

Introduction to Health Services Research Methods (HS237)  
Overview of Methods

*Imputation*

Imputation methods are used to “fill in” missing data values in a dataset with incomplete cases.

*Sampling theory and survey weights*

Sampling is conducted to select observations from a population. Depending on the goals of the research and other practical considerations, sampling can follow different designs with different implications for deriving sample estimates and errors. Survey weights are often used to weight observations in statistical analysis of datasets based on non-random samples of the population, to facilitate our ability to generalize the results to the underlying population.

*Multinomial logit (MNL)*

MNL is an extension of the logit model used for dichotomous outcomes. MNL is used when the dependent variable is categorical but does not have a natural ordering, e.g., choice of insurance coverage. The dependent variable is specified to be a function of characteristics that vary across decisionmakers. For example, the choice of insurance coverage could be modeled as a function of the decisionmaker’s age, sex, race, income, etc.

*Multinomial probit (MNP)*

MNP is similar to MNL but used when the assumption of the Independence of Irrelevant Alternatives (IIA) fails. It is generally more difficult to run complex categorical choice models using MNP than MNL, so MNL is the first choice if the IIA assumption holds.

*Ordered logit*

Ordered logit is used when the dependent variable is categorical and follows a natural ordering, e.g., self-assessed health = poor, fair, good, very good or excellent.

*Transformations and retransformations*

Transformations (such as log) are often used to improve efficiency of the estimates in cases in which the dependent variable in a linear regression is skewed. In order to put the estimates back on the original scale (e.g., interpret effects of age on dollars instead of log dollars), an appropriate retransformation algorithm must be used. These algorithms can be complex if the regression residuals are heteroskedastic or non-normally distributed.

*Generalized linear models (gamma)*

Gamma models (a special case of generalized linear models, or GLMs) are often used as an alternative to estimating linear regressions with log transformations for continuous dependent variables that are skewed. Gamma models have the advantage that no complicated retransformation algorithm is required, so they are simpler to estimate, especially in the case of heteroskedasticity. However, under certain circumstances they are less efficient.

*Count data models (Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial)*

Count data models are used to model discrete event counts. The Poisson can be thought of as a series of independent Bernoulli trials; the negative binomial relaxes some of the assumptions

of the Poisson. Zero-inflated versions of these models can be used when the number of zero values is greater than would normally be expected.

### *Two-part models*

Two-part models are used for limited-dependent variables, e.g., inpatient costs, where a high fraction of the sample has zero values for the dependent variables. They are especially useful when the conditional sample of users (in this example, those who had any inpatient costs) exhibits a skewed distribution for the level of use. Two-part models offer a very flexible specification because the choice of whether or not the person is a user is estimated separately from the choice of how much the person uses, once they have become a user. Zero-inflated count data models are an alternative to two-part models, with different pros and cons.

### *Taylor series, bootstrapping and simulation methods*

These methods can be used to derive standard errors and/or confidence intervals for certain statistics for which there otherwise would be no formulae.

### *Fixed and random effects models*

Fixed and random effects models are used to account for having multiple observations within a cluster. For example, they can be used with panel data (multiple time observations on a single person) or multi-level data (observations on multiple individuals within facilities, providers, schools, geographic areas, etc.). Both types of models address the issue of biased standard errors due to within-cluster correlation of the residuals. Fixed effects models also addresses a specific omitted-variable bias issue, but at the expense of a sometimes-large efficiency loss (relative to random effects models).

### *Multi-level models*

Multi-level models are a generalization of fixed and random effects models. As such, they are more difficult to estimate, but allow a more flexible specification.

### *Generalized Estimating Equations (GEE)*

GEE is an alternative method for dealing with clustering, i.e., the correlation of the error terms among observations within the same cluster. GEE is often easier to implement than random effects or multi-level models but is a less desirable option under certain circumstances.

### *Survival models*

Survival analysis is used to study the timing of events. It is a family of techniques dealing with the time it takes for something to happen such as death, onset of disease, a reoccurrence (e.g. rehospitalization), or a cure. One feature of survival analyses is that it can deal with censored data – the observation period ends before everyone has experienced the event, therefore leaving the dependent variable unknown for a number of cases.

### *Generalized Tobit models*

Tobit models and generalized Tobit models are used as an alternative to two-part models for estimating limited-dependent outcomes. Generalized Tobit models are conceptually appealing because they allow the two parts of the model to be correlated, but they are extremely sensitive to untestable distributional assumptions.

### *Sample selection models*

Sample selection models are used when the dependent variable is observed only for a non-random sample of the population, e.g., wage rates are observable only for persons who work. They are also used to address non-random attrition bias, e.g., in RCTs. The dependent variable for these models may be either continuous or dichotomous (0/1).

#### *Treatment effects model*

Treatment effects models are used when the dependent variable is continuous and the regressor of interest is dichotomous (0/1) and suffers from potential selection bias, e.g., is correlated with the error term in the regression. A good example is estimation of the effectiveness of a particular medical treatment when the study was observational, so treatment was not randomly assigned.

#### *Bivariate probit model*

Bivariate probit models are similar to treatment effects models, but used when the dependent variable is dichotomous (0/1) rather than continuous.

#### *Two-stage least squares (2SLS)*

2SLS is a method used to address endogeneity bias, i.e., correlation of a regressor with the error term, due to reverse causality, measurement error or omitted-variable bias.

#### *Instrumental variables (IV) with nonlinear variables*

IV with non-linear variables is a generalization of the two-stage least squares method to cases in which the dependent variable and/or endogenous regressor are not continuous.

#### *Propensity scores*

Propensity scores are an alternative to treatment effects or nonlinear IV models when dealing with selection bias. To date the software applications have been limited to the case in which the dependent variable is continuous and the regressor of interest (which suffers from possible omitted-variable bias) is dichotomous. Propensity scores are thought to do a better job of “balancing” treatment groups in terms of observable respondent characteristics, but does not necessarily attenuate (and could actually worsen) bias due to selection on the basis of *unobservable* characteristics.