Advanced Quantitative Statistics

By Scott Oatley

Table of Contents

Confounding, Mediation, and Moderation

Confounding

A key strength of multiple regression analysis is that it allows us to examine potential confounders (exclusionary) or mediators (inclusionary) of the association we are looking at. By always starting with a simple bivariate model (or null model) mediators and confounders can then be added to the model. Then if we compare the slopes of each subsequent model as well as its R-squared we can begin to assess the utility of the models.

By controlling for a confounding variable (like neighbourhood poverty when looking at foreclosure rate and suicide) we can assess an alternative hypothesis that even when controlling for a confounding variable the association between Y and X1 will still persist.

Suppression (Negative Confounding)

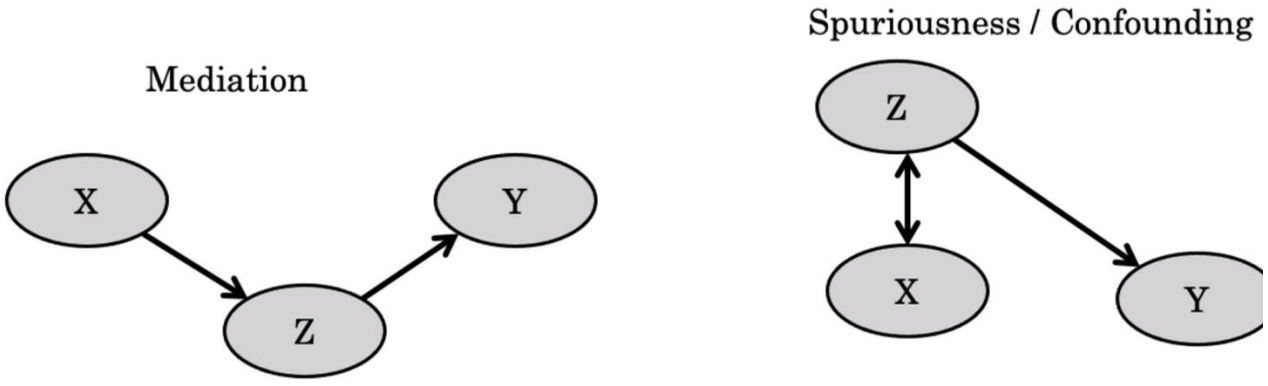
When our primary association turns out to be larger if we controlled for a cofounding factor that is known to be a negative confounding scenario- also known as suppression. In this scenario the confounder actually provides support for your theory.

Mediation

Mediation or a mediating variable explains the relationship of a dependent variable and independent variable.

Mediation vs Confounding

Statistically speaking mediation and confounding look exactly the same- the distinction instead comes from a theoretical role in your focal association and time ordering. Diagrammatically speaking mediation and confounding look very different:



Moderating (Interaction) Effects

Moderating effects are also known as interaction effects. Interaction effects allow us to explore whether the association between a variable of interest and our dependent variable on a third variable.

The non-interactional model is expressed as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Whereas the interaction model is expressed as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_1 (X_1 * X_2)$$

Regression Models for Categorical Dependent Variables: Logistic Regression, Odds Ratios, and Marginal Effects

Logistic Regression

Ordinary Least Squares regression assumes that there are linear assumptions, there is no heteroscedasticity, there are normally distributed errors, and there is an equal distance between response categories. Only continuous dependent variables can meet all of these assumptions.

When wanting to model dependent variables that are not continuous in nature there are a number of possibilities: for binary categorical outcomes logistic regression is required, for ordinal outcomes ordinal logistic regression, and nominal outcomes with >2 response categories multinominal regression is required.

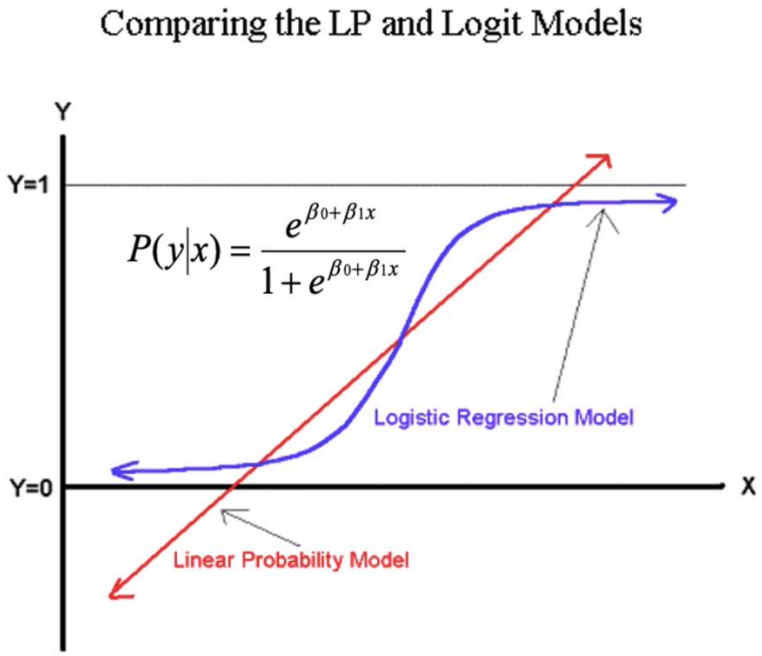
Within binary outcomes, logic dictates that the probability of experiencing a binary outcome is bounded between 0 and 1. Any association therefore is non-linear and limited (bounded). The effect of X on Y depends upon your baseline risk of experiencing Y. This binary outcome violates OLS assumptions: namely normally distributed errors (Y only takes on two values) and homoscedasticity.

