you explore other approaches for verifying your data. R also provides good plotting functions to quickly obtain a visual indication of some of the key relationships in your data set. We will see some examples of these functions in Section 3.1. If you discover obvious errors or flaws in your data, you may have to eliminate portions of that data. For instance, you may find that the perfor- mance reported for a few system configurations is hundreds of times larger than that of all of the other systems tested. Although it is possible that this data is correct, it seems more likely that whoever recorded the data simply made a transcription error. You may decide that you should delete those results from your data. It is important, though, not to throw out data that looks strange without good justification. Sometimes the most interesting conclusions come from data that on first glance appeared flawed, but was actually hiding an interesting and unsuspected phenomenon. This process of checking your data and putting it into the proper format is often called *data cleaning.* It also is always appropriate to use your knowledge of the system and the relationships between the inputs and the output to inform your model building. For instance, from our experience, we expect that the clock rate will be a key parameter in any regression model of computer systems per- formance that we construct. Consequently, we will want to make sure that our models include the clock parameter. If the modeling methodology sug- gests that the clock is not important in the model, then using the methodology is probably an error. We additionally may have deeper insights into the physical system that suggest how we should proceed in developing a model. We will see a specific example of applying our insights about the effect of caches on system performance when we begin constructing more complex models in Chapter 4. These types of sanity checks help you feel more comfortable that your data is valid. However, keep in mind that it is impossible to prove that your data is flawless. As a result, you should always look at the results of any regression modeling exercise with a healthy dose of skepticism and think carefully about whether or not

the results make sense. Trust your intuition. If the results don't feel right, there is quite possibly a problem lurking somewhere in the data or in your analysis.

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 col- lected 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 understand- ing and comparing the performance of the various processors. Not all of these parameters will be useful as predictors in the regression models, how- ever. 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 num- ber 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 num- ber of transistors and the die size are rough indicators of the size and com- plexity 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 process- ing 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 perfor- mance, 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 perfor- mance 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 smaller number of results are reported for each individual benchmark.

2.4 || Data Frames

The *fundamental* object used for storing tables of data in R is called a *data frame.* We can think of a data frame as a way of organizing data into a large table with a row for each system measured and a column for each parameter. An interesting and useful feature of R is that all the columns in a data frame do not need to be the same data type. Some columns may consist of numerical data, for instance, while other columns contain textual data. This feature is quite useful when manipulating large, heterogeneous data files. To access the CPU DB data, we first must read it into the R environ- ment. R has built-in functions for reading data directly from files in the csv (comma separated values) format and for organizing the data into data frames. The specifics of this reading process can get a little messy, depend- ing on how the data is organized in the file. We will defer the specifics of reading the CPU DB file into R until Chapter 6. For now, we will use a function called `extract_data()` , which was specifically written for reading the CPU DB file. To use this function, copy both the **all-data.csv** and **read-data.R** files into a directory on your computer (you can download both of these files from this book's web site shown on p. ii). Then start the R environment and set the local directory in R to be this directory using the *File -> Change dir* pull-down menu. Then use the *File -> Source R code* pull-down menu to read the **read-data.R** file into R. When the R code in this file completes, you should have six new data frames in your R environment workspace: `int92.dat` , `fp92.dat` , `int95.dat` , `fp95.dat` , `int00.dat` , `fp00.dat` , `int06.dat` , and `fp06.dat` . The data frame

`int92.dat` contains the data from the CPU DB database for all of the processors for which performance results were available for the SPEC Integer 1992 (Int1992) benchmark program. Similarly, fp92.dat contains the data for the processors that executed the Floating-Point 1992 (Fp1992) benchmarks, and so on. I use the `.dat` suffix to show that the corresponding variable name is a data frame. Simply typing the name of the data frame will cause R to print the en- tire table. For example, here are the first few lines printed after I type `int92.dat` , truncated to fit within the page:

	nperf	perf	clock	threads	cores	...
1	9.662070	68.60000	100	1	1	...
2	7.996196	63.100000	125	1	1	...
3	16.363872	90.72647	166	1	1	...
4	13.720745	82.00000	175	1	1	...
...						

The first row is the header, which shows the name of each column. Each subsequent row contains the data corresponding to an individual processor. The first column is the index number assigned to the processor whose data is in that row. The next columns are the specific values recorded for that parameter for each processor. The function `head(int92.dat)` prints out just the header and the first few rows of the corresponding data frame. It gives you a quick glance at the data frame when you interact with your data. Table 2.1 shows the complete list of column names available in these data frames. Note that the column names are listed vertically in this table, simply to make them fit on the page.

