Simple Linear Regression

R^2, The Coefficient of Determination

We'll now take a look at a commonly used statistic for assessing how well a linear model fits some data.

We'll illustrate that statistic in the context of an example, in particular a model for predicting the average crawling age of babies from the average temperature when they were 6 months old. From our 12 data points, one for each month of the year, the crawling age ranged from 28.6 weeks to 33.8 weeks. We could summarize the variability we see in crawling age by its standard deviation:

$$s_y = \sqrt{\tfrac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

We are going to focus on a related quantity, the Total Sum of Squares:

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Sums of squares, play an important role in statistics. SS_{Total} can be broken down into two components: the part of the variation in the y's that can be explained by the regression line and the variation in the y's that is not explained by the line. The variation in the y's that is explained by the regression line is the called the Regression Sum of Squares:

$$SS_{Regression} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The Residual Sum of Squares is the the variation in the y's that the regression line does not explain:

$$SS_{Residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

In order to find our regression line, we find the slope and intercept to give us a line that minimized Residual Sum of Squares. Figure 1 shows the components that go into each of these sums of squaries.

This decomposition of the variation in the y's leads to a statistic that can be used to assess how well the line fits the data.

Definition: The *coefficient of determination*, called R^2, is the proportion of variation in the response variable that is explained by the regression line, that is, the proportion of variation in the y's that is explained by the linear relationship with x.

$$R^2 = SS_{Regression}/SS_{Total} = 1 - SS_{Residual}/SS_{Total}$$

R^2 is a proportion, so it's a number between 0 and 1. A small value of R^2, near 0, means the line is not explaining much of the variation in the response variable. And a large value

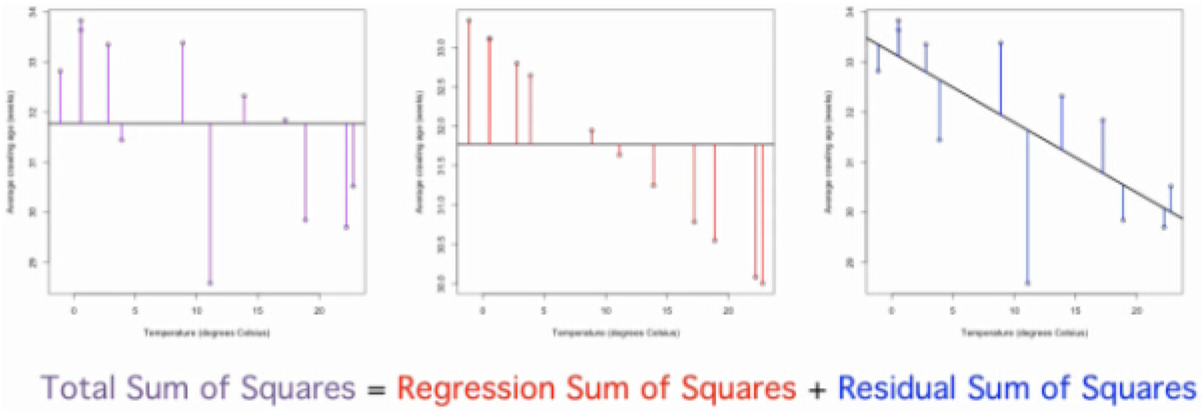

Figure 1: The components of the sums of squares for regression for the babies crawling data. Total Sum of Squares = Regression Sum of Squares + Residual Sum of Squares

of R^2, close to 1, means the line is explaining much of the variation in the response.

Why is the coefficient of determination called R^2? r is the symbol commonly used for the correlation between two quantitative variables and mathematically it can be shown that $R^2 = r^2$.

Example 1
For the babies crawling data (Figure 2) $R^2 = 0.49$. So 49% of the variation in the crawling age is explained by its linear relationship with temperature, and 51% is unexplained. The correlation between average crawling age and temperature is $r = -\sqrt{R^2} = -0.70$. Note that the correlation is negative, because the relationship between crawling age and temperature is negative. R^2, on the other hand, is always positive. So R^2 doesn't indicate whether the relationship is positive or negative.

