StatNews #07

**Assessing Logistic Regression**
August 1996
Revised: May 2012

In many different situations we have to rely on indicators to assess the existence of a specific condition that cannot itself be measured. For example, in nutritional sciences, indicators such as weight/height and skinfold are used to assess the presence of malnutrition or obesity. When an indicator only has two different responses possible, the following table can be set up to illustrate its accuracy:

```
                TRUTH
                 Yes    |     No
               _____|_____
      High   |          |          |
             |    a     |    b     |
 INDICATOR   |_____|_____|
             |          |          |
      Low    |    c     |    d     |
             |_____|_____|
```

Sensitivity is defined as the ratio a/a+c and specificity as d/b+d. To be useful, an indicator should have high sensitivity and specificity.

We often encounter indicators that can take on a continuous range of values. In this case, we can generate a large number of these tables by choosing different cut-off points between "high" and "low". Each of these tables is then accompanied by a measure of sensitivity and specificity. When sensitivity is plotted against 1-specificity we obtain a curve which is called an ROC (Receiver Operating Characteristic) curve.

```
         |      * /
         |    *  /
         |  *   /
 sens    | *   /
         |* /
         |*/
         |/_____
            1-spec
```
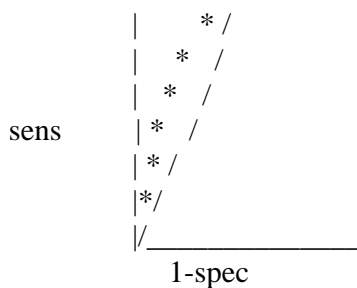
The diagonal line represents chance. A curve that is well above the diagonal line means that an indicator is accurate. Graphing an ROC curve gives a good visual representation of accuracy, but often a numerical measure of accuracy is useful as well. Several different measures of accuracy have been developed, but the easiest one is the area under the ROC curve. This measure will vary between 0.5 and 1. An area of 0.5 represents the diagonal, attained when no discrimination exists or by chance alone. An area of 1 represents the perfect indicator.

In logistic regression, ROC curves are very useful for evaluating the predictive accuracy of a chosen model. The predicted values generated by the logistic model can be viewed as a continuous indicator to be compared to the observed binary response variable.

Most commonly used statistical software packages will give you ROC curves and the corresponding area under the curve.   For example, in the statistical software STATA, you can request the ROC curve and the area under the curve after running a logistic regression. In SAS, the area under the curve is labeled "c", and is given in the proc logistic output . In SPSS you can save the predicted probabilities and then plot the ROC curve and obtain the area under the curve in the Graphs>ROC Curves menu.

A good reference is Swets John A. (1988). Measuring the Accuracy of Diagnostic Systems.   Science, Vol 240, 1285-1293.

Do not hesitate to contact us in case you need to use one of these programs or if you would like more information.

Author: Francoise Vermeylen