Simple Linear Regression

Some Cautions

Statistical methods work well in some situations, but care must be taken to use them appropriately. In this section, we'll take a look of some examples of situations that illustrate when taking that care will help ensure that sensible conclusions are made.

Example 1

Chlorofluorocarbons or CFCs are chemicals that were widely used in refrigerants and aerosol cans. However they are also known to contribute to ozone depletion in the under atmosphere. Concentrations of CFC in the atmosphere have been measured in a variety of locations around the globe since 1977, as can be seen in Figure 1. There is a strong linear pattern in the growth of CFCs over time. Using the method of least squares, we can find the equation of a line to describe that growth and we get:

$$\widehat{\text{CFC}} = -19064.03 + 9.71 \times \text{year}$$

The slope is approximately 9.7, so on average, the concentration of CFCs is growing by 9.7 parts per trillion each year. The intercept is the estimated value of the concentration at time 0. Of course in the year 0, CFCs hadn't yet been invented, and the negative intercept is meaningless. Interpreting the intercept in a practical way only makes sense when having 0 as the predictor variable is in the range of our data.

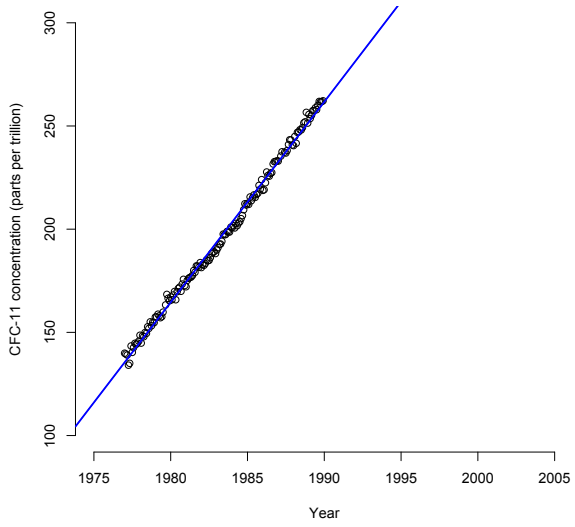

Figure 1: Scatter plot of CFC concentration (parts per trillion) versus year (Example 1)

We can also use regression lines to make predictions. Suppose we wanted to predict the CFC concentration in 1995. Then

$$\widehat{\text{CFC}} = -19064.03 + 9.71 \times 1995 \doteq 309.4$$

However we might not be willing to trust this prediction because 1995 is 5 years outside the range of our data. In fact, in 1987 the Montreal Protocol was developed, which laid out a schedule for phasing out the manufacturing of CFCs. Figure 2 shows the CFC concentrations extended beyond 1990. We can see the effect of the Montreal Protocol, as the atmospheric concentrations of CFCs began to decrease. Making a prediction at 1995, which was beyond the range of the data we used for our line, is called **extrapolation**. Extrapolation is always dangerous. We may think a linear model is a reasonable way to describe what we see in the data, but that doesn't mean the pattern of a linear relationship exists outside the range of the observed data.

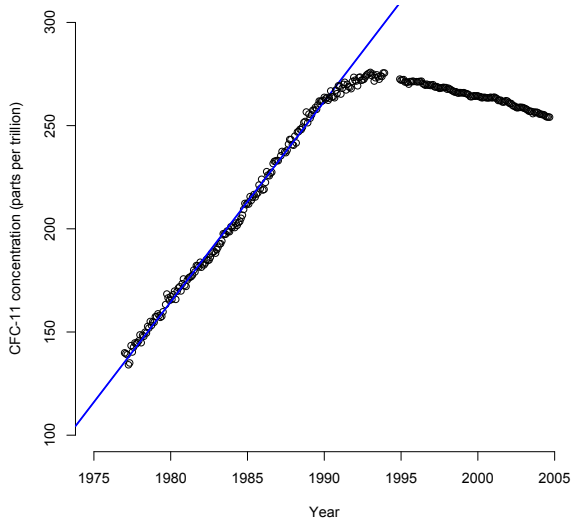

Figure 2: Scatter plot of CFC concentration (parts per trillion) versus year extended beyond 1990 (Example 1)

For the range of data in Figure 1, it seems that the straight line model fits well, but on a closer look the data seem to follow a pattern of points above the line, and then below the line, and then above the line. A systematic pattern like this is an indication that another, more complicated model, might fit the data better than the straight line. But how can we make sure we don't miss a subtle pattern like this? Residual plots are useful tools as they can exaggerate these kinds of patterns so we are less likely to miss them.

A residual is the vertical distance from a point to the regression line. Some of the residuals are negative, corresponding to points that are below the line, and some are positive. In the residual plot of the CFCs data in Figure 3 we can see the systematic pattern of mostly

positive residuals then negative then positive, corresponding to points below, above, then below, then above the line. If a line is an appropriate model for our data, the residual plot should look like random scatter about 0.

