Valuing Environmental Impacts: Practical Guidelines for the Use of Value Transfer in Policy and Project Appraisal

Annex 3 – Glossary of Econometric Terminology

Submitted to

Department for Environment, Food and Rural Affairs

February 2010

eftec
73-75 Mortimer Street
London W1W 7SQ
tel: 44(0)2075805383
fax: 44(0)2075805385
eftec@eftec.co.uk
www.eftec.co.uk

Report prepared for the Department for Environment, Food and Rural Affairs

by:

Economics for the Environment Consultancy (eftec)
73 – 75 Mortimer St, London, W1W 7SQ
Tel: 020 7580 5383
Fax: 020 7580 5385
www.eftec.co.uk

Study team (in alphabetical order):

Ian Bateman (University of East Anglia)
Roy Brouwer (Institute for Environmental Studies, VU University, Amsterdam)
Matthew Cranford (eftec)
Stephanie Hime (eftec)
Ece Ozdemiroglu (eftec)
Zara Phang (eftec)
Allan Provins (eftec)

Acknowledgements:

The study team would like to thank Prof. Stale Navrud (Norwegian University of Life Sciences), Prof. Ken Willis (University of Newcastle upon Tyne) and members of the Steering Group for their comments on the previous versions of the Guidelines, Case Studies and Technical Report.

eftec offsets its carbon emissions through a biodiversity-friendly voluntary offset purchased from the World Land Trust (http://www.carbonbalanced.org) and only prints on 100% recycled paper.

ANNEX 3: GLOSSARY OF ECONOMETRIC TERMINOLOGY

- *This annex is intended to help analysts interpret results presented by primary valuation studies.*
- *Key terms from econometric analysis are explained, including different model types, interpretation of results and model statistics.*

Introduction – glossary of basic econometric terminology

The purpose of this annex is to provide a brief review of key econometric terminology to assist analysts in interpreting statistical and econometric results reported by primary valuation studies. The summary covers:

- Some 'basics' of statistical and econometric analysis;
- Types of econometric models;
- Descriptive statistics of models; and
- Interpretation of model coefficients.

Some basics

Econometric analysis: In relation to economic valuation of non-market goods and services, econometric analysis focuses on identifying the separate effects of different factors that act in common to determine the economic value generated by a particular good or service. This is relevant to revealed preference (e.g. hedonic pricing, travel cost and multi-site recreation demand models), stated preference (e.g. contingent valuation, choice experiments) and other valuation methods such as the production function approach. Typically analysis is via regression techniques that enable a statistical analysis of the quantitative relationship (correlation) between one or more independent variables and one dependent variable. Approaches and methods are determined by the nature of the data analysed.

Sampling: All economic analyses attempt to make conclusions about a particular population. It is infeasible to survey everyone in a given population, so a sample of that population is used for research purposes. A sample should be representative: it should have similar socio-economic characteristics as the entire population (e.g. if population is 53 percent female, sample should ideally be 53 percent female). There are a few different sampling methods:

- ***Simple random sampling*** – When the sample is wholly randomly chosen from the population.
- ***Quota/probability sampling*** – When information about the population is known and a sample is collected to match that (e.g. if the population is known to be 53 percent female, the sample is explicitly chosen to be 53 percent female).
- ***Stratified sampling*** – If sub-populations vary significantly, the sample should be weighted to proportionally reflect the mixture of subpopulations within the population.

Variables: A variable is a value that may vary and in the context of economics refers to any characteristic of the subject matter that may vary (e.g. willingness-to-pay, income level of respondent, etc.):

- ***Dependent variable*** – The variable of interest (e.g. household WTP): a variable that is observed to arise based on the levels of all independent variables.
- ***Independent variable*** – Variables that are considered changeable and affecting the dependent variable.

Transformation of variables: Often there is not a linear relationship between the dependent and independent variables as they are measured, but there may still be a tangible relationship. In such cases, variables can be transformed by applying a deterministic function such as squaring or taking the log of a variable to explore a different functional relationship (besides linear) between the dependent and independent variables. One of the most common transformations is a log transformation used for dealing with variables that occur on very different scales:

- ***Semi-log transformation*** – The log of one variable (independent or dependent) is taken and used instead of the initial observations.
- ***Double-log transformation*** – The log of both the independent and dependent variables are used in analysis.