Table 2.1: The names and definitions of the columns in the data frames containing the data from CPU DB.

Column Number	Column name	Definition
1	(blank)	Processor index number
2	nperf	Normalized performance
3	perf	SPEC performance
4	clock	Clock frequency (MHz)
5	threads	Number of hardware threads available
6	cores	Number of hardware cores available
7	TDP	Thermal design power
8	transistors	Number of transistors on the chip(M)
9	dieSize	The size fo the chip
10	voltage	Nominal operating voltage
11	featureSize	Fabrication feature size
12	channel	Fabrication channel size
13	FO4delay	Fan-out-four delay
14	L1icache	Level 1 instruction cache size
15	L1dcache	Level 1 data cache size
16	L2cache	Level 2 cache size
17	L3cache	Level 3 cache size

2.5 || Accessing a Data Frame

We access the individual elements in a data frame using square brackets to identify a specific cell. For instance, the following accesses the data in the cell in row 15, column 12:

```
int92.dat[15,12]
```

```
## [1] 180
```

We can also access cells by name by putting quotes around the name:

```
int92.dat["71","perf"]
```

```
## [1] 105.1
```

This expression returns the data in the row labeled 71 and the column labeled perf . Note that this is not row 71, but rather the row that contains the data for the processor whose name is 71 . We can access an entire column by leaving the first parameter in the square brackets empty. For instance, the following prints the value in every row for the column labeled clock :

```
int92.dat[,"clock"]
```

```
##  [1] 100 125 166 175 190 200 225 233 266 275 231 233  99 250 266 291 300
## [18] 333 350 110  60  70  85 101 118  50 100 125  50  80  90 100  48  60
## [35]  64  80  96 125  99 100 120  66  77  66  75  66 100 120 133  66 100
## [52] 100 120 133  60  66  75  90 150 166 180 200 200 250 100 150 175 200
## [69] 100 133 133  75 100 125 150  50  60  75
```

Similarly, this expression prints the values in all of the columns for row 36:

```
int92.dat[36,]
```

```
##       nperf     perf clock threads cores TDP transistors dieSize voltage
## 36 13.07378 79.86399    80       1     1  NA          NA      NA      NA
##    featureSize channel FO4delay L1icache L1dcache L2cache L3cache
## 36        0.75    0.75      270        1       NA      NA      NA
```

The functions nrow() and ncol() return the number of rows and columns, respectively, in the data frame:

```
nrow(int92.dat)
```

```
## [1] 78
```

```
ncol(int92.dat)
```

```
## [1] 16
```

Because R functions can typically operate on a vector of any length, we can use built-in functions to quickly compute some useful results. For ex- ample, the following expressions compute the minimum, maximum, mean, and standard deviation of the perf column in the int92.dat data frame:

```
min(int92.dat[,"perf"])
```

```
## [1] 36.7
```

```
max(int92.dat[,"perf"])
```

```
## [1] 366.857
```

```
mean(int92.dat[,"perf"])
```

```
## [1] 124.2859
```

```
sd(int92.dat[,"perf"])
```

```
## [1] 78.0974
```

This square-bracket notation can become cumbersome when you do a substantial amount of interactive computation within the R environment. R provides an alternative notation using the $ symbol to more easily access a column. Repeating the previous example using this notation:

```
min(int92.dat$perf)
```

```
## [1] 36.7
```

```
max(int92.dat$perf)
```

```
## [1] 366.857
```

```
mean(int92.dat$perf)
```

```
## [1] 124.2859
```

```
sd(int92.dat$perf)
```

```
## [1] 78.0974
```

This notation says to use the data in the column named perf from the data frame named int92.dat . We can make yet a further simplification using the attach function. This function makes the corresponding data frame local to the current workspace, thereby eliminating the need to use the potentially awkward $ or square-bracket indexing notation. The following example shows how this works:

```
attach(int92.dat)
min(perf)
```

```
## [1] 36.7
```

```
max(perf)
```

```
## [1] 366.857
```

```
mean(perf)
```

```
## [1] 124.2859
```

```
sd(perf)
```

```
## [1] 78.0974
```

To change to a different data frame within your local workspace, you must first detach the current data frame:

```
detach(int92.dat)
attach(fp00.dat)
min(perf)
```

```
## [1] 87.54153
```

```
max(perf)
```

```
## [1] 3369
```

```
mean(perf)
```

```
## [1] 1217.282
```

```
sd(perf)
```

```
## [1] 787.4139
```

Now that we have the necessary data available in the R environment, and some understanding of how to access and manipulate this data, we are ready to generate our first regression model.