A quick demonstration of the importance of graphing your data: A polite nod of thanks to F. J. Anscombe.

As published in Benchmarks RSS Matters, July 2015

http://web3.unt.edu/benchmarks/issues/2015/07/rss-matters

Jon Starkweather, PhD

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and **S**tatistical **S**upport

http://www.unt.edu

http://www.unt.edu/rss

RSS hosts a number of “Short Courses”.
A list of them is available at:
http://www.unt.edu/rss/Instructional.htm

Those interested in learning more about R, or how to use it, can find information here:
`http://www.unt.edu/rss/class/Jon/R_SC`

A quick demonstration of the importance of graphing your data: A polite nod of thanks to F. J. Anscombe.

This month's article presents a rather interesting reminder of the importance of graphing data. The article also serves as an important reminder that model specification (i.e. a model's form; e.g. *linear model*, *quadratic model*, *cubic model*, etc.) should not be chosen without thought. As an example, consider the linear model; which is so popular it is often the default model form for many researchers. However, a linear model may not always be the most appropriate model (see: Starkweather, 2010).

Many data analysts (me included) often place more emphasis on the precision of textual / numeric output rather than more subjectively interpreted graphical output. This behavior is not recommended because; graphs can often convey rather glaringly the nuances of the data which are not readily conveyed in textual or numeric output. As a reminder to me and others, this article demonstrates a truly ingenious way of illustrating the fact that graphs are equally important with computation when working with data. I have occasionally re-learned this lesson over the years and was truly astonished recently when I came across *Anscombe's Quartet* (Anscombe, 1973). I still find it hard to believe I had not been made aware of this brilliant quartet of data earlier. However, now that I am aware of it, I felt compelled to pass it along. Anscombe's Quartet consists of four simple data sets, each with 11 cases and each with 2 variables (see Table 1).

Table 1: Anscombe's Quartet

	x1	y1	x2	y2	x3	y3	x4	y4
1	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
2	8.00	6.95	8.00	8.14	8.00	6.77	8.00	7.71
3	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
4	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
5	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
6	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
7	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
8	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
9	12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
10	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
11	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Note: x1, x2, & x3 are identical.

1 Are these data pairs the *same*?

First, we can import the data into R[1] using the function (and file path) listed below.

```
a.df <- read.table("http://www.unt.edu/rss/class/Jon/Benchmarks/Anscombe_df.txt",
  header = TRUE, dec = ".", sep = ",")
```

[1] http://cran.r-project.org/

Next, we take a look at some descriptive statistics, such as the mean, variance, and standard deviation of each variable.

```
apply(a.df, 2, mean)
    x1     y1     x2     y2     x3     y3     x4     y4
9.0000 7.5009 9.0000 7.5009 9.0000 7.5000 9.0000 7.5009
apply(a.df, 2, var)
     x1      y1      x2      y2      x3      y3      x4      y4
11.0000  4.1273 11.0000  4.1276 11.0000  4.1226 11.0000  4.1232
apply(a.df, 2, sd)
    x1     y1     x2     y2     x3     y3     x4     y4
3.3166 2.0316 3.3166 2.0317 3.3167 2.0304 3.3166 2.0306
```

Clearly, each pair is virtually identical with respect to the mean (M = 9.00), variance (V = 11.00), and standard deviation (SD = 3.12) of the x variables; and the mean (M = 7.50), variance (V = 4.13), and standard deviation (SD = 2.03) of the y variables. Next, we take a look at the correlations of each pair.

```
cor(a.df[,1:2])
          x1        y1
x1 1.0000000 0.8164205
y1 0.8164205 1.0000000
cor(a.df[,3:4])
          x2        y2
x2 1.0000000 0.8162365
y2 0.8162365 1.0000000
cor(a.df[,5:6])
          x3        y3
x3 1.0000000 0.8162867
y3 0.8162867 1.0000000
cor(a.df[,7:8])
          x4        y4
x4 1.0000000 0.8165214
y4 0.8165214 1.0000000
```