Sometimes, instead of being plotted against the predictor variable, residuals are plotted against the corresponding predicted values of the response variable. For regression with one predictor variable, this has the effect of simply changing the scale on the horizontal axis, and we interpret the plot in the same way. See Figure 3.

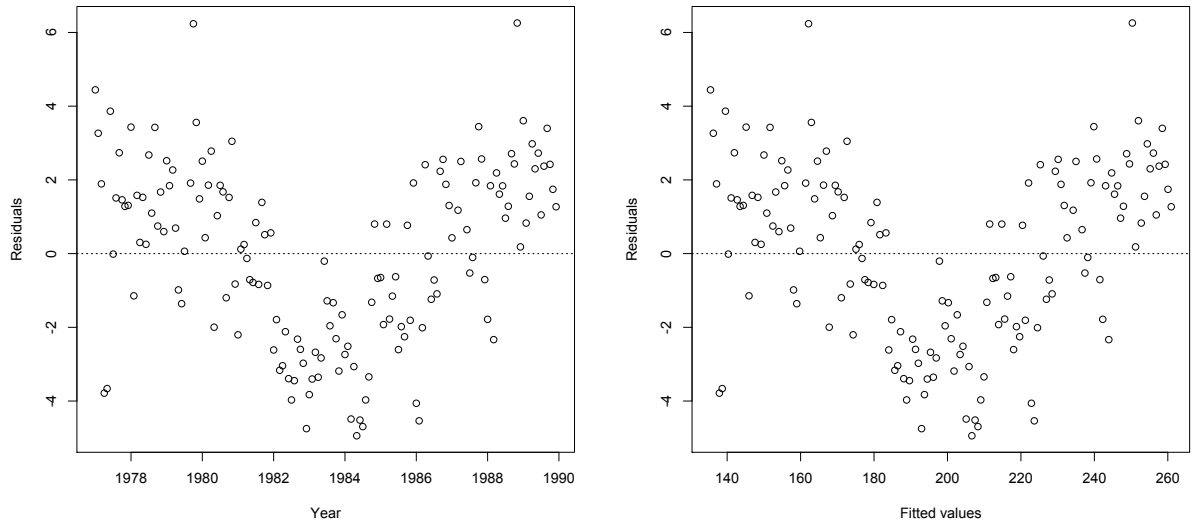

Figure 3: Plots of residuals versus the explanatory variable and versus the predicted values for Example 1

Example 2
Table 1 and Figure 4 show data and the corresponding scatterplots of 4 data sets, each with 11 points. They aren't real data, but were created by a the statistician Anscombe to illustrate a point. For each of these data sets, summary statistics, including the means of x and y, and their standard deviation and the correlation between x and y are the same. As a result, the estimated regression line would be the same. However, a linear model is only appropriate for the first set of data. The second would be better described by a curve. The third looks like a nice linear relationship, but whose regression line is shifted by one stray point. And the fourth would have no linear relationship, if it wasn't for the single outlying point. The statistician's point in creating these data sets was to emphasize that the numbers are not the whole story. Make sure you plot the data first to see if a linear model is a sensible choice.

The third and the fourth scatterplots in Figure 4 include **influential points**. Both of these examples have one point that, if we removed it from the data, the resulting regression line would change by a substantial amount.

x_1	y_1	x_1	y_2	x_1	y_3	x_2	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
7	4.82	7	7.26	7	6.42	8	6.89
5	5.68	5	4.74	5	5.73	19	12.50
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
s.d.	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
correlation	0.82		0.82		0.82		0.82	

Regression line for all 4 data sets: $\hat{y} = 3.0 + 0.5x$

Table 1: The 4 Anscombe datasets and their summary statistics (Example 2)

Influential point: A point is influential if the regression line changes substantially when the point is removed from the data.