By presenting probabilities as odds it gets rid of the bounding problem. Taking the log odds allows us to model the association as non-linear log odds: ln(P/1-P). The logistic regression model is expressed as:

$$Log\ Odds(Y) = \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k$$

The Logistic Curve

Although log odds are bounded to 0-1, they have an infinite range. Log odds follow an S-shaped curve. At the extremes, changes in log odds produce very little change in probabilities. In the middle, changes in the log odds can produce large changes in probabilities. Linear changes in the log odds thus produce non-linear changes in probabilities. Another thing to note is that there are floor and ceiling effects- if things are very good it is very hard for them to get much better and vice versa. The logistic curve can be compared to the Linear Probability (OLS model) below:



Odds Ratio

The dependent variable is expressed in terms of log(odds) of Y occurring. This changes the interpretation of the coefficients: the change in log(odds) of Y occurring is associated with a one unit change in the predictor. Most papers thus transform log(odds) into either odds ratios or marginal effects. The logit model can be expressed as:

$$l = log_b \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

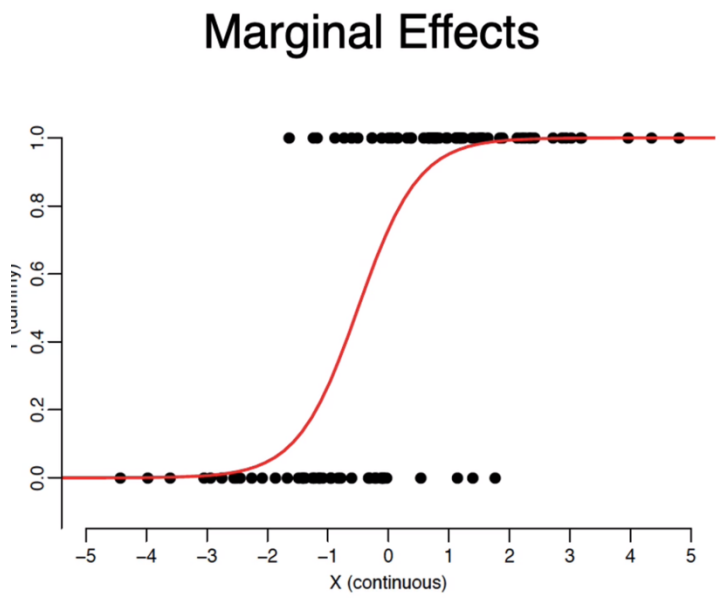
We can recover the odds and use the odds ratio by exponentiating the log(odds):

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 X_1+\beta_2 X_2}$$

We interpret the odds ratio as the change in the odds. Between 0 and 1 of the association with X is negative. 1 if there is no association. Larger than 1 if the association with X is positive.

Marginal Effects

The effect on the conditional mean of Y associated with a change in the independent variable is the marginal effects. Marginal effects are necessary to compare variables within and between models. For non-linear models, the effect of a unit change in an independent variable depends on the value of all independent variables and of all model parameters. An example of marginal effects in a non-linear scenario is seen below:



Marginal Effects in Logit Models

The average effect of going from one value of Xi to another value of Xi across all observations. For example, if the marginal effect of females on smoking is -0.20, then females have a 20% point lower probability of smoking than men on average. In Stata average marginal effects are estimated using the *margins* postestimation command.

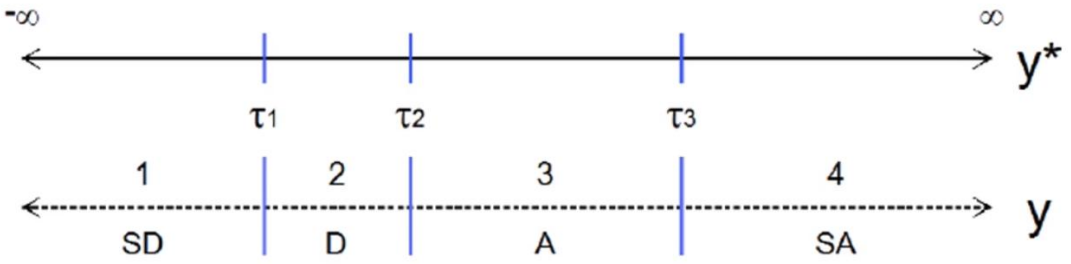
Regression Models for Categorical Dependent Variables: Ordinal and Nominal Outcomes

Ordinal Logistic Regression

Ordinal categories are those variables that are ordered on a single dimension with unkown distance between categories. Observed y is theoretically obtained from y* by dividing it into segments by thresholds τj, expressed as:

$$y_i = j \; if \; \tau_{j-1} \leq y_i^* < \tau_j$$

Representing these thresholds graphically:



And finally, mathematically:

$$y_i = \begin{cases} 1 \Rightarrow \text{SD-Strongly Disagree} & \text{if } \tau_0 = -\infty \leq y_i^* < \tau_1 \\ 2 \Rightarrow \text{D-Disagree} & \text{if } \tau_1 \leq y_i^* < \tau_2 \\ 3 \Rightarrow \text{A-Agree} & \text{if } \tau_2 \leq y_i^* < \tau_3 \\ 4 \Rightarrow \text{SA-Strongly Agree} & \text{if } \tau_3 \leq y_i^* < \tau_4 = \infty \end{cases}$$

Ordered Logit Model

The ordered logit model (OLM) has the slope coefficients of the X regressors the same in each category, only their intercepts differ. Traditionally this is why the OLM is also called the proportional odds model. The model can be expressed as:

$$\log \frac{p_i}{1 - p_i} = a_i + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

