

GLOSSARY OF TECHNICAL TERMS

A2 and B2 scenarios:

The A2 scenario envisions that global population will reach 15 billion by the year 2100 and that economic and technological development will be rather slow. It projects slightly lower GHG emissions than the IS92a scenario, but also slightly lower aerosol loadings, such that the warming response differs little from that of the earlier scenario. The B2 scenario envisions slower global population growth (10.4 billion by 2100) with a more rapidly evolving economy and more emphasis on environmental protection. It therefore produces lower emissions and less future warming. A2 and B2 are scenarios as described in the IPCC Special Report on Emission Scenarios (refer to SRES below) (Environment Canada 2004d).

Average linkage clustering procedure:

Average linkage clustering procedure is one of the hierarchical clustering procedures. In this method, the distance between two clusters is the average distance between pairs of observations, one in each cluster. Average linkage tends to join clusters with small variances, and it is slightly biased toward producing clusters with the same variance (SAS Institute Inc. 1999a). This procedure attempts to minimize the within-cluster variances and maximize the between-cluster variances (Boyce 1996, Cheng and Lam 2000).

Centroid clustering procedure:

Centroid clustering procedure is one of the hierarchical clustering procedures. In this method, the distance between two clusters is defined as the Euclidean distance between their centroids or means. The centroid method is more robust to outliers than most other hierarchical methods but in other respects may not perform as well as Ward's method or average linkage (SAS Institute Inc. 1999a).

Cumulative logit regression:

Logistic regression is also applicable to multilevel responses. The response may be ordinal or nominal. For ordinal response outcomes, you can model functions called *cumulative logits* by performing ordered logistic regression using the proportional odds model (McCullagh 1980). For nominal response outcomes, it is possible to model *generalized logits* and perform a logistic analysis, except that this involves modeling multiple logits per subpopulation (Stokes et al. 2000).

SAS Institute Inc. (2004) provides very useful information for understanding of cumulative logit regression as follows.

PROC CATMOD can compute three different types of logits with the use of keywords in the RESPONSE statement. Other types of response functions can be generated by specifying appropriate transformations in the RESPONSE statement.

- Generalized logits are used primarily for nominally scaled dependent variables, but they can also be used for ordinal data modeling. Maximum likelihood estimation is available for the analysis of these logits.

- Cumulative logits are used for ordinally scaled dependent variables. Except for dependent variables with two response levels, only weighted least-squares estimation is available for the analysis of these logits.
- Adjacent-category logits are equivalent to generalized logits, but they have some advantages for ordinal data analysis because they automatically incorporate integer scores for the levels of the dependent variable. Except for dependent variables with two response levels, only weighted least-squares estimation is available for the analysis of these logits (SAS Institute Inc. 2004).

Discriminant function analysis (one of the nonhierarchical classification methods):

Discriminant function analysis utilizes the mean vectors and the covariance matrix, based on differences and relationship between the selected variables of different groups, to develop discriminant classification functions. The functions are in turn used to assign each individual day to the most appropriate group (Cheng 1995).

EOF (Empirical Orthogonal Functions) See PCA

Hierarchical clustering: Hierarchical clustering joins the most similar observations, then successively connects the next most similar observations to these. First an $n \times n$ matrix of similarities between all pairs of observations is calculated. Those pairs having the highest similarities are then merged, and the matrix recomputed. This is done by averaging the similarities that the combined observations have with other observations. The process iterates until the similarity matrix is reduced to 2×2 . The levels of similarity at which observations are merged are used to construct a dendrogram (Davis 1986).

Hypothesis test:

A procedure for deciding between two hypotheses on the basis of the value of a statistic called the test statistic, which is a function of the observations in a random sample (Upton and Cook 2002).

The t -test, χ^2 -test, semi-partial R^2 , pseudo- F , and pseudo- t^2 tests are used in this study. The t -test, based on the t probability distribution, is useful for establishing the likelihood that a given sample could be a member of a population with specific characteristics, or for testing hypotheses about the equivalency of two samples (Davis 1986).

The χ^2 -test is used to test if a sample of data came from a population with a specific distribution (Snedecor and Cochran 1989). An attractive feature of the chi-square goodness-of-fit test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The chi-square goodness-of-fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the χ^2 -test. However, the values of the χ^2 -test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid. The χ^2 -test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to

continuous distributions (<http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm>, accessed November 2004).

The semi-partial R^2 is the ratio of the increased within-cluster variance after joining two clusters to the variance for the entire dataset. The pseudo- F is the ratio of between-cluster to within-cluster variances. The pseudo- t^2 is the ratio of the increased within-cluster variance after joining two clusters to the variance within each of two clusters (Cheng et al. 2004).

