University of Waterloo W. H. Cherry

REGRESSION: Anscombe's Four Simple Linear Regression Data Sets

Of interest in the context of simple linear regression is an article in *The American Statistician* **27**(#1) 17-21 (1973) by F. J. Anscombe entitled *Graphs in Statistical Analysis.* The introduction of this article reminds us of the usefulness of graphs and pictorial displays in statistics and is as follows:

> Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:
>
> - numerical calculations are exact, but graphs are rough;
> - for any particular kind of statistical data, there is just one set of calculations constituting a correct statistical analysis;
> - performing intricate calculations is virtuous, whereas actually looking at the data is cheating.
>
> A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.
>
> Graphs can have various purposes, such as:
>
> * to help us perceive and appreciate some broad features of the data;
> * to let us look behind those broad features and see what else is there.
>
> Most kinds of statistical calculation rest on assumptions about the behaviour 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; if they are wrong, we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.
>
> Good statistical analysis is not a purely routine matter, and generally calls for more than one pass through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Of relevance to simple linear regression are four fictitious data sets given by Anscombe and reproduced (with minor refinements) at the right; the x values are the *same* for the first three data sets. Numerical summaries relevant to simple linear regression are essentially the *same* for all four data sets and are:

Data set 1:

x	4	5	6	7	8	9	10	11	12	13	14
y	4.25	5.68	7.24	4.82	6.98	8.80	8.03	8.32	10.84	7.58	9.96

Data set 2:

x	4	5	6	7	8	9	10	11	12	13	14
y	3.10	4.74	6.13	7.26	8.14	8.76	9.14	9.26	9.13	8.74	8.10

Data set 3:

x	4	5	6	7	8	9	10	11	12	13	14
y	5.39	5.69	6.08	6.44	6.80	7.12	7.44	7.83	8.14	12.74	8.83

Data set 4:

x	8	8	8	8	8	8	8	8	8	8	19
y	6.57	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

$\Sigma x_j = 99$, $\Sigma x_j^2 = 1{,}001$; $\Sigma x_j y_j = 797.5$;

$\Sigma y_j = 82.5$, $\Sigma y_j^2 \simeq 660.0013$; $(n = 11)$

$\overline{x} = 9$, $\overline{y} = 7.5$;

$SS_{xy} = 55$, $SS_x = 110$, $SS_y \simeq 41.2513$;

hence, the *estimates* of β_1 (the slope) and β_0 (the **Y**-intercept) of the regression of **Y** on **X** are: $\hat{\beta}_1 = 0.5$, $\hat{\beta}_0 = 3$,
so the equation of the straight-line model is: ${}_{\text{reg}}\overline{y} = 3 + 0.5x = 7.5 + 0.5(x - 9)$; also: $\hat{\sigma} \simeq 3.7083$, $r^2 \simeq 0.66665$.

Scatter diagrams of the four data sets, with the estimated regression of **Y** on **X** superimposed on each, are given overleaf on page HL53.2; *visual* assessment shows the following:

Data set 1: The scatter of the points about the line seems to meet the following three (visual) criteria for acceptability:

- the points should lie (reasonably) *close* to the fitted line – we may be able to tolerate a *small proportion* (typically one or two points) with larger estimated residuals, but it is then important to follow up on such observations to try to find out *why* they are more deviant than the other observations;
  - *if* the reason(s) for one or two deviant observations can be found, it *may* be useful to recalculate the estimated regression of **Y** on **X** with the more deviant observation(s) omitted;
- there should be no obvious *pattern* in the way the observations fall on the two sides of the fitted line;
- there should not be any systematic *change* in the magnitude of the estimated residuals with increasing values of **X** – this is consistent with the modelling assumption of *constant standard deviation about the regression line.*

Data set 2: The dependence of **Y** on **X** appears to be *non*-linear and, in place of a *straight-line* model, a *quadratic* polynomial would be a better choice.

Data set 3: The *outlier* at $\mathbf{X} = 13$ pulls the fitted line upwards so it is *not* a good fit *either* to this observation *or* to the remaining ten; if these were real data, we should try to find out *why* this observation differs from the other ten.

Data set 4: The estimate of the slope of the line depends mainly on the *one* observation with $\mathbf{X} = 19$, so it would be important to:

- check that this influential observation is *accurate*:
- understand *why* the data set has unusual **X** values – all but one are the *same.*

In addition, with observations at only *two* **X** values, we do not know if the relationship between **Y** and **X** is *linear* over the interval $\mathbf{X} = 8$ to $\mathbf{X} = 19$.