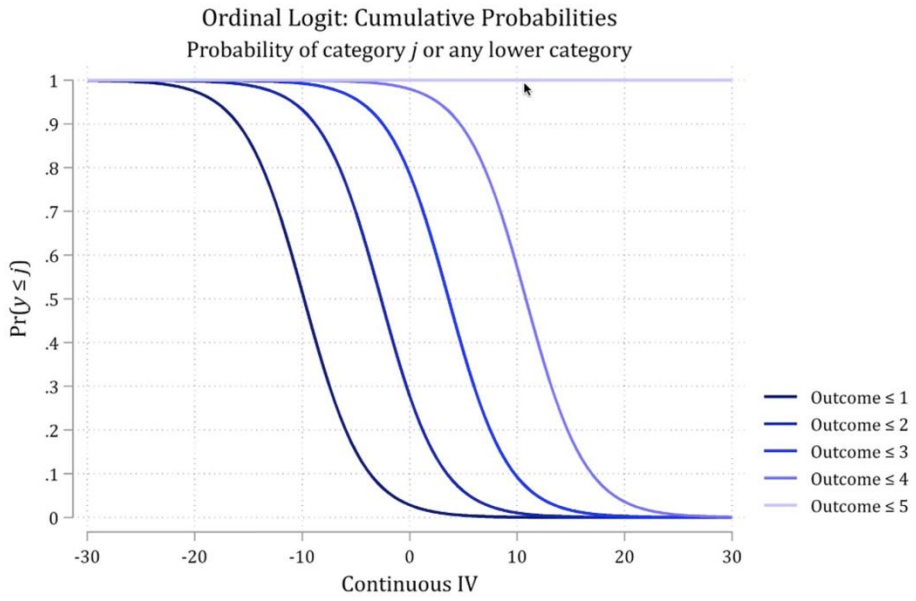
The proportional odds model is used to estimate the cumulative probability of being at or below a particular level of the response variable, or its complimentary, the probability of being beyond a particular level.

The assumptions of the proportional odds model involve: multiple intercept/cut off points, the assumption that beta coefficients do not vary across outcome categories (all independent variables exert the same effect on each cumulative logit regardless of cut off), and finally the Brant test (if we collapsed some of our adjacent categories, would the estimated coefficients be approximately equivalent?).

The cumulative probability model predicts the probability of being in a category j or higher and is expressed as:

$$\Pr(y \leq j) = \frac{\exp\,(x_i(\beta - \tau_j)}{1 + \exp\,(x_i(\beta - \tau_j)}$$

Graphically this is represented as:



Continuation Ratio Model

As already mentioned, the proportional odds model is a form of ordinal logistic regression- but it is not the only form. The key properties of a proportional odds model are that there are cut points where results can be reversed, and the substantive meaning would not change (Palindromic invariance). Alternatively, the continuation ratio model is another

form of ordinal logistic regression. With this model, parametrization is slightly different to the proportional odds model. Results and substantive meaning are changed when the cut points are reversed- meaning no palindromic invariance. If there is a natural baseline that all individuals start from then it is advisable to go for this model over the proportional odds model.

The continuation ratio model estimates the odds of being in a particular category *j* relative to being that category or beyond. In this situation, the continuation ratio model can be formulated as:

$$\ln\left(\frac{\pi(Y = j\ x_1, x_2, \ldots x_p)}{\pi(Y \geq j\ x_1, x_2, \ldots x_p)}\right) = a_j + (-\beta_1 X_1 - \beta_2 X_2 - \cdots \beta_P X_P)$$

Where $Y = j\ x_1, x_2, \ldots x_p$ is the conditional probability of being in category *j,* conditional on being that category or beyond, given a set of predictors. j=1,2,…J-1. $a_j$are the cut points and $\beta_1, \beta_2, \ldots \beta_p$are the logit coefficients. The stata command *ocratio* fits the continuation ratio model with the command *eform* used to estimate the odds ratios and corresponding standard errors and confidence intervals.

Multinominal Logistic Regression

The multinominal logistic regression requires a reference category to compare the probability of each outcome category to the probability of the reference category. Probabilities of all outcomes amount to 1. The independence of irrelevant alternatives assumption must be satisfied- if A is preferred to B out of a set choice {A, B}, introducing a third option X, expanding the choice set to {A, B, X}, must not make B preferable to A.

Multinominal logit is a set of binary logits that are simultaneously estimated: L vs M using nl+nm observations. It is difficult to judge statistical significance using tests from a minimal set (which is what *mlogit* defaults to). The stata command *listcoef* will list all coefficients.

The hypothesis that xk has no effect involves a joint test of all J-1 coefficients for a given IV. We can use either a Wald or likelihood-ratio test. The stata command *mlogtest* will calculate joint tests for all IVs.

Plots of Predicted Probabilities

Plots are especially helpful with multiple outcome categories. Use *margins* to make predictions, *marginsplot* to graph predictions, and *noci* will omit the many confidence

intervals making it easier to read. The interpretation of marginal effects in multinominal logit is all but identical to the interpretation in binary logit.

Fixed and Random Effects Models

Fixed Effects Models

In nonexperimental research, the classical way to control for potentially confounding variables is to measure them and put them in some kind of regression model. No measurement, no control. The basic idea of fixed effects models is very simple: use each individual as their own control. If we average those differences across all persons in the population, we get an estimate of the ''average treatment effect''.

There are two basic data requirements for using fixed effects methods. First, the dependent variable must be measured for each individual on at least two occasions. Those measurements must be directly comparable, that is, they must give the same meaning and metric. Second, the predictor variables of interest must change in value across those multiple occasions for some substantial portion of the same.

The term fixed effects is usually contrasted with random effects model. In a classic view, a fixed effects model treats unobserved differences between individuals as a set of fixed parameters that can either be directly estimated or patialed out of the estimating equations. In a random effects model, unobserved differences are treated as random variables with a specified probability distribution.

In a more modern framework, (Wooldridge, 2002) argues the unobserved differences are always regarded as random variables. Then, what distinguishes the two approaches is the structure of the associations between the observed variables and the unobserved variables. In a random effects model, the unobserved variables are assumed to be uncorrelated with all the observed variables. In a fixed effects model, the unobserved variables are allowed to have any associations whatever with the observed variables.

There are some serious disadvantages of fixed effects models. A classical fixed effects approach will not produce any estimates of the effects of variables that don't change over time. Second, In many cases, fixed effects estimates may have a substantially larger standard errors than random effects estimates, leading to higher p values and wider confidence intervals. This is because fixed effects models use only within-individual

differences, essentially discarding any information about differences between individuals. If predictor variables vary greatly across individuals but have little variation over time for each individual, then fixed effects estimates will be very imprecise.