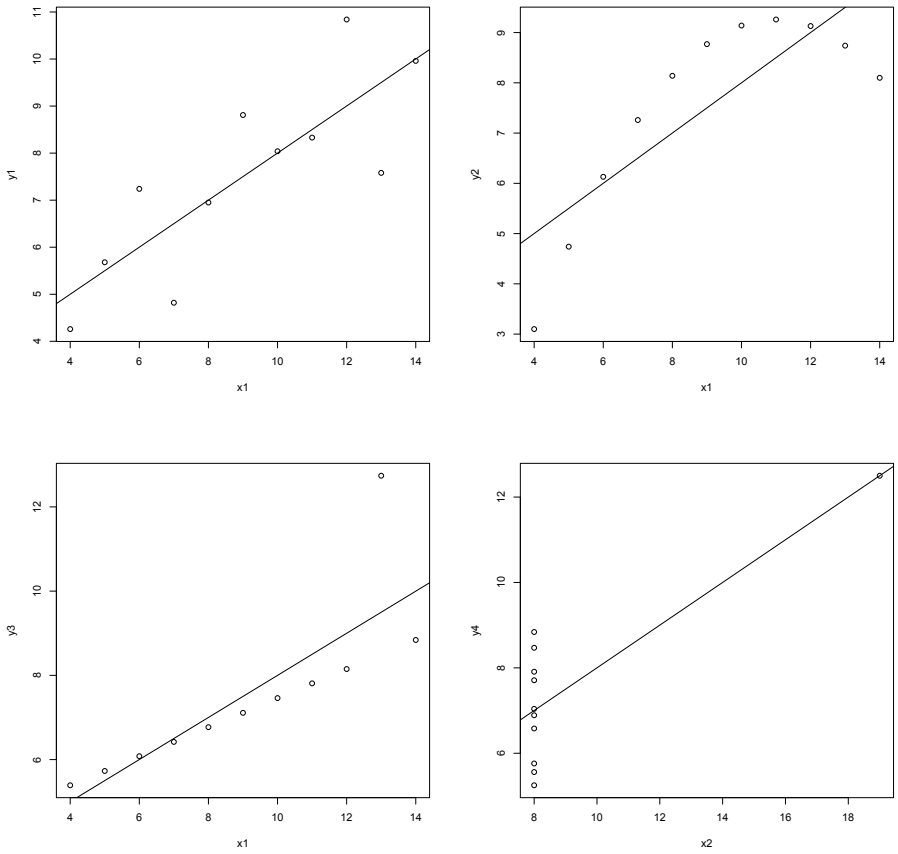

Figure 4: Scatterplots and corresponding regression lines of the 4 Anscombe datasets (Example 2)

Example 3

Figure 5 shows the relationship between average crawling age of babies and temperature. There is one point that is far away from the regression line. The magnitude of its corresponding residual is much larger than the magnitudes of all of the other residuals. In the distribution of residuals, it is an outlier. We can remove that point from the dataset, and refit the regression line to the remaining points to see if the line changes by much. We see that the slope is very similar and the intercept is shifted up a small amount. So this point doesn't fit the data well, but it's not influential.

