A Nonparametric Lack-of-Fit Test Using Arc

Sanford Weisberg
Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108-6042.
Supported by National Science Foundation Grant DUE 96-52887

February 24, 2000, slightly revised October 2, 2001

Abstract

Arc is a computer program for the analysis of regression problems. It is described in Cook and Weisberg (1999). This paper describes an *Arc* add-on called `nptest.lsp` that uses a bootstrap procedure described by Hart (1997, p. 150) to obtain a nonparametric lack-of-fit test for the fit of a linear model to a 2D graph.

Arc is a menu driven computer program of the analysis of regression data, as described in Cook and Weisberg (1999). This paper describes an *Arc* add-on to perform a bootstrap for a nonparametric lack-of-fit test of a polynomial linear regression model to a 2D graph.

1 Getting the Add-on

The file `nptest.lsp` is available from www.stat.umn.edu/arc. To use it, download the correct file for your system, and put it in your Extras directory in the same directory as the *Xlisp-Stat* program. The add-on will be loaded automatically every time you start *Arc.* If you want to remove it, simply rename it so its name does not end in .lsp, or move it to another directory.

A conflict between nptest.lsp and the *Arc* updates.lsp file was corrected on October 2, 2001. If the menu item described below does not appear when expected, you need to get the newer version of nptest.lsp that fixes this problem.

2 Example

Load the file `lakemary.lsp`. Shown in Figure 1 is a plot of *Length* versus *Age* for a sample of 78 bluegills captured in Lake Mary, Minnesota. Also shown on the graph are two estimates of the mean function for these data: a straight line estimated via OLS from the OLS slide bar, and a *lowess* smooth, with smoothing parameter 0.6. The first of these is based on the assumption that simple linear regression is appropriate for these data, while the second requires no assumptions at all. One might hope that an hypothesis test of the null hypothesis that the simple linear regression model is appropriate versus a general alternative could be tested by comparing these two fits;

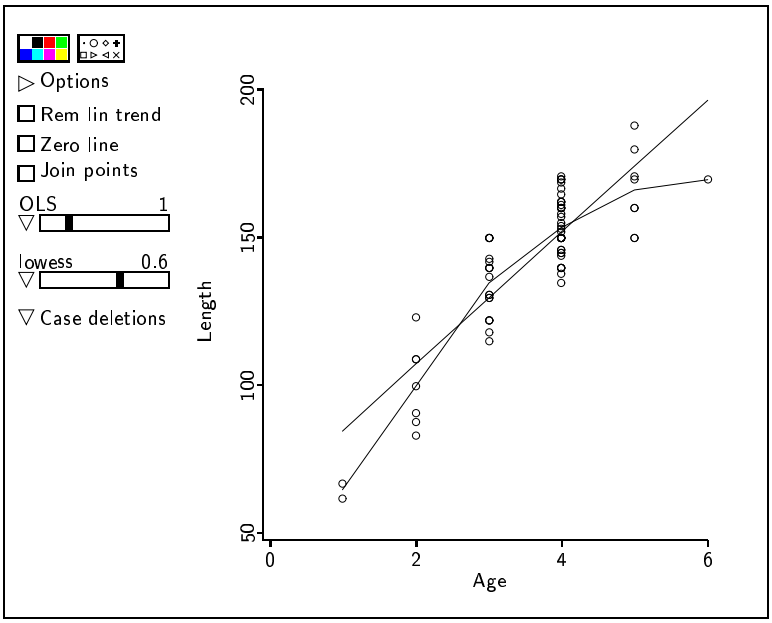

Figure 1: The Lake Mary data, with the fit of the parametric model and the fit of the *lowess* smooth.

we would have evidence against simple linear regression if the *lowess* fit is much better than the fit under the null hypothesis.

To test this hypothesis, select the item "Lack-of-fit test" from the OLS slide bar's pop up menu. You will then get the following output printed in the text window:

```
Computing bootstrap significance level...
[Plot2]LakeMary:V:Length H:Age
Lack of fit p-value = 0.01 based on 99 bootstraps.
```

The program computed a test statistic for this hypothesis test, and then to obtain a significance level, a simulation called a *bootstrap* was performed. For all 99 simulations, the value of the test obtained was smaller than the value for the actual data giving a significance level of $1/(99+1) = 0.01$; the significance level is random, and will vary from use to use.

You can change the number of bootstraps by first selecting Settings from the Arc menu, selecting the item `*nptest-num-bootstraps*` from the settings list, and then selecting the item Update selection from the Settings menu.

3 Details

1. Fit the parametric model specified by the slide bar. This is a polynomial regression model that will use weights if they are specified when setting up the graph. Obtain the residuals $\hat{e}$ from this parametric fit.

2. Smooth the 2D graph of $\hat{e}$ versus the horizontal axis in the plot, using the same smoother and smoothing parameter as shown on the graph. If the null hypothesis were true, then this smooth should approximate a straight line of slope and intercept zero.

3. Let $\hat{g}_i$ be the fitted values from the smooth in step 2. We will have evidence against the null hypothesis if $\sum \hat{g}_i^2$ is too large. Compute the statistic (Hart, 1997, p. 150)

$$R_s = \frac{n^{-1} \sum_{i=1}^{n} \hat{g}_i^2}{\tilde{\sigma}^2}$$

where $\tilde{\sigma}^2$ is a "model-free" estimate of variance. (Hart, 1997, p. 129)

$$\tilde{\sigma}^2 = \frac{1}{a_n} \sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2$$

and a_n is a normalizing constant given by Hart. We use the normalizing constant in the add-on that assumes that a simple regression model is fit, so this will be slightly in error for higher-order polynomials.

4. A significance level for R_s is computed using a bootstrap procedure as follows (Hart, 1997, Sec. 5.4.3). Repeat the following B times:

   (a) Draw a sample of size n with replacement from $\hat{e}_1, \ldots \hat{e}_n$.

   (b) Repeat Steps 2-3 above, substituting the resampled residuals for the observed residuals, to obtain a resampled value of R_s^*.

   We now have a sample of B values of R_s^*. The significance level is then the rank of R_s among the R_s^* divided by $B + 1$.

4 Limitations

This code works only for linear models, not for any other generalized linear model. The value of the smoothing parameter is as specified on the slidebar, so there is no optimization over the smoothing parameter. The method requires that the true variance function is constant.

5 References

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer-Verlag.