

## Chapter 9

# Anscombe's Quartet: Graphs Can Reveal

**Abstract** We look at Anscombe's, (Am Stat 27(1):17–21, 1973) data which is one of the datasets that come with the R software. We compute different linear regressions of Anscombe's four sets of data—they give us the same results. When we look at the scatterplots corresponding to Anscombe's four sets of data we see that they are very different.

**Keywords** Anscombe · Graphs

### 9.1 Introduction

In 1973, the statistician Anscombe published a lovely paper in which he argued that we should use statistical graphs (p.17):

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding. ... Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.

Anscombe (1973) provided some synthetic data to illustrate this. Anscombe's data illustrates the importance of visualization, of examining the data graphically.

### 9.2 The Data: 4 Sets of xs and ys

R comes with some datasets; one of these is `anscombe`. We call the dataset `ans` and see what it contains:

```
> ans <- anscombe
> ans
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

The  $x$ s have means = 9 and standard deviations = 3.317 as can be verified by using the functions *mean* and *sd*. The  $y$ s have means = 7.5 and standard deviations = 2.03.

### 9.3 Same Regressions of $y$ s on $x$ s

When we regress  $y_1$  on  $x_1$  and  $y_2$  on  $x_2$  etc. using *lm*, and see the output with *display* the coefficients and other regression output are the same:

```
> fit1 <- lm(y1 ~ x1, data = ans)
> fit2 <- lm(y2 ~ x2, data = ans)
> fit3 <- lm(y3 ~ x3, data = ans)
> fit4 <- lm(y4 ~ x4, data = ans)
> library/arm)
> display(fit1)

lm(formula = y1 ~ x1, data = ans)
      coef.est coef.se
(Intercept)  3.00    1.12
x1           0.50    0.12
---
n = 11, k = 2
residual sd = 1.24, R-Squared = 0.67

> display(fit2)

lm(formula = y2 ~ x2, data = ans)
      coef.est coef.se
(Intercept)  3.00    1.13
x2           0.50    0.12
---
n = 11, k = 2
residual sd = 1.24, R-Squared = 0.67
```

```

> display(fit3)

lm(formula = y3 ~ x3, data = ans)
      coef.est coef.se
(Intercept)  3.00    1.12
x3           0.50    0.12
---
n = 11, k = 2
residual sd = 1.24, R-Squared = 0.67

> display(fit4)

lm(formula = y4 ~ x4, data = ans)
      coef.est coef.se
(Intercept)  3.00    1.12
x4           0.50    0.12
---
n = 11, k = 2
residual sd = 1.24, R-Squared = 0.67

```

## 9.4 Very Different Scatter Plots

We load the mosaic package.

```
> library(mosaic)
```

We use `xyplot` to plot scatters for the 4 sets of ys and xs and choose points (p) and regression lines (r):

```

> xyplot(y1 ~ x1, data = ans, type = c("p", "r"))
> xyplot(y2 ~ x2, data = ans, type = c("p", "r"))
> xyplot(y3 ~ x3, data = ans, type = c("p", "r"))
> xyplot(y4 ~ x4, data = ans, type = c("p", "r"))

```

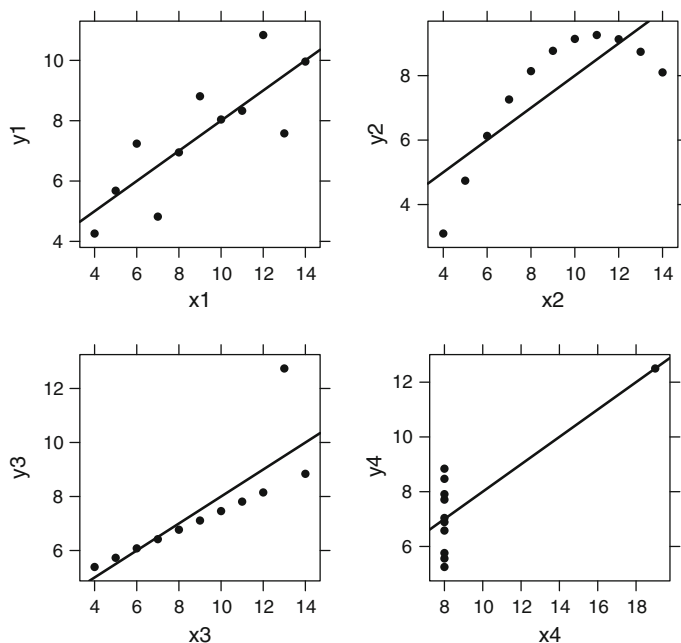
When we examine the lines of fit they are the same, but the scatter plots are very different (Fig. 9.1). In Fig. 9.1, a quadratic term should be used for  $y_2$  versus  $x_2$  because there is curvature. In the scatter of  $y_3$  against  $x_3$ , one outlying point is tilting the fitted line upwards. In the scatter of  $y_4$  against  $x_4$ , if we drop the extreme point, there is no relationship between  $y_4$  and  $x_4$ .

We choose (p) and loess smoother lines (smooth) to get Fig. 9.2.

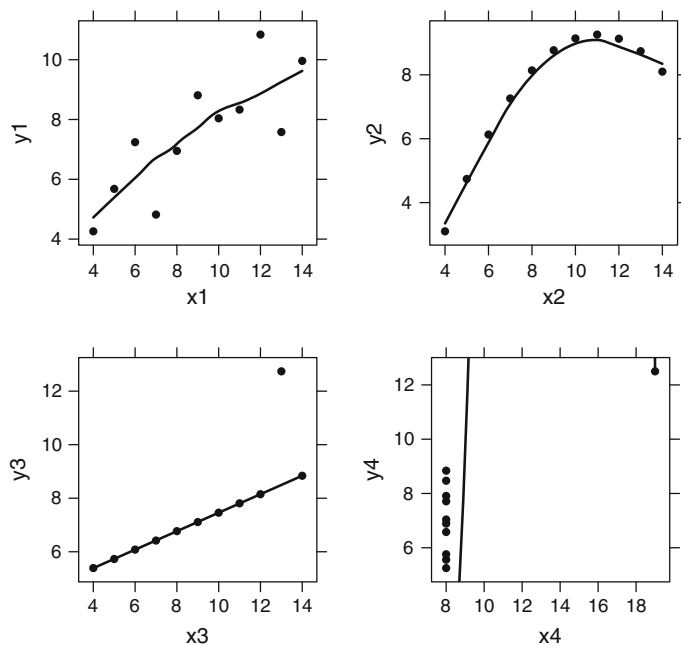
```

> xyplot(y1 ~ x1, data = ans, type = c("p", "smooth"))
> xyplot(y2 ~ x2, data = ans, type = c("p", "smooth"))
> xyplot(y3 ~ x3, data = ans, type = c("p", "smooth"))
> xyplot(y4 ~ x4, data = ans, type = c("p", "smooth"))

```



**Fig. 9.1** Scatterplots and linear regressions for the four datasets in Anscombe (1973)



**Fig. 9.2** Scatterplots and loess smoothers for the four datasets in Anscombe (1973)

What is also notable is that the smoother lines in Fig. 9.2 do not deceive us the way linear regression did.

## 9.5 Exploring Further

It is really worth reading Anscombe ([1973](#)) original paper.

## Reference

Anscombe FJ (1973) Graphs in statistical analysis. *Am Stat* 27(1):17–21