IS92a scenarios:

The IPCC IS92a scenario specifies equivalent greenhouse gas (GHG) concentrations and sulphate aerosol loadings from 1850 to 2100. The scenario has effective CO₂ concentration increasing at 1% per year after 1990. In the model, the concentrations are specified by linear interpolation between specified values at 2000, 2025, 2050 and 2100. Climate change simulations based on this scenario have been performed by a number of climate modeling groups who have contributed to the IPCC Third Assessment Report (Environment Canada 2004d).

Linear regression:

Regression techniques enable you to investigate the relationship between a dependent variable (also called a *response* variable) and one or more explanatory variables (also called *predictor*, or *independent*, variables). In linear regression, the dependent variable is modeled as a linear function of the quantitative independent variables. For example, you can write the simple linear regression equation as

$$Y = b_0 + b_1 X$$

where Y represents the single dependent variable, X is the explanatory variable, and b_0 and b_1 are regression coefficients (SAS Institute Inc. 2004).

Logistic regression model:

Logistic regression enables you to investigate the relationship between a categorical outcome and a set of explanatory variables. The outcome, or response, can be dichotomous (yes, no) or ordinal (low, medium, high). When you have a dichotomous response, you are performing standard logistic regression. When you are modeling an ordinal response, you are fitting a proportional odds model (SAS Institute Inc. 2004).

Multiple stepwise regression:

The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the F statistic for a variable to be added must be significant at the SLENTY= level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an F statistic significant at the SLSTAY= level. Only after this check is made and the necessary deletions accomplished can another variable be added to the model. The stepwise process ends when none of the variables outside the model has an F statistic significant at the SLENTY= level and every variable in the model is significant at the SLSTAY= level, or when the variable to be added to the model is the one just deleted from it (SAS Institute Inc. 1999a).

The forward-selection technique begins with no variables in the model. For each of the independent variables, the FORWARD method calculates F statistics that reflect the variable's contribution to the model if it is included. The p -values for these F statistics are compared to the SLENTY= value that is specified in the MODEL statement (or to 0.50 if the SLENTY= option is omitted). If no F statistic has a significance level greater than the SLENTY= value, the FORWARD selection stops. Otherwise, the FORWARD method adds the variable that has the largest F statistic to the model. The FORWARD method then calculates F statistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant F statistic. Once a variable is in the model, it stays (SAS Institute Inc. 1999a).

The backward elimination technique begins by calculating F statistics for a model, including all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce F statistics significant at the SLSTAY= level specified in the MODEL statement (or at the 0.10 level if the SLSTAY= option is omitted). At each step, the variable showing the smallest contribution to the model is deleted (SAS Institute Inc. 1999a).

Orthogonal regression or principal component regression:

Orthogonal regression or principal component regression is a simple extension of multiple linear regression and principle component analysis. In the first step, the principal components are calculated. The scores of the most important principal components are used as the basis for the multiple linear regression with the target data y (Lohninger 1999). The advantage of this method can produce more accurate estimates than other regression procedures for very “ill-conditioned” data (collinearity) (SAS Institute Inc. 1999a).

Principal components analysis (PCA):

A technique for making multivariate data is easier to understand. The idea of PCA is to replace the original intercorrelated n variables by m ($\leq n$) uncorrelated component variables, each of which is a linear combination of the original variables, so that the bulk of the variation can be accounted for using just a few explanatory variables (Upton and Cook 2002).

Poisson regression:

Poisson regression is one type of regression method. It is appropriate when the mean function and the variance function are equal, as will occur if y is a count and $y | \mathbf{x}$ has a Poisson distribution (Cook and Wiesberg 1999). An advantage of Poisson regression is that it can be precisely tailored to the highly-skewed distribution of the dependent variable, a disadvantage is that it is susceptible to over-dispersion problems that do not affect ordinary regression (Allison 2000).

R^2 :

R^2 also called multiple correlation or the coefficient of multiple determination is the percent of the variance in the dependent explained uniquely or jointly by the independents (<http://www2.chass.ncsu.edu/garson/pa765/regress.htm>, accessed November 2004).

Robust regression:

The main purpose of robust regression is to detect outliers and provide resistant (stable) results in the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers. Historically, three classes of problems have been addressed with robust regression techniques:

- problems with outliers in the y -direction (response direction)
- problems with multivariate outliers in the x -space (i.e., outliers in the covariate space, which are also referred to as leverage points)
- problems with outliers in both the y -direction and the x -space (SAS Institute Inc. 2004).

SRES Scenarios:

The IPCC Special Report on Emission Scenarios (SRES) provides 40 different scenarios which are deemed “equally likely”. For the Third Assessment Report, the IPCC facilitated the conversion of two of these emission scenarios (A2 and B2) into concentration scenarios for use in climate simulations (Environment Canada 2004d).

Ward’s clustering:

Ward’s clustering procedure is one of the hierarchical clustering procedures. It tends to join clusters with a small number of observations, and it is strongly biased toward producing clusters with roughly the same number of observations. It is also very sensitive to outliers (SAS Institute Inc. 1999a).