We sacrifice efficiency to reduce bias.

Fixed effects models are one way in which we study longitudinal data. Fixed effects are regression models for panel data that ''differences out'' between person differences, sweep away all stable between-unit (person) differences that are stable over time (wi). Fixed effects restrict analysis to within-unit change. It models within-unit change in the dependent variable as a function of time-varying independent variables. Fixed effects models are formulated as:

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + a_i + \varepsilon_{it}$$

The μ_t is an intercept term that can be difference for each time-period. μ_t does not vary across cases, it only varies across time. x stands for the independent variables whose values can vary across time (income, marital status etc). z stands for the independent variables whose values do not change across time (race, gender etc). We can say that these variables have time-invariant values or measure stable characteristics. β and γ are the coefficients for the xs and zs. The model assumes that these effects are time-invariant (the effect of x1 is the same at time 1 as it is at time 4). Interactions with time can be added is the effects of the xs or zs are thought to vary with time (race may be thought to have less effect at time 1 than time 4). a_i and ε_{it} are both error terms. The latter is different for each individual at each point in time. the former only varies across individuals but not across time- we can think of it as representing the effects of all the time invariant variables that have not been included in the model.

The two error terms a_i and ε_{it} behave somewhat different from each other. There is a different ε_{it} for each individual at each point in time, a_i only varies across individuals, not over time. We regard a_i as representing the combined effect on y of all unobserved variables that are constant over time. On the other hand, ε_{it} represents purely random variation at each point in time.

At this point I'll make some rather strong assumptions about ε_{it}, namely, that each ε_{it} has a mean of zero, has a constant variance (for all I and t), and is statistically independent of everything else (except for y). The assumption of zero mean is not critical as it is only relevant for estimating the intercept. The constant variance assumption can sometimes be

relaxed to allow for different variances for different t. Note that the ε_{it} at any one period is independent of x_{it} at any other period, which means that x_{it} is strictly exogenous.

As for a_i, the traditional approach in fixed effects analysis is to assume that this term represents a set of n fixed parameters that can either be directly estimated or removed in some way from the estimating equations.

Although we'll assume statistical independence of a_i from ε_{it}, we allow for any correlations between a_i and x_{it}, the vector of time-varying predictors. If we are not interested in γ, we can also allow for any correlations between a_i and z_i.

The inclusion of such correlations distinguishes the fixed effects approach from a random effects approach and allows us to say that the fixed effects method ''controls'' for time-invariant unobservables.

The Two-Period Case

Estimation is particularly easy when the variables are observed at only two periods (T=2). The two equations are then:

$$y_{i1} = \mu_1 + \beta x_{i1} + \gamma z_i + a_i + \varepsilon_{i1}$$

$$y_{i2} = \mu_2 + \beta x_{i2} + \gamma z_i + a_i + \varepsilon_{i2}$$

By subtracting the first equation from the second, we get the 'first difference' equation:

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

Which can be rewritten as:

$$\Delta y_i = \Delta\mu + \beta\Delta x_i + \Delta\varepsilon_i$$

Where Δ indicates a difference score. Note that both $a_i \; and \; \gamma z_i$ have been ''differenced out'' of the equation. Hence, we no longer have to be concerned about a_i and its possible correlation with Δx_i. On the other hand, we also lose the possibility of estimating γ. Since $x_{i2} \; and \; x_{i1}$ are each independent of $\varepsilon_{i2} \; and \; \varepsilon_{i1}$, it follows that Δx_i is independent of $\Delta\varepsilon_i$. This implies that one can get unbiased estimates of β by doing ordinary least squares (OLS_ regression on the difference scores.

Whenever conventional regression produces a significant coefficient but fixed effects regression does not, there are two possible explanations: (a) the fixed effects coefficient is

substantially smaller in magnitude and/or (b) the fixed effects standard error is substantially larger. Standard errors for fixed effects models are often larger than those for other methods, especially when the predictor variable has little variation over time. Whenever p values differ from other methods always check the standard errors and coefficients.

Extending the Difference Score for the Two-Period Case

The basic effects model can be extended to allow for the effects of x and z to vary over time. In the two-period case, we can write the equations with distinct coefficients at each period:

$$y_{i1} = \mu_1 + \beta_1 x_{i1} + \gamma_1 z_i + a_i + \varepsilon_{i1}$$

$$y_{i2} = \mu_2 + \beta_2 x_{i2} + \gamma_i z_i + a_i + \varepsilon_{i2}$$

Taking first differences and rearranging terms produces:

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta_2(x_{i2} - x_{i1}) + (\beta_2 - \beta_1)x_{i1} + (\gamma_2 - \gamma_1)z_i + (\varepsilon_{i2} - \varepsilon_{i1})$$

Which could also be written as:

$$\Delta y_i = \Delta\mu + \beta_2 \Delta x_i + \Delta\beta x_i + \Delta\gamma z_i + \Delta\varepsilon_i$$

There are three things about this equation. First, as before, a_i has dropped out, so we don't have to be concerned about its potential confounding effects. Second, z has not dropped out, and its coefficient vector is the difference in the coefficient vectors for the two time points. From this we learn that time-invariant variables whose coefficients vary over time must be explicitly included in the regression equation. Fixed effects only controls for time-invariant variables with time-invariant effects. Third, the equation now includes x_1 as a predictor, and its coefficient vectyor is the difference in the coefficient vectors for the two time periods. Thus, for z and x_1, tests for whether their coefficients are 0 are equivalent to testing whether $\beta_1 = \beta_2$ or $\gamma_1 = \gamma_2$.

A First-Difference Method for Three of More Periods per Individual

When each individual is observed at three or more points in time (T>2), it's not so obvious how to extend the methods we just considered. One possible approach is to construct and estimate two first-difference equations.