Continuous variable: A variable which in the simplest term can take 'any' value. For example, survey respondent's age or distance from a respondent's home to a visitor site.

Categorical variable: A variable which can take a value from a finite set of values. Survey questions that require respondents to answer on the basis of a Likert scale result in categorical variables.

Dummy variable: Categorical variables believed to influence the outcome of a regression model can be coded as 1 or 0 for existing or not existing respectively. Inclusion of dummy variables can help to increase the fit of a model, but at a loss of generality of the model (i.e. more dummy variables means a more case-specific model). If multiple categorical variables are related, one must be excluded in the regression analysis, to act as the baseline from which the effect of the other categories is measured.

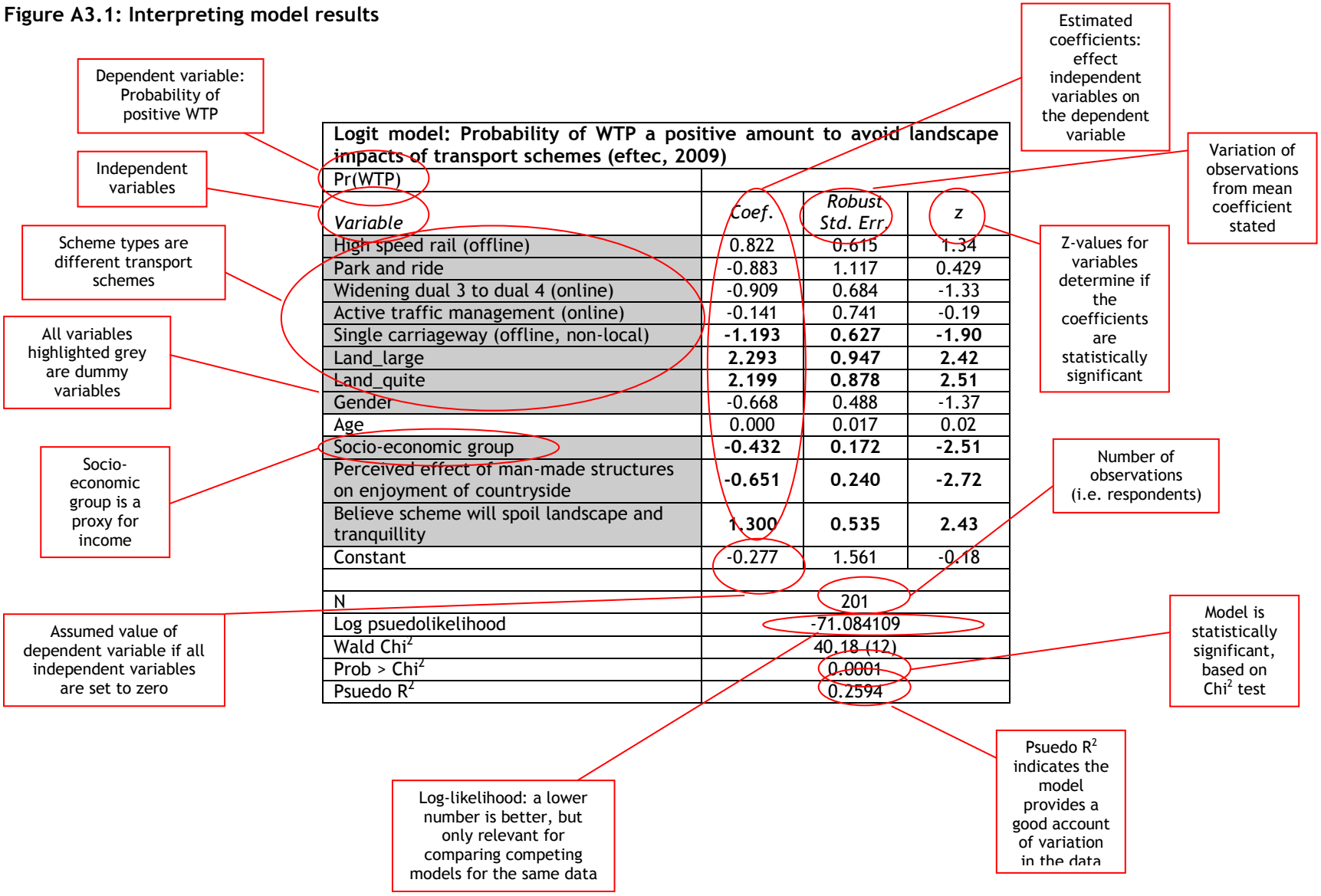
An example

Figure A3.1 shows a typical table reporting econometric analysis results from a stated preference study[1]. The components of the estimated model are labelled along with a basic interpretation of the key findings:

- The higher the socio-economic group, the higher the probability of paying.
- The greater the perceived negative impact of the scheme, the higher the probability of paying.
- The more negative the perceived effect of man-made structures on the enjoyment of the countryside, the higher the probability of paying.

