3.2 The Anscombe Quartet

Data Set A	
x	*y*
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Data Set B	
x	*y*
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

Data Set C	
x	*y*
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Data Set D	
x	*y*
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Frank Anscombe invented the data sets above to teach his students some very important lessons. After you are done, see if you can determine what lesson(s) Mr. Anscombe is teaching us.
For each data set, do the following.

1. Find the least-squares linear regression equation, and the values of r and r^2.
2. Obtain and record a scatterplot of the data with the regression equation shown on the scatterplot.
3. Obtain and record the residual plot for each data set.

LSR Eqn.				
r				
r^2				

PURPOSE AND BACKGROUND

The purpose of this activity is to "teach" students some lessons about least-squares regression through having them compare seemingly different data sets. Frank Anscombe originally developed these data sets, which are now known as Anscombe's quartet.

TEACHING IDEAS

I prefer to give this handout to students soon after we've discovered that technology can spit out the LSR model quickly. Students are often tempted to skip looking at the scatterplot and go straight to the "real stuff" (equations, r, r^2).

I usually give this handout to groups of four students and have them each do one set, then share their results. In the end, all students are expected to have all information on their paper. This is not something I use as a summative assessment, but it could be used as a formative assessment. Ideally, I think that these data are best used for investigative purposes.

ANSCOMBE'S "LESSONS"

Here are some lessons I've learned from Frank Anscombe. I'm sure you can add to the list.

- Always plot your data! Calculators and software are powerful and can quickly give us lots and lots of numbers and equations, but make sure you look at your data visually.
- Linear relationships are revealed by a "random scatter" of residuals: See Data Set A. We expect randomly collected data to show no pattern in the residuals.
- Non-linear data can have high values of "r". See Data Set B. It follows that linear models of non-linear data can have high values of r^2. This simply means the model predicts the data well – it *doesn't* mean that the model we've chosen is the most appropriate one.
- Plotting data can help identify outliers. See Data Set C. The value of r would be 1 without the outlier. The value of r would be 1, if not for that single point.
- Influential observations can greatly change the equation of the regression model, r, and r^2. See Data Set D. Look for these "influential observations" at the extreme left and right of your data. The value of r would be 0 if not for that single point!