Example 2
As another example, Figure 3 shows how atmospheric concentration of CFCs was growing over time from 1977 to 1990. The points are very tightly scattered about the regression line so we expect a very high value of R^2. And indeed for this example, $R^2 = 0.996$, so 99.6% of the variability in CFCs is explained by its linear relationship with time. Although the line is explaining almost all of the variability, that doesn't mean that there isn't a better model for these data. For these data we have a systematic pattern of points above then below then above the line. So a model that captures that pattern would be preferred.

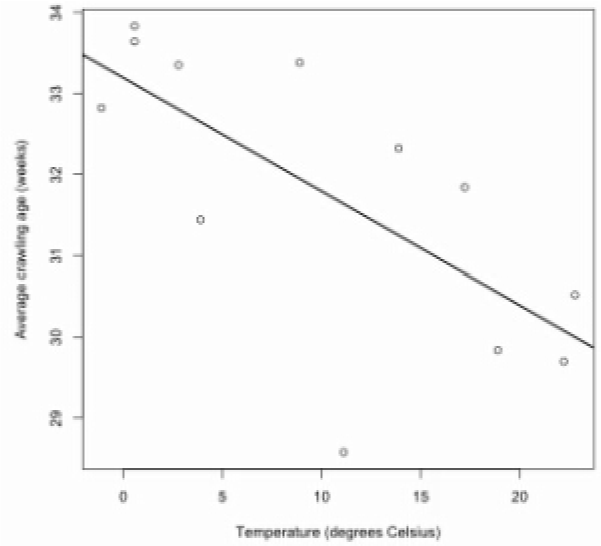

Figure 2: Scatterplot and regression line for the babies crawling data (Example 1).

Cautions when interpreting R^2:

- There are no general rules about how large R^2 should be.
- It is possible to have a large value of R^2 even when the linear model is not ideal.
- R^2 is not meaningful if a linear model is not appropriate.
- It is possible to get very high R^2 by fitting complicated models to data using multiple regression. This is called *overfitting* and the resulting models typically fit poorly to new data.

R^2 gives us one indication about how closely our linear model describes our data but it should never be relied on as the whole story. It is essential to use other tools, such as scatterplots, in order to fully understand the nature of the relationship between two variables.

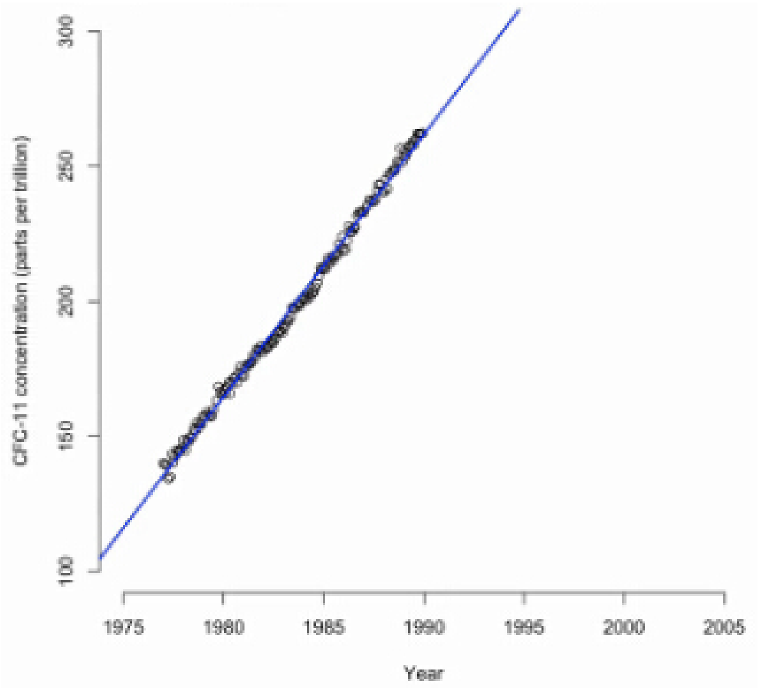

Figure 3: Scatterplot and regression line for the CFCs data (Example 2).