[1] Source study: eftec (2009) *Valuing Transport's Impact on the Natural Landscape*: *Phase 2 Final Report*, Report to the Department for Transport.

Figure A3.1: Interpreting model results



Logit model: Probability of WTP a positive amount to avoid landscape impacts of transport schemes (eftec, 2009)

Pr(WTP)			
Variable	*Coef.*	*Robust Std. Err.*	*z*
High speed rail (offline)	0.822	0.615	1.34
Park and ride	-0.883	1.117	0.429
Widening dual 3 to dual 4 (online)	-0.909	0.684	-1.33
Active traffic management (online)	-0.141	0.741	-0.19
Single carriageway (offline, non-local)	**-1.193**	**0.627**	**-1.90**
Land_large	**2.293**	**0.947**	**2.42**
Land_quite	**2.199**	**0.878**	**2.51**
Gender	-0.668	0.488	-1.37
Age	0.000	0.017	0.02
Socio-economic group	**-0.432**	**0.172**	**-2.51**
Perceived effect of man-made structures on enjoyment of countryside	**-0.651**	**0.240**	**-2.72**
Believe scheme will spoil landscape and tranquillity	**1.300**	**0.535**	**2.43**
Constant	-0.277	1.561	-0.18
N		201	
Log psuedolikelihood		-71.084109	
Wald Chi^2		40.18 (12)	
Prob > Chi^2		0.0001	
Psuedo R^2		0.2594	

- Respondents who agree the scheme will "spoil the landscape and tranquillity" are more likely to pay.
- Scheme type, age and gender are not found to have a statistically significant effect on the probability of paying some amount to prevent the landscape impact.

The model is statistically significant with a Chi^2 test statistic of 0.0001, meaning that there is a 0.0001 percent probability that the model fit occurred by chance. Additionally the psuedo-R^2 value is 0.2594, which is an acceptable level of fit for a logit model.

Econometric models

Bi-variate regression: The 'simplest' regression model, which predicts the change in a dependent variable based on the effect of a single independent variable.

Multiple regression: Predicts the change in a dependent variable based on the effects of multiple independent variables.

Ordinary least squares (OLS) regression: OLS regression is a linear model. The dependent variable is continuous and the independent variables are generally continuous, although there are methods of including categorical variables as dummy variables. Additionally, the linear relationship can be modified by transforming variables. For example, the relationship between income and value of water quality may not be significant, but the relationship between the square of income and value of water quality may be. OLS refers to the statistical method of fitting the line of best fit to the data available. Specifically the distance between each data point and the regression line is squared and added together. The OLS model minimises this number, essentially minimising the total distance of data points from the line of best fit.

Random utility models (RUM): RUMs are the basis for discrete choice analysis methods (such as choice experiments and multi-site recreation demand models). An equation is used to estimate the total utility of a choice based on independent variables, which each have their own utility coefficients. RUMs assume that there is some uncertainty in the decision-making process of a respondent and this is accounted for as an error term that deals with:

- Unobserved attributes of the choice options;
- Unobserved characteristics of the respondent;
- Measurement errors; and
- Instrumental variables.

Logistic regression (logit): Logit analysis is a regression based on a logistic model with a dependent variable that is categorical and independent variables that are continuous or categorical. The basic logit analysis is a ***binary logit model***, where the categorical dependent variable has two options (i.e. it is dichotomous), where the option chosen is based on characteristics of the respondent. This means that logit analysis allows the prediction of which of two categories an individual will choose. For example, whether or not they accept a given price or whether they choose one conservation policy over another. Where OLS actually measures changes in the level of dependent variable, logit analysis estimates the likelihood of a category occurring or being chosen.