Again, we see that each pair of variables displays the same correlation coefficient (r = 0.816). Next, as one might expect, we see below the linear regression intercepts and coefficients are the same as well.

```
summary(lm(y1 ~ x1, a.df))$coef
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629
summary(lm(y2 ~ x2, a.df))$coef
            Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816
summary(lm(y3 ~ x3, a.df))$coef
             Estimate Std. Error  t value    Pr(>|t|)
```

```
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305
summary(lm(y4 ~ x4, a.df))$coef
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

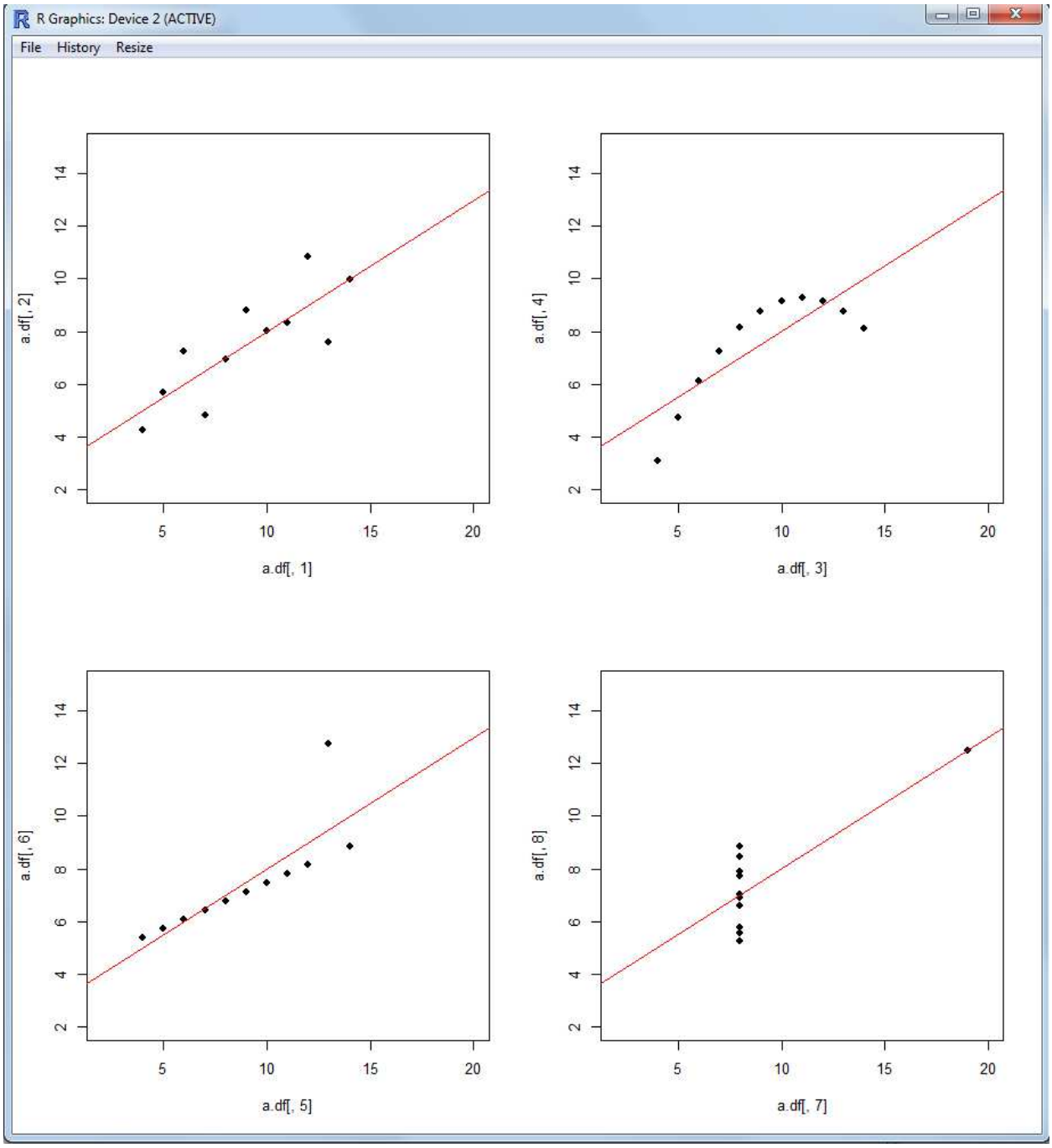
So, all four pairs of data result in the same linear regression equation:

$$y = 3.00 + 0.50 * x \tag{1}$$

2 Are these data pairs *really* the same?

Based on the above textual / numeric computations and output we might think these four pairs of data are *the same*. However, we quickly see that they are not at all the same once we do a simple scatterplot of each pair.

```
 par(mfrow = c(2,2))
plot(a.df[,1], a.df[,2], ylim = c(2,15), xlim = c(2,20), pch = 16)
abline(summary(lm(y1 ~ x1, a.df))$coef[,1], col = "red")
plot(a.df[,3], a.df[,4], ylim = c(2,15), xlim = c(2,20), pch = 16)
abline(summary(lm(y2 ~ x2, a.df))$coef[,1], col = "red")
plot(a.df[,5], a.df[,6], ylim = c(2,15), xlim = c(2,20), pch = 16)
abline(summary(lm(y3 ~ x3, a.df))$coef[,1], col = "red")
plot(a.df[,7], a.df[,8], ylim = c(2,15), xlim = c(2,20), pch = 16)
abline(summary(lm(y4 ~ x4, a.df))$coef[,1], col = "red")
```



So, let this be a reminder to us all - graphing data really matters. Graphing data is as important as computation when doing initial data inspection. A version of the R script used in this article can be found on the RSS Do-It-Yourself Introduction to R website[2] in the Module 12 section.

Until next time; *always look on the bright side of life...*

[2] `http://www.unt.edu/rss/class/Jon/R_SC/`

3 References and Resources

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17 - 21.
Available at:
`http://www.unt.edu/rss/class/Jon/MiscDocs/1973_Anscombe_theQuartet.pdf`

Starkweather, J. (2010). Model Specification Error...Are you straight, or do you have curves?
Benchmarks: *RSS Matters*, April 2010. Available at:
`http://www.unt.edu/rss/class/Jon/Benchmarks/Model%20Specification%20Error_J`

This article was last updated on June 4, 2015.

This document was created using LaTeX