



Some Basics of Simple Linear Regression

by Katrien Declercq

Academic

Assume a club would like to study the fitness of its members based on a measure of oxygen consumption. Besides the oxygen consumption, they also recorded (amongst other measurements) the time to run 1.5 miles. In order to find out how this time measurement influences the oxygen consumption, it is always useful to start by making a scatterplot (using PROC GPLOT) of oxygen consumption (the variable of primary interest, the dependent variable) versus the time needed to run 1.5 miles (the explanatory or independent variable or predictor). This plot shows a linear decreasing trend of oxygen consumption as function of the runtime. So, as the time to run 1.5 miles increases, the fitness of the person decreases.

It would now be of interest to be able to describe the linear relationship between both variables and find out whether the time needed to run 1.5 miles explains a significant amount of variability in the oxygen consumption. Once the best line is found, it can then be used to do predictions of a person's fitness (oxygen consumption) based on the knowledge of his time needed to run 1.5 miles. These things are exactly what a regression analysis can do.

The model can be written as $\text{oxygen_consumption} = \text{intercept} + b * \text{runtime} + \text{error}$. The assumptions are that the oxygen consumption is normally distributed for each value of the runtime, with equal variances and that the responses are independent at each value of the predictor variable and that the mean oxygen consumption is linearly related to the value of runtime. The following SAS code will do a simple linear regression between the two variables in our example.

```
PROC REG DATA=SASUSER.B_FITNESS;  
MODEL OXYGEN_CONSUMPTION=RUNTIME;  
RUN;
```

The output contains a lot of information, from which the most important information will be described here. The best line (least squares method) that fits the data in our sample of 31 club members can be found from the parameter estimates at the bottom of the output. So, the line is estimated as $\text{oxygen consumption} = 82.43 - 3.31 * \text{runtime}$. To find out how much of the variability in the oxygen consumption is explained by this model (with a only explanatory variable the runtime), the R-square may be used. In this case it has a value of 0.74 meaning that 74% of the variability in the oxygen consumption is explained by the runtime. In order to do a formal test to find out whether the runtime explains a significant amount of variability of the oxygen consumption, the P-value next to the parameter for runtime can be used. The corresponding null hypothesis is that the parameter is zero. Since in this case, the P-value is < 0.0001 which is very significant, we can conclude that runtime explains a significant amount of variability of the response (and the parameter is significantly different from 0).

Of course, these are only the very basics and if more measurements are available, a better model can be constructed to explain the variability in the response. A battery of techniques exists to find good models to fit the data.