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

$$y_{i3} - y_{i2} = (\mu_3 - \mu_2) + \beta(x_{i3} - x_{i2}) + (\varepsilon_{i3} - \varepsilon_{i2})$$

These equations can be estimated separated by OLS, and each will give unbiased estimates of β.Under the assumption that β does not vary over time, the two equations should be estimated simultaneously for optimal efficiency. This can be accomplished by creating a single data set with two records for each person, one with the difference scores for the first equation and the other with the difference scores for the second equation. There should also be a dummy variable distinguishing the first record from the second record. Also there should be a variable with a common ID number for the two records from each person.

The intercept can be interpreted as an estimate of $\mu_2 - \mu_1$ while the coefficient for the equation dummy is an estimate of $(\mu_3 - \mu_2) - (\mu_2 - \mu_1)$. Although the combined OLS estimates should be unbiased, they ignore the fact that $\varepsilon_2 - \varepsilon_1$is likely to be negatively correlated with $\varepsilon_3 - \varepsilon_2$ because they share a common component, ε_2, with opposite signs. This implies that the coefficient estimates may not be fully efficient and the standard error estimates may be biased. We can solve this problem by estimating the correlation between the error terms and then using generalised least squares (GLS) to take account of that correlation.

Such GLS programs typically require the specification of an ID variable so that observations from the same individual can be identified. You can use the *xtreg* command in Stata with the *pa* option, which estimates the linear model using GLS.

The first-difference method can be easily extended to more than three periods per individual. For T periods per individual, T-1 records are created, each with difference scores between adjacent time points for all variables. Additionally, there should be a variable containing a common ID number for all observations from the same individual and a variable or a set of dummy variables to distinguish the different records. The regression is then estimated on the entire set of records, using GLS to adjust for correlations among the error terms. Unless T is large, for example, greater than 10, it's probably best to allow the error correlation matrix to be unstructured. That is, the matrix would allow for a different correlation between each pair of error terms. With larger T, it may be preferable to impose a simplified structure to reduce the number of distinct correlations that need to be estimated.

Dummy Variable Method for Two or More Periods per Individual

Although the multiple-difference-score method is a reasonable way to estimate a fixed effects model for the multi-period case, the name ''fixed effects'' is usually reserved for a different method, one that can be implemented either by dummy variables or by constructing mean deviations. The results while not identical are usually very similar to the difference-score method. In the two-period case, the two methods give identical results.

The dummy variable method requires a data set with a rather different structure: one record for each period for each individual. The time-varying variables have the same variable names on each record but different values. For any time-invariant variables, their values are simply replicated across the multiple records for each individual. There should also be an ID variable with a common value for all the records for each individual. Last, there should be a variable distinguishing the different periods for each individual.

To implement the method, it's necessary to construct a set of dummy variables to distinguish the individuals in the data set. Many statistical packages can do this automatically by specifying the UD variable as a categorical variable. If the TIME variable is also specified as a categorical variable, two dummy variables will be created to distinguish the here different years. One can then use OLS to estimate the coefficients. The coefficients for the dummy variables created from the ID variable are actually the estimates of the a_i, under the contrast that one of them is equal to 0.

The best situation for a fixed effects analysis is when all the variation on a time-varying predictor is within person, but there's still a lot of between-person variation on the response variable.

The problem with the dummy variable method is that the computational requirement of estimating coefficients for all dummy variables can be quite burdensome, especially in large samples where it may be beyond the capacity of the software or the machine memory. Fortunately, there is an alternative algorithm- the mean deviation method- that produces exactly the same results. The one drawback is that it doesn't five estimates for the coefficients of the dummy variables representing different persons, but those are rarely of interest anyway.

The mean deviation algorithm works like this. For each persons and for each time-varying variable (both response and predictor variables), we compute the means over time for that person:

$$\overline{y_i} = \frac{1}{n_i}\sum_t y_{it}$$

$$\overline{x_i} = \frac{1}{n_i}\sum_t x_{it}$$

Where n_i is the number of measurements for person i. Then we subtract the person-specific means from the observed values of each variable.

$$y_{it}^* = y_{it} - \overline{y_i}$$

$$x_{it}^* = x_{it} - \overline{x_i}$$

Finally, we regress y* on x*, plus variables to represent the effect of time. This is sometimes called a 'conditional method' because it conditions out the coefficients for the fixed effects dummy variables.

If you construct the deviation scores yourself and then use OLS, you will get the correct OLS estimates for all the coefficients but not the standard errors and p values. That's because the calculation of the degrees of freedom is based on the number of variables in the specified model, when it should actually include the number of dummy variables implicitly used to represents different persons in the sample. The *xtreg* command in Stata does the correct calculations for a fixed effects model.

The *xtreg* command also reports several additional statistics that are specific to a fixed effects model:

An F test of the null hypothesis that all the coefficients for the fixed effects dummy variables are zero. This is equivalent to saying that there is evidence for person-level unobserved heterogeneity. That is, there are stable differences in antisocial behaviour between persons that are not fully accounted for by the measured predictor variables.

An estimate of the proportion of variance in the dependent variable that is attributable to the fixed effects (the $a_i s$), labelled ''rho (fraction of variance due to u_i).''

An estimate of the correlation between the fixed effects a_i and $\hat{\beta}x_{it}$, the estimated linear combination of the time-varying predictors. In a random effects model, this correlation is assumed to be 0.

Three R^2s: within, between, and overall. The within R^2 is just the usual R^2 calculated for the regression using the mean deviation variables. The between R^2 is the squared correlation between the person-specific mean of y and the predicted person-specific mean of y. Finally, the overall $R^2$is the squared correlation between y itself and the predicted value of y. All three of these R^2s are calculated using predicted values based on the estimated regression coefficients but not using the coefficients for the fixed effects dummy variables.

Interactions with Time in the Fixed Effects Method

For each of the interactions with time, the t statistic tests whether a coefficient at Time 2 or Time 3 is different from the coefficient at Time 1.

