Linear Regression Analysis for STARDEX

Malcolm Haylock, Climatic Research Unit

The following document is an overview of linear regression methods for reference by members of STARDEX. While it aims to cover the most common and relevant methods for calculating trends and their levels of statistical significance, there will inevitably be omissions. Please send any corrections or comments to Malcolm Haylock (M.Haylock@uea.ac.uk).

Trend Calculation

Least squares (used in diagnostic tool)

Least squares linear regression is a maximum likelihood estimate i.e. given a linear model, what is the likelihood that this data set could have occurred? The method attempts to find the linear model that maximises this likelihood.

Suppose each data point y_i has a measurement error that is independently random and normally distributed around the linear model with a standard deviation σ_i.

The probability that our data (± some fixed Δy at each point) occurred is the product of the probabilities at each point:

$$P \alpha \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - (a + bx_i)}{\sigma_i} \right)^2 \right] \Delta y \right\}$$

Maximising this is equivalent to minimising:

$$\sum \left(\frac{y_i - (a + bx_i)}{\sigma_i} \right)^2$$

If the standard deviation σ_i at each point is the same, then this is equivalent to minimising:

$$\sum (y_i - (a + bx_i))^2$$

Solving this by finding a and b for which partial derivatives with respect to a and b are zero, gives the best fit parameters for the regression constant and coefficient (α and β):

$$\alpha = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}$$

$$\beta = \frac{N S_{xy} - S_x S_y}{\Delta}$$

where $\Delta = N S_{xx} - (S_x)^2$

and $S_x = \sum x_i$, $S_y = \sum y_i$, $S_{xy} = \sum x_i y_i$, $S_{xx} = \sum x^2$

For further information, see Wilks (1995) or Press *et al.* (1986).

Minimum Absolute Deviation

Least squares linear regression, like many statistical techniques, assumes that the departures from the linear model (errors) are normally distributed. Techniques that do not rely on such assumptions are termed *robust*.

Least squares regression is also sensitive to outliers. Although most of the errors may be normally distributed, a few points with large errors can have a large affect on the estimated parameters. Techniques that are not so sensitive to outliers are termed *resistant*.

A more resistant method for linear trend analysis is to assume that the errors are distributed as a two-sided exponential. This distribution, with its larger tails, allows a higher probability of outliers:

$$\Pr\{y_i - (a + bx_i)\} \sim \exp(-|y_i - (a + bx_i)|)$$

Similarly to the process of least squares, this requires that we minimise:

$$\sum_{i=1}^{N} |y_i - (a + bx_i)|$$

The solution to this needs to be found numerically. Example code can be found in Press *et al.* (1986).

Three-group resistant line

This method derives its resistance from the fact that one of the simplest resistant measures of a sample is the median.

Data are divided into three groups depending on the rank of the x values. The *left* group contains the points with the lowest third of x values. In a time series this is equivalent to the first third of the series. Similarly, the *middle* and *right* groups contain points with the middle and highest third of ranked x values respectively.

Next the x and y median values are determined for the three groups to give the points (x_L, y_L), (x_M, y_M) and (x_R, y_R).

The slope of the line is taken as the gradient of the line through the medians of the left and right groups:

$$b_0 = \frac{y_R - y_L}{x_R - x_L}$$

The intercept of the line is calculated by finding the three lines with slope b_0 that pass through each of the points (x_L, y_L), (x_M, y_M) and (x_R, y_R), then averaging their intercept:

$$a_0 = \frac{1}{3}[(y_L - b_0 x_L) + (y_M - b_0 x_M) + (y_R - b_0 x_R)]$$

The three-group resistant line method usually requires iteration. After the first pass to find a_0 and b_0, the process can be repeated on the residuals to find a_1 and b_1. The iterations are continued until the adjustment to the slope is sufficiently small in magnitude (at most 1%). The final slope and intercept is the sum of those from each iteration.

Further information can be found in Hoaglin *et al.* (1983).

Logistic Regression

Linear regression has been generalised under the field of generalised linear modelling, of which logistic regression is a special case. This method utilises the binomial distribution and can therefore be used to model counts of extreme events.

Often in a series, the variance of the residuals (from the linear model) varies with the magnitude of the data. This goes against the assumptions of least squares regression, which assumes residuals to have constant variance, but is a natural element of the binomial distribution and logistic regression. Therefore data do not need to be normalised.

The logistic regression model expresses the probability π of a success (e.g. an event above a particular threshold) as a function of time:

$$\eta(\pi) = \alpha + \beta \cdot t$$

Since the probability of a success is in the range [0,1], it needs to be transformed to the range $(-\infty, \infty)$ using a link function:

$$\eta(x) = \log\left(\frac{x}{1-x}\right)$$

Solving for π gives:

$$\pi(t;\alpha,\beta) = \frac{e^{\alpha+\beta\cdot t}}{1+e^{\alpha+\beta\cdot t}}$$

We are not fitting a straight line to the counts and therefore can not refer to a single trend value. The odds ratio is used to express the relative change in the ratio of events to non-events over the period (t_1, t_2):

$$\Theta \equiv \frac{\pi(t_2)}{1-\pi(t_2)} \Big/ \frac{\pi(t_1)}{1-\pi(t_1)} = e^{\beta\cdot(t_2-t_1)}$$

Model fitting can be done using a maximum likelihood method.

Further information about logistic regression, together with an example using extreme precipitation in Switzerland, can be found in Frei and Schär (2001).

Significance Testing

Confidence intervals for least squares

Often the standard deviations σ_i for the observations are not known. If we assume that the linear model does fit well and that all observations have the same standard deviation σ, the assumption that the residuals are normally distributed around the linear model implies that:

$\sigma^2 = \frac{\sum(y_i - (a+bx_i))^2}{N-2}$, with *N-2* appearing in the denominator because two parameters are estimated.