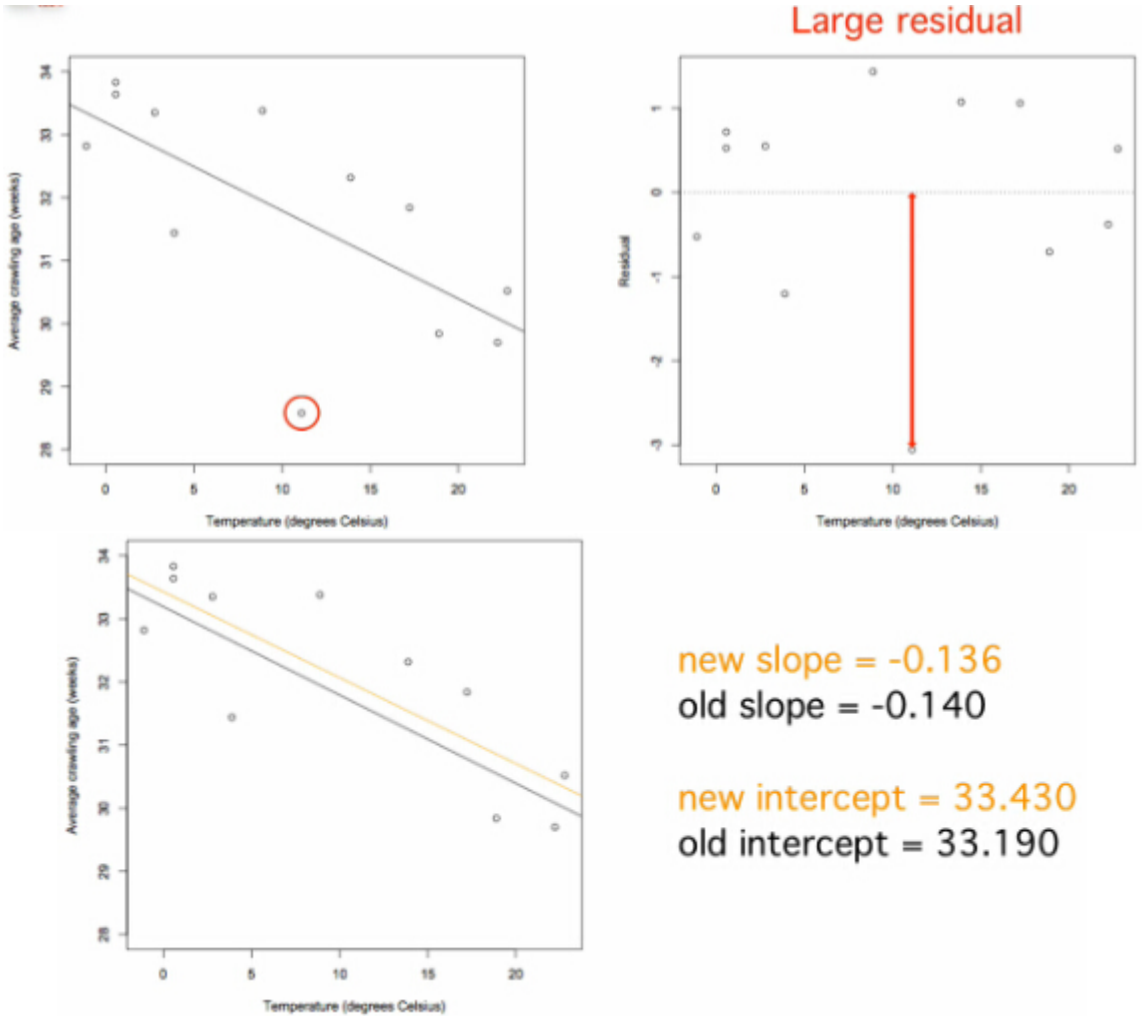

Figure 5: Scatterplot, residual plot, and effect of point of large residual for babies crawling data (Example 3)

Points that are influential are often outliers in the horizontal direction. Outliers in the horizontal-direction are called **leverage points**. Figure 6 shows some examples of leverage points. Sometimes a leverage point falls into the pattern of the rest of the points, and thus it is not influential as the line would be the same with or without it. An example of this is illustrated in the first sketch in Figure 6. Sometimes a leverage point can completely determine whether or not there appears to be a linear relationship, as illustrated in the second sketch in Figure 6. Without the leverage point, there does not appear to be a relationship between x and y. But with the point up in the top right, we get a regression line that shows a relationship. So this leverage point is influential. Influential points don't often have large residuals; they often are influential because they draw to point close to them. Be careful not to read too much into a regression line if one point determines its value. We wouldn't want to conclude that there is a linear relationship between the variables, when it is only due to

one single point.

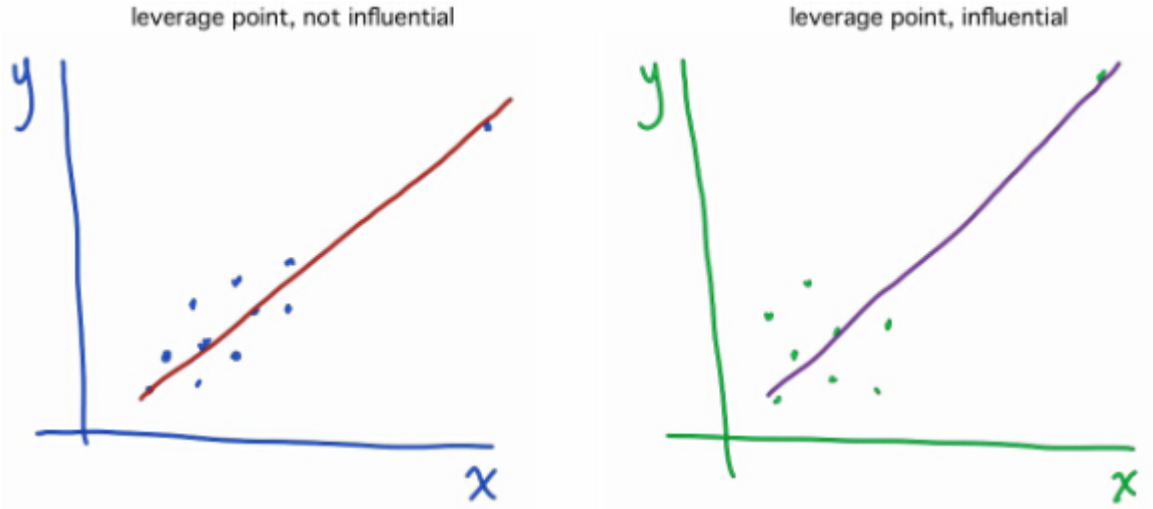

Figure 6: Influential points

Leverage point: A point is a leverage point if it is an outlier in the horizontal direction.

While regression is commonly used, it can also be easily abused. Careful consideration of the range of the data, whether a straight line is appropriate, and whether individual points might will help in making sensible conclusions from linear models fit to data. In summary:

- The intercept may only have a practical interpretation if a value of 0 for the predictor variable is within the range of the data.
- Don't extrapolate beyond the range of the data. The linear relationship may not continue.
- A plot of the residuals versus the predictor variable (or versus the predicted values of the response variable) can show whether a straight line is a good model for the data. When a straight line model is appropriate, the plot should look like random scatter.
- Start with a scatterplot.
- Watch for influential points.