Comparison with Random Effects Models

A popular alternative to the linear fixed effects model is the random effects or mixed model. This model is based on the same equation that we used for the fixed effects model:

$$y_{it} = \mu_i + \beta x_{it} + \gamma z_i + a_i + \varepsilon_{it}$$

The crucial difference is that now, instead of treating a_i as a set of fixed numbers, we assume that a_i is a set of random variables with a specified probability distribution. For example, it is typical to assume that each a_i is normally distributed with a mean of 0, constant variance, and is independent of all the other variables on the right hand side of the equation.

Contrary to popular belief estimating a random effects model does not really ''control'' for unobserved heterogeneity. That's because the conventional random effects model assumes no correlation between the unobserved variables and the observed variables. The fixed effects model, on the other hand, allows for any correlations between time-invariant predictors and the time-varying predictors. It does so, however, at the cost of some efficiency in the event that those correlations are really zero.

The simpler model (random effects) will lead to more efficient estimates, but those estimates may be biased if the restrictions of the model are wrong. The more complex model (fixed effects) is less prone to bias but at the expense of greater sampling variability.

The Hausman test compares the random effects and fixed effects models- this helps determine whether the biases inherent in the random effects method are small enough to ignore, or whether we need to move to the less restrictive fixed effects mode. The Hausman

test of the null hypothesis that the random effects coefficients are identical to the fixed effects coefficients.

The fixed effects model has the benefits of: being better able to deal with unobserved heterogeneity, allows us to study change, is a relatively straightforward extension of the OLS model, and allows for multiple levels of fixed effects to be modelled. Some of the costs of fixed effects are it limiting our focus to time varying independent variables, as well as limiting our analysis to units that experience change on time-varying variables- this limits statistical power and raises external validity issues, unobserved heterogeneity is still an issue (confounders that are time-varying remain a problem).

If your goal is to study change, then fixed effects models are for you. Also, if you care a great deal about unobserved heterogeneity then fixed effects models are your best bet. Since fixed effects models discard a lot of information standard errors tend to be large. These large standard errors can sometimes be tolerated through a tradeoff- other models like random effects models will suffer from omitted variable bias which fixed effects models can control for.

A Hybrid Method

In the hybrid method, we combine elements of a fixed effects and random effects model. The time-varying x variables are again transformed into deviations from their person-specific means, but the response variable y is not. Furthermore, unlike previous fixed methods fixed effects methods, we now include a time-invariant z variables in the regression model. In addition, we also include variables that are the person-specific means for each of the time-varying variables. Finally, instead of doing OLS regression, we estimate a random effects model to ensure that the standard errors reflect the dependence among the multiple observations for each person.

In the multilevel model literature (Bryk and Raundenbush 1992; Goldstein 1987; Kreft and DeLeeuw 1995) the practice of subtracting person-specific means from each time varying variable is called group mean centering. Although group mean centering can produce substantially different results, the literature has not made the connection to fixed effects models, nor has it been recognized that group mean centering controls for all time-invariant predictors.

Another attraction of the hybrid approach is that it can be extended in interesting ways that are not easily handled with the conventional methods for fixed effects estimation. The random effects models that we have been considering so far are all random intercept models though it is also possible to estimate random slope models.

Using the hybrid model, it's also possible to estimate models with more complex error structures than the rather simple structure implied by the conventional fixed effects model (See Singer and Willett 2003).

Fixed Effects logistic Models

The notation for logistic regression fixed effects models is as follows:

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \mu_t + \beta x_{it} + \gamma z_i + a_i, \text{ t=1,1,…,T}$$

Where p_{it} is the probability that the response variable is equal to 1. As before, x_{it} is a vector of time-varying predictors, z_i is a vector of time-invariant predictors, and a_i represents the combined effects of all unobserved variables that are constant over time. We shall treat a_i as a set of fixed constants, one for each individual. This is equivalent to assuming that a_i is random with unrestricted associations between a_i and x_{it}.

The Two-Period Case

An analgous procedure is available for logistic regression (comparable to that of fixed effects linear models for the two-period case). We apply a conventional maximum likelihood to estimate the model:

$$\log\left(\frac{p_i}{1-p_i}\right) = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1})$$

This is actually a form of conditional maximum likelihood estimation. As in the linear case, both z_i and a_i drop out of the equation.

Three or More Periods

Both conditional and unconditional maximum likelihood (using dummy variables and multiple records using a deviation from its person-specific mean respectively) are available for logistic regression of dichotomous outcomes, but in this case, they do not produce the same results. As in the linear case, unconditional maximum likelihood is implemented by creating multiple records per person and estimating a conventional logistic regression

model with dummy variables for persons. Unfortunately, this method produces biased estimates of the coefficients (Hsiao 1986). In fact, in the two-period case, the coefficient estimates are exactly twice as large as they should be (Abrevaya 1997). The reason for this bias is something called the incidental parameters problem (Kalbfleisch and Sprott 1970; Lancaster 2000). What happens is the number of parameters increases directly with the sample size, thus violating one of the conditions that underlie the asymptotic theory of maximum likelihood estimation.

The solution is to do conditional maximum likelihood, which conditions the a_i parameters out of the likelihood function (Chamberlain 1980). This is accomplished by conditioning the likelihood function on the total number of events observed for each person. In effect, each person's contribution to the likelihood function is the answer to a question: 'Given that a girl was in poverty 2 out of the 5 years, what is the probability that this happened in, say, Years 2 and 4 rather than in one of the nine other possible pairs of years?'. This conditioning approach only works for the logistic regression model for dichotomous response variables, not for other ''link'' functions such as probit or complementary log-log.

Within Stata, the routines to maximize the conditional likelihood for logistic regression can be accomplished using either the *xtlogit* or *clogit* command.

The Stata command *xtlogit* fits logistic regression models to panel data using three quite different methods: fixed effects (conditional likelihood), random effects, and generalized estimating equations.