From the above, it can be shown that the regression coefficient *b* will be normally distributed with variance:

$$Var[b] = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Since the variance of b is estimated, Student's t-distribution is used to define the multiplier *t* for the confidence limits for the regression coefficient:

$$b = \beta \pm t\sqrt{Var(b)}$$

The assumption that the residuals are normally distributed can be tested with a quantile-quantile (Q-Q) plot of the residuals against the quantiles from a Gaussian distribution.

For further information, see Wilks (1995) or Press *et al.* (1986).

Linear Correlation

The *linear correlation coefficient* (Pearson product-moment coefficient of linear correlation) is used widely to assess relationships between variables and has a close relationship to least squares regression.

The correlation coefficient is defined by:

$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$ i.e. the ratio of the covariance of x and y to the product of their standard deviations.

In a least squares linear model, the variance of the predictand can be proportioned into the variance of the regression line and the variance of the predictand around the line:

SST=SSR+SSE

Sum of Squares Total = Sum of Squares Regression + Sum of Squares Error

In a good linear relationship between the predictor and predictand, SSE will be much smaller than SSR i.e. the spread of points around the line will be much smaller than the variance of the line. This goodness of fit can be described by the coefficient of determination:

$R^2 = \frac{SSR}{SST}$ = variance of predictand *explained* by the predictor

It can be shown that the coefficient of determination is the same as the square of the correlation coefficient. The correlation coefficient can therefore be used to assess how well the linear model fits the data. Assessing the significance of a sample correlation is difficult, however, as there is no way to calculate its distribution for the null hypothesis (that the variables are not correlated). Most tables of significance use the approximation that, for a small number of points and normally distributed data, the following statistic is distributed for the null hypothesis like Student's t-distribution:

$$t = r\sqrt{\frac{N-2}{1-r^2}} \qquad (1)$$

The common basis of the correlation coefficient and least squares linear regression means that they share the same shortcomings such as limited resistance to outliers.

See Wilks (1995) or Press *et al.* (1986) for further information.

Spearman rank-order correlation coefficient

Non parametric correlation statistics are an attempt to overcome the limited resistance and robustness of the linear correlation coefficient, as well as the uncertainty in determining its significance.

If x and y data values are replaced by their rank, we are left with the set of points (i,j), i,j=1,N which are drawn from an accurately known distribution. Although we are ignoring some information in the data, this is far outweighed by the benefits of greater robustness and resistance.

The Spearman rank-order correlation coefficient is just the correlation coefficient of these ranked data. Significance is tested as for the linear correlation coefficient using (1), but in this case the approximation does not depend on the distributions of the data.

See Press *et al.* (1986) for further information.

Kendall-Tau (used in diagnostic tool)

Kendall's Tau differs from the Spearman rank-order correlation in that it only uses the relative ordering of ranks when comparing points. It is calculated over all possible pairs of data points using the following:

$$\tau = \frac{concordant - discordant}{\sqrt{concordant + discordant + sameX}\sqrt{concordant + discordant + sameY}}$$

where *concordant* is the number of pairs where the relative ordering of x and y are the same, *discordant* where they are the opposite, *sameX* where the x values are the same and *sameY* where the y values are the same.

τ is approximately normally distributed with zero mean and variance:

$$Var(\tau) = \frac{4N + 10}{9N(N-1)}$$

One advantage of Kendall's tau over the Spearman coefficient is the problem of assigning ranks when data are tied. Kendall's tau is only concerned whether a rank is higher or lower than another, and can therefore be calculated by comparing the data themselves rather than their rank. When data are limited to only a few discrete values, Kendall's tau is a more suitable statistic.

See Press *et al.* (1986) for further information.

Resampling

Resampling procedures are used extensively by climatologists and could be used to assess the significance of a linear trend. The bootstrap method involves randomly resampling data (with replacement) to create new samples, from which the distribution of the null hypothesis can be estimated. Therefore no assumption needs be made about the sample distribution. If enough random samples are generated, the significance of an observed linear trend can be assessed by where it appears in the distribution of trends from the random samples.

A problem, however, is that the maximum likelihood derivation of the least squares estimate for the linear trend assumed that data residuals about the line were normally distributed. Therefore if the distribution of the residuals is not Gaussian, then the least squares estimate is not valid. Still, bootstrapping could be used to test the significance of a least squares linear trend, given that this may not be the best trend estimate.

An important assumption in resampling is that observations are independent. Zwiers (1990) showed that, for the case of assessing the significance of the difference in two sample means, the presence of serial correlation greatly affected the results. A method has been proposed by Ebisuzaki (1997) whereby random samples are taken in the frequency domain (with random phase) to retain the serial correlation of the data in each sample.

References

Ebisuzaki, W., 1997: A method to estimate the statistical significance of a correlation when the data are serially correlated. *J. Clim.*, **10**, 2147–2153.

Frei, C. and C. Schär, 2001. Detection probability of trends in rare events: theory and application to heavy precipitation in the alpine region. *J. Clim.*, **14**, 1568-1584.

Hoaglin, D.C., F. Mosteller and J.W. Tukey, 1983. *Understanding robust and exploratory data analysis*. Wiley. 129-165.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, 1986. *Numerical recipes: The art of scientific computing.* Cambridge Univ. Press, 488-493.

Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences.* Academic Press. 160-176.

Zwiers, F. W., 1990. The effect of serial correlation on statistical inferences made with resampling procedures. *J. Clim.*, **3**, 1452-1461.