Multinomial logistic regression: The multinomial logit (MNL) model expands the binary logit model allowing a dependent variable to have more than two categories (i.e. not dichotomous, but polytomous). As with binary logit, multinomial logit assumes that the probability of an outcome is dependent on the characteristics of the respondent making the choice. There are two primary categories of multinomial logit:

- ***Ordinal*** - Categories of the dependent variable are ordered (i.e. can be ranked); and
- ***Nominal*** – Categories of the dependent variable are not ordered.

Conditional logistic regression (clogit): In contrast to binomial or MNL models, clogit models assume that the probability of an outcome is dependent on the characteristics of the outcome options. Clogit models can be combined with multinomial models to account for characteristics of both the respondent and outcome options in estimating the probability of each outcome.

Probit model: Probit models are very similar to logit models. The only difference is the type of distribution the probability of the dependent categories is assumed to follow. In logit models, probabilities are based on a natural logarithm function whereas in probit models it is based on a cumulative normal distribution. In practice there is typically little difference between the outcomes of the two types of models.

Tobit model: Tobit models are similar to probit models but based on the assumption that the distribution is truncated. This means that the distribution of the sample does not go below or above a certain value. Often this is used in analysis of valuation when for example willingness to pay is asked and the distribution of values approximates a 'normal' distribution, but is censored at zero; i.e. negative WTP is not observed.

Nested logit model: The nested logit model is an extension of the MNL model specifically designed to capture correlations among alternative dependent options. Simply, it accounts for unobserved dependencies between alternatives and then treats decisions as a hierarchical choice. For example, for a model analysing choice of recreation site to visit, a primary concern of some visitors might be whether to visit a site that is 'free' or has an 'entry fee'. Nesting the choice on 'free' or 'priced entry' can be useful in better determining the effects of site attributes, such as size, activities, facilities, etc. A nested logit would allow a more clear observation of the utility actually gained from the more relevant site attributes.

Mixed logit model: Mixed logit (MXL) models are an advance on standard logit (or MNL) models. They address three primary limitations of logit and MNL models by allowing for:

- ***Random taste variation*** – Standard logit models assume the effect of independent variables on the dependent variable is fixed over the entire population. MXL models allow these effects to vary between individuals. This is the ***Random Parameter Logit*** **(RPL)** model.
- **Correlation in unobserved factors** – Standard logit models to not take into account unobserved factors that continue to affect an individual's decisions over time. MXL models allow for this, which is particularly important for panel data (i.e. multiple decisions over time from each individual). This is *the* ***Random Paramter Logit with correlation*** **(RPL-correlated)** model.
- **Unrestricted substitution patterns** - Standard logit models assume independence of irrelevant alternatives, meaning that a percentage decrease in the chance of one alternative correlates to a proportional percentage increase in the chance of other alternatives. MXL models relax this

assumption, recognising that substitution patterns may not always be so clear and there may be additional variance induced by non-experienced alternatives (affecting the error component). This is the ***Error Corrected*** model.

Latent class model: These models recognises that individuals may come from particular classes (or groups/types) of individuals. Where other models recognise heterogeneity in parameters between respondents, latent class models go one step further to allow for not only general heterogeneity, but for homogeneity within classes that are heterogeneous from each other.

Fixed effects estimation: Fixed effects estimation controls for unobserved heterogeneity in observations by assuming that all unobserved variables create an individual specific effect on the depdenant variable that is fixed between observations.

Random effects estimation: Random effects estimation assumes that the individual specific effects are uncorrelated with the independent variables, and so change between observations.

Descriptive statistics

R^2 (**R-squared**): The R^2 value is a measure of fit of a linear model (e.g. OLS regression). Essentially it describes how much of the variance data is explained by the model, based on how much variation there was to explain in the first place. It does so by comparing the differences between the data points and the mean (original variation) to the differences between the model function and the mean (model variation). The value ranges from 0 to 1, with values closer to one representing greater explanatory power of the model.

Pseudo-R^2: these values are similar to R^2 values, but for log-likelihood models (e.g. the discrete choice models described above). They are not calculated in the same manner as R^2 values (hence the name 'pseudo'), but still offer an estimate of how much variation a model accounts for. There are a number of different methods for calculating pseudo-R^2, but they are all similar and generally accepted equally. However, pseudo-R^2 values should be interpreted with caution and used simply as a rough gauge of the fit of a model. Importantly, the goodness-of-fit of any models must only be compared using the same Ppeudo-R^2 measure. Types of pseudo-R^2 measures include: McFadden LRI, Estrella, Cragg-Uhler and Veall-Zimmerman goodness-of-fit measures.