Both GEE and random effects estimates do nothing to control for unmeasured predictors. In contrast, fixed effects estimation (conditional likelihood) controls for all constant predictors. It also produces appropriate estimates of standard errors that are corrected for dependence. On the downside, (looking at data from page 36 of Allison 2009) those standard errors are larger than the random effects or GEE standard errors because a substantial amount of information in the data is not used. In applications where the withing-person variation is small relative to the between-person variation, the standard errors of the fixed effects coefficients may be too large to tolerate.

Another point worth noting is that both conditional likelihood and random effects estimates are ''subject specific'' while the GEE estimates are ''population averaged''. A subject-specific coefficient tells us what would happen to a single individual if that person's predictor variable were increased by one unit. In contrast, a population-averaged

coefficient tells us what would happen to the whole population if everyone's predictor variable were increased by one unit. If the model is linear, there is no distinction between the two kinds of coefficients. For logistic regression models, however, subject-specific coefficients are typically larger than population-averaged coefficients.

Interactions With Time

Another downside of the conditional likelihood method is that coefficients can't be estimated for variables that don't vary over time (though these variables are implicitly controlled). Interactions between time-varying and time-constant variables can be included in the model.

The interpretation of these interactions is somewhat different for the time-varying and time-constant predictors. For the time-varying predictors, it's usually best to consider how the effect of each predictor arise with time. For time-constant predictors, the best way to interpret the interactions is to examine how the effect of *time* varies with these variables.

A Hybrid Method

As in the linear case, the attraction is that we can (a) include time-constant variables in the mode, (b) perform test for comparing fixed effects and random effects, and (c) fit a wider class of models. Unlike the conditional likelihood, the hybrid approach can be used with other link functions such as probit or complimentary log-log.

The coefficients for the mean variables are not very interesting in themselves. A conventional random effects model implicitly assumes that the deviation coefficients are identical to the mean coefficients. We can easily test that assumption within the hybrid model by directly testing for equality across the pairs of coefficients.

Another advantage of the hybrid approach is the ability to get estimates for the time-constant predictors.

In Stata, logit models with random coefficients require a different command, *xtmelogit*.

Methods for More Than Two Categories on the Response Variable

Consider a categorical response variable y_{it} that can take on more than two values. Suppose that those values are the integers ranging from 1 to J, with the running index j. Let $p_{itj} = Prob(y_{it} = j)$. We then need a model of how this probability depends on our predictor variables x_{it} and z_i.

In the case in which the categories of the dependent variable are ordered a fixed effects version of the model can be written as:

$$\log\left(\frac{F_{ijt}}{1-F_{ijt}}\right) = \mu_{tj} + \beta x_{it} + \gamma z_i + a_i, \text{ j=1,…,J-1}$$

Where $F_{ijt} = \sum_{m=j}^{J} p_{imt}$ is the cumulative probability of being in category j or higher. This equation can be viewed as as et of simultaneous binary logistic regression equations, each equation comparing one category with the last category.

The more complicated situation in which the categories of the dependent variable are not ordered calls for the use of a multinomial logit model, also known as the generalized logit model. The firxed effecs version of that model is:

$$\log\left(\frac{p_{ij}}{p_{iJ}}\right) = \mu_{tj} + \beta_j x_{it} + \gamma_j z_i + a_{ij,} \text{ j=1,…,J-1}$$

The fixed effects $a_{ij,}$ vary both over individuals and over possible response values, but not over time.

The fixed effects multinomial logit model is like the binary logit model in that it has reduced sufficient statistics for the a_{ij}'s, namely, the frequency counts of the different response values for each individual. In principle, the model can be estimated by conditional maximum likelihood with conditioning on those counts (Chamberlain 1980). However, there is no software available to do this. If the time-varying predictors are categorical, the model can be reformulated as a log-linear model and estimated in that framework (Conaway 1989).

Another approach to estimation is to decompose the multinomial model into a set of binary models, one model for each comparison of a particular category with a reference category (Allison 1999a). Each binary model can then be estimated using the conditional logistic regression methods. This approach produces approximately unbiased estimates of the coefficients though results will differ depending on the choice of the reference category.

Estimating the multinomial logit model in Stata, you can use the *mlogit* command with robust standard errors to correct for dependence in the repeated observations for each person.

Fixed Effects Models For Count Data

Many researchers treat count variables as continuous variables and do their analysis with ordinary least squares regression. Count variables are necessarily discrete and cannot have values less than zero. Usually, their distributions are highly skewed.

Poisson Models for Count Data With Two Periods per Individual

A fixed effects Poisson regression model can be estimated with an ordinary logistic regression program for grouped data. To illustrate this (example used in pp. 50 of Allison 2009) Let y_{i1} be be the patent count for firm i, in 1975 and y_{i2} the patent count in 1979. Each of these variables is assumed to have a Poisson distribution with an expected value of λ_{it}. That is, the probability that $y_{it} = r$ is given by:

$$\Pr(y_{it} = r)\frac{\lambda_{it}^{r} e^{-\lambda_{it}}}{r!}, \text{ r=0,1,2,…}$$

The Poisson distribution is perhaps the simplest probability distribution that is appropriate for count data. It may be derived form a stochastic process model under the assumptions that (a) events cannot occur simultaneously and (b) events are independent (Cameron and Trivedi 1998). The independence assumption means that the occurrence of an event neither raises nor lowers the probability of future events.

An unusual property of the Poisson distribution is that its mean and variance are equal:

$$E(y_{it}) = var(y_{it}) = \lambda_{it}$$

This does of course lead to issues of overdispersion that can seriously compromise the estimation of Poisson regression models.

Next, we let λ_{it} be a log-linear function of the predictor variables:

$$log\lambda_{it} = \mu_t + \beta x_{it} + \gamma z_i + a_i$$

The objective is to estimate the parameters in the above equation. TO do this, we can use conditional maximum likelihood, the same method used to estimate the fixed effects logistic regression model. Consider the distribution of y_{i2} conditional on the total event count for the two time periods combined, denoted by $w_i = y_{i1} + y_{i2}$. It can be shown that $y_{i2}|w_i \sim B(p_i, w_i)$. That is, conditional on the total count w_i, the 1970 count y_{i2} has a binomial distribution with parameters $p_i \; and \; w_i$ where:

$$p_i = \frac{\lambda_{i2}}{\lambda_{i2} + \lambda_{i1}}$$

It follows, that:

$$\log\left(\frac{p_i}{1 - p_i}\right) = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1})$$

Thus, we have been able to convert our Poisson regression model into a logistic regression model in which the predictor variables are difference scores for the original predictors.

To implement this conditional method using Stata, I can use the *blogit* command, which does ML estimation of grouped binomial data. The *blogit* command expects the dependent variable to come in two parts: the number of ''events'' and the number of ''trials''.

Poisson Models for Data With More Than Two Periods per Individual

There are two approaches to estimation of a fixed effects Poisson model when individuals are observed at more than two periods. The first is a conditional ML and the second an unconditional ML. In conditional ML the likelihood function is conditioned on the sum of all counts for each individual, which eliminates the fixed effects (a_i). The resulting conditional likelihood (Cameron and Trivedi 1998) is proportional to:

$$\prod_i \prod_t \left(\frac{\exp(\mu_t + \beta x_{it})}{\sum_s exp(\mu_s + \beta x_{is})}\right)^{y_{it}}$$

In Stata, this likelihood can be maximized with the *xtpoisson* command. This command requires that the data be restructured so that there is one record for each year, with a common ID variable linking together the records for each year.

The default in Stata is to assume that a_i has a log-gamma distribution, but it's possible to specific a normal distribution. The population-averaged model, on the other hand, does not postulate an additional disturbance term in the Poisson regression equation, but merely allows the multiple observations for each to be correlational. This model is estimated by the GEE method, which, as in the logistic case, is a kind of iterated generalized least squares. Both random effects and GEE are vulnerable to overdispersion, so the conventional standard errors are biased.

The fixed effects Poisson regression model can also be estimated by unconditional ML. This is accomplished by estimating a conventional Poisson regression model that includes

dummy variables. For the Poisson regression model unlike the logistic regression model, conditional and unconditional estimation always produce identical results. Consequently, the choice between one or the other is purely a matter of computational convenience.

Fixed Effects Negative Binomial Models for Count Data

Fixed effects Poisson regression models are quite vulnerable to the effects of overdispersion. That's somewhat unexpected because fixed effects already allow for unobserved heterogeneity across individuals by way of the a_i parameters. That heterogeneity is presumed to be time invariant, however, and there may still be unobserved heterogeneity that is specific to particular points in time, leading to observed overdispersion. The standard errors can be corrected for overdispersion by using the bootstrap or jackknife methods. A better method is by directly building overdispersion into the model for event counts.

To model the overdispersion we now assume that counts are drawn from a negative binomial distribution for each count at each point in time The negative binomial distribution is a generalization of the Poisson distribution that allows for overdispersion by way of an additional parameter. The appeal of the negative binomial model is that the estimated regression coefficients may be more efficient, and the standard errors and test statistics may be more accurate than those produced by such empirical, after-the-fact corrections.

The traditional NB2 model is given by:

$$\Pr(y_{it} = r) = \frac{\Gamma(\theta + r)}{\Gamma(\theta)\Gamma(r+1)}\left(\frac{\lambda_{it}}{\lambda_{it} + \theta}\right)^r \left(\frac{\theta}{\lambda_{it} + \theta}\right)^\theta$$

In this equation λ_{it} is the expected value of y_{it}, θ is the overdispersion parameter, and $\Gamma(.)$ is the gamma function. As $\theta \to \infty$, this distribution converges to the Poisson distribution. As with the Poisson midel, we assume that the expected value of y_{it} is described by the log-linear regression:

$$log\lambda_{it} = \mu_t + \beta x_{it} + \gamma z_i + a_i$$

Where a_i are treated as fixed effects. Unlike the Poisson model conditional likelihood is not an option. In technical terminology, the total count for everyone is not a ''complete sufficient statistic'' for a_i, so conditional on the total does not remove a_i from the likelihood function. Whilst Hausman, Hall, and Griliches (1984) proposed a rather different fixed

effects negative binomial regression model derived from a conditional ML estimator and their method has been incorporated in to the Stata command *xtnbreg* Allison and Waterman (2002) have shown this is not a true fixed effects regression model and the method does not control for all stable predictors.

An unconditional ML by estimating negative binomial regression models that include dummy variables for all individuals (except one) is the preferred option. In Stata, this can be done with the *nbreg* command. Computation is quite slow because of the many coefficients that must be estimated.

Using Monte Carlo simulations, Allison and Waterman (2002) found that the unconditional negative binomial estimator did not show any substantial bias from incidental parameters. They also showed that negative binomial estimators had substantially smaller true standard errors that Poisson estimators. Unconditional negative binomial estimation did have one flaw however: confidence intervals tended to be too small. Under many conditions, the nominal 95% confidence intervals covered the true value only about 85% of the time. This problem can be easily corrected by adjusting the standard errors for overdispersion using a formula based on the deviance statistic. Although Stata does not report the deviance statistic required for this correction, the standard errors produced by the *vce(opg)* option in Stata are about the same as those produced by the deviance correction.

A Hybrid Approach

A hybrid approach is also available to Poisson regression methods. Here, it is best to use negative binomial regression models because they are less prone to overdispersion. To get correct standard errors, it's important to use an estimation method that allows for dependence among the multiple observations for everyone. Either a random effects model or a population-averaged (GEE) model can accomplish that.

These can be estimated using Stata's *xtnbreg* command.

As usual, one of the attractions fo the hybrid method is the ability to include time-invariant predictors. The other attraction is the ability to test the fixed effects model against the more restricted random effects model

Fixed Effects Models For Events History Data

Event history analysis is the name given to a set of statistical methods that are designed to describe, explain, or predict the occurrence of events. These methods are also called survival analysis. These methods are appropriate for a vast array of social phenomena such as births, marriages, divorces, job terminