CHAPTER 7

Real-World Data

So far, we've been working purely in the abstract. It's time to take a look at some real data, and see if we can make any observations about it.

Try R is Sponsored By:

Complete to Unlock

Some Real World Data 7.1

Modern pirates plunder software, not silver. We have a file with the software piracy rate, sorted by country. Here's a sample of its format:

```
Country,Piracy
Australia,23
Bangladesh,90
Brunei,67
China,77
...
```

We'll load that into the `piracy` data frame for you:

```
> piracy <- read.csv("piracy.csv")
```

We also have another file with GDP per capita for each country (wealth produced, divided by population):

```
Rank    Country         GDP
1       Liechtenstein   141100
2       Qatar           104300
3       Luxembourg      81100
4       Bermuda         69900
...
```

That will go into the `gdp` frame:

```
> gdp <- read.table("gdp.txt", sep="  ", header=TRUE)
```

We'll merge the frames on the country names:

```
> countries <- merge(x = gdp, y = piracy)
```

Let's do a plot of GDP versus piracy. Call the `plot` function, using the `"GDP"` column of `countries` for the horizontal axis, and the `"Piracy"` column for the vertical axis:

```
> plot(countries$GDP, countries$Piracy)
```

It looks like there's a negative correlation between wealth and piracy - generally, the higher a nation's GDP, the lower the percentage of software installed that's pirated. But do we have enough data to support this connection? Is there really a connection at all?

R can test for correlation between two vectors with the cor.test function. Try calling it on the GDP and Piracy columns of the countries data frame:

```
> cor.test(countries$GDP, countries$Piracy)

        Pearson's product-moment correlation

data:  countries$GDP and countries$Piracy
t = -14.8371, df = 107, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8736179 -0.7475690
sample estimates:
       cor
-0.8203183
```

The key result we're interested in is the "p-value". Conventionally, any correlation with a p-value less than 0.05 is considered statistically significant, and this sample data's p-value is definitely below that threshold. In other words, yes, these data do show a statistically significant negative correlation between GDP and software piracy.

We have more countries represented in our GDP data than we do our piracy rate data. If we know a country's GDP, can we use that to estimate its piracy rate?

We can, if we calculate the linear model that best represents all our data points (with a certain degree of error). The `lm` function takes a *model formula*, which is represented by a *response variable* (piracy rate), a tilde character (`~`), and a *predictor variable* (GDP). (Note that the response variable comes *first*.)

Try calculating the linear model for piracy rate by GDP, and assign it to the `line` variable:

```
> line <- lm(countries$Piracy ~ countries$GDP)
```

You can draw the line on the plot by passing it to the abline function. Try it now:

```
> abline(line)
```

Now, if we know a country's GDP, we should be able to make a reasonable prediction of how common piracy is there!

ggplot2 7.2

The functionality we've shown you so far is all included with R by default. (And it's pretty powerful, isn't it?) But in case the default installation doesn't include that function you need, there are still more libraries available on the servers of the Comprehensive R Archive Network, or CRAN. They can add anything from new statistical functions to better graphics capabilities. Better yet, installing any of them is just a command away.

Let's install the popular `ggplot2` graphics package. Call the `install.packages` function with the package name in a string:

```
> install.packages("ggplot2")
--- Please select a CRAN mirror for use in this session ---
Loading Tcl/Tk interface ... done
trying URL 'http://rweb.quant.ku.edu/cran/src/contrib/ggplot2_0.9.2.1.tar.gz'
Content type 'application/x-gzip' length 2310996 bytes (2.2 Mb)
opened URL
==================================================
downloaded 2.2 Mb

* installing *source* package 'ggplot2' ...
** package 'ggplot2' successfully unpacked and MD5 sums checked
** R
** data
**  moving datasets to lazyload DB
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded

* DONE (ggplot2)
```

You can get help for a package by calling the help function and passing the package name in the package argument. Try displaying help for the "ggplot2" package:

```
> help(package = "ggplot2")
                Information on package 'ggplot2'

Description:

Package:            ggplot2
Type:               Package
Title:              An implementation of the Grammar of Graphics
Version:            0.9.1

...
```

Here's a quick demo of the power you've just added to R. To use it, let's revisit some data from a previous chapter.

```
> weights <- c(300, 200, 100, 250, 150)
> prices <- c(9000, 5000, 12000, 7500, 18000)
> chests <- c('gold', 'silver', 'gems', 'gold', 'gems')
> types <- factor(chests)
```

The qplot function is a commonly-used part of ggplot2. We'll pass the weights and values of our cargo to it, using the chest types vector for the color argument:

```
> qplot(weights, prices, color = types)
```

Not bad! An attractive grid background and colorful legend, without any of the configuration hassle from before!

`ggplot2` is just the first of many powerful packages awaiting discovery on CRAN. And of course, there's much, much more functionality in the standard R libraries. This course has only scratched the surface!

Chapter 7 Completed

Captain's Log: The end of chapter 7. Supplies are running low. Luckily, we've spotted another badge!

We've covered how to take some real-world data sets, and test whether they're correlated with `cor.test`. Then we learned how to show that correlation on plots, with a linear model.

Share your plunder:
Tweet

Continue