Log-likelikood statistic: The log-likelihood statistic is similar to R^2, but instead of measuring the variation based on the mean value, it compares the predicted outcome based on the model to the actual observed outcomes. When comparing between models using the same data, a greater (or less negative) log-likelihood statistic indicates a better fitting model.

AIC: - Akaike Information Criterion (AIC) is a measure of fit for log-likelihood models that penalises larger models (i.e. models that include more variables). The aim is to judge the fit of a model not only based on minimisation of variance, but to also reward a parsimonious (i.e. more simple) model. When comparing models, a lower AIC indicates a preferred model.

F-test: A test of significance similar to a t-test, but rather than testing one coefficient or parameter, an F-test is 'global' evaluating the significance of the entire model. It is used for simple bi-variate and multiple regression. The F-ratio compares the average variability in the data that a given model can

explain to the average variability unexplained by that same model, essentially helping to find the best-fitting model.

Chi-square goodness of fit: A test of whether or not data follow the assumed distribution. This can be used to determine if the model data fit the observed data. If the test statistic is large, then the observed data does not fit that expected based on the model and the model should be rejected.

Interpreting model coefficients

OLS coefficients: Each independent variable of an OLS model has a ß-value associated with it. These values are the model coefficients. If the independent variable increases by one unit, then the dependent variable should increase by the ß-value. However, only coefficients that are significantly different from zero should be accepted (see p-value).

Logistic regression coefficients: Each independent variable of a logistic model has a ß-value associated with it. At first glance, the only meaningful information is whether the sign is positive or negative (i.e. whether the attribute has utility or disutility associated with it) and whether it is statistically significant (see p-value). However, these coefficients are more meaningful compared to each other (i.e. larger coefficients mean that variable has a greater effect on the probability of an outcome). Caution must be used in any such comparison however, since coefficients are only directly comparable if the independent variables they are associated with are measured on the same scale. In the context of economic valuation, a money attribute is often included in utility functions of logistic regressions. Comparing the ß-values of other independent variables with that of the money attribute can provide marginal WTP for those other independent variables. The coefficients of logistic regression can be further analysed to better understand their effect on the probability of an outcome/choice:

- ***Elasticities*** – Can be determined that evaluate the percentage change in probability of an outcome/choice based on a 1 percent change in the independent variable; and
- ***Marginal Effects*** – Are the same as elasticities, but based on a unit change in the independent variable rather than a percentage change.

P-value: This provides a measure of the statistical significance of coefficients by stating the probability of that coefficient estimate being equal to zero; i.e. the variable has no effect on the dependent variable. A p-value ≤0.1 indicates that the coefficient is statistically significant at the 10% level; i.e. there is a 10 percent that the coefficient is equal to zero. Normally a p-value of 0.05 or below is the criteria for significance, with thresholds for categorical levels of significance often set at 0.05, 0.01, and 0.001.

Standard deviation: A measure of variability (or dispersion) of the population from the mean.

Standard error: Standard deviation of the sampling distribution around the estimated mean (note that this can be different from the true mean and standard deviation of the population).

Robust standard error: Robust statistics provide alternative methods of calculating estimators (such as mean, standard error, etc) that are not affected by small departures from model assumptions or outliers in a sample. Robust standard errors are adjusted correlations of error terms across observations.

Confidence interval: These indicate the within which the estimated parameter lies based on the set confidence level. For example, if an interval is given for 95 percent confidence, there is a 95 percent probability that the true value lies within the range of the estimated parameter.

T-test: A statistical test used to determine the p-value of a parameter (including coefficients). For regressions, it is generally used to determine if the parameter is statistically different from zero. Once the t-value is determined, it can be compared to the threshold value on a t-test lookup table to determine what level of significance the parameter has based on the p-values of those threshold t-values (e.g. the chance of being wrong is 10 percent, 5 percent, 1 percent, etc). Alternatively, the p-value can be specifically calculated from the t-value. The t-test is based on the student's t-distribution.

Z-test: A statistical test similar to a t-test, but based on a normal distribution (i.e. is used when the distribution of the parameter under the null hypothesis can be approximated by a normal distribution).