(*continued overleaf*)

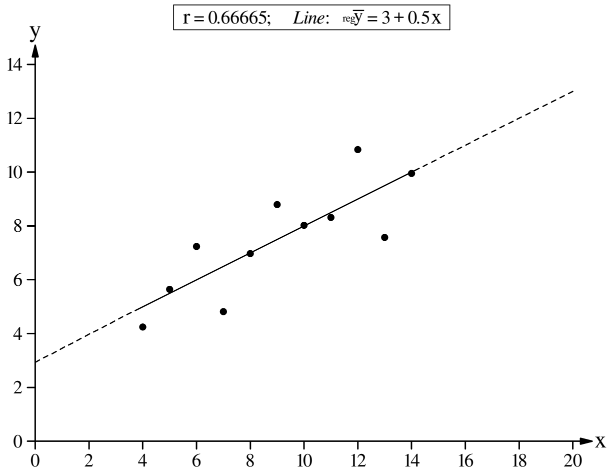
Thus, the straight-line model is *un*successful for three of the four data sets. The reader is referred to Anscombe's article for more detailed discussion and other matters of statistical interest.

The major lesson from Anscombe's four data sets – with their vastly differing *visual* implications but virtually identical *numerical* summaries – for this Highlight #53 is the importance of *visual inspection* as our *primary* method of assessing the fit of a regression model, and the *supplementary* nature of the (still useful) information provided by the ANOVA table and the value of r^2. We are also reminded that, essential as numerical summaries are for data analysis, they must be used with due care.

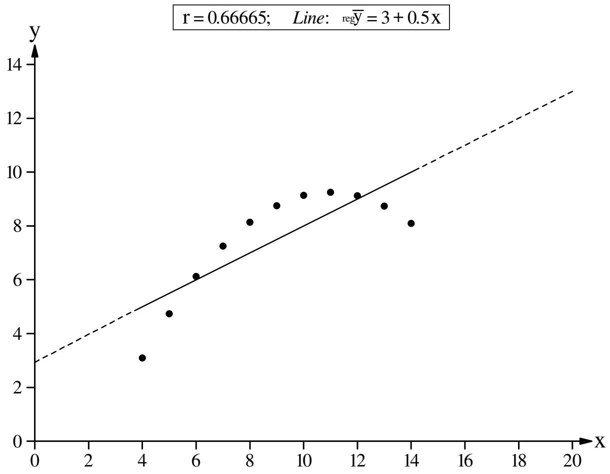
Exercise: For Data set 3, if the outlier at $\mathbf{X} = 13$ is omitted, show that the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ for the remaining ten points is: ${}_{\text{reg}}\overline{y} = 4.003 + 0.3457x = 6.976 + 0.3457(x - 8.6)$ – this line is shown in *longer* dashes on the lower left diagram below.

- For the recalculated line, confirm that $r^2 = 0.999\,607$ – this is a dramatic *in*crease from the previous value of 0.666 647 and reflects the fact we can *see* that the ten points lie almost *on* the recalculated line.

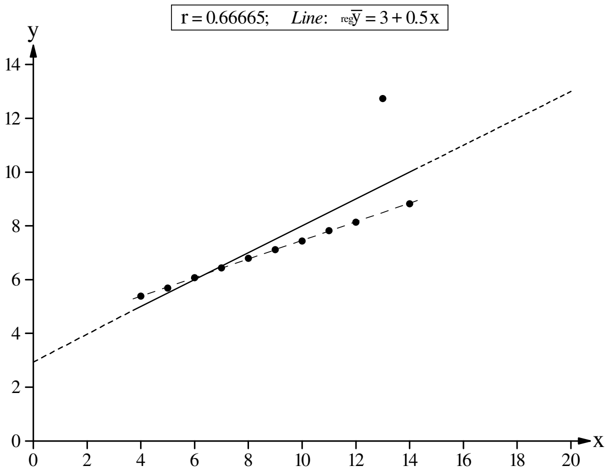
Data set 1



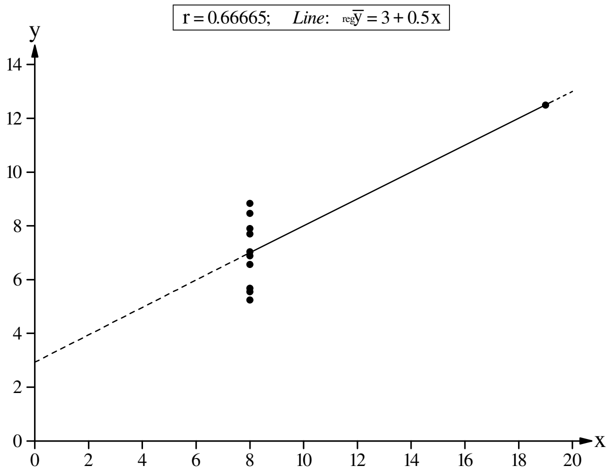
Data set 2



Data set 3



Data set 4



NOTE: The respective values of Σy_j^2 for the four data sets are 660.0022, 660.001, 660.0012 and 660.001; the approximate figure of 660.0013 given overleaf on page HL53.1 is close to the average of these four values.

The information in this Statistical Highlight #53 is also used in Figure 16.5 of the STAT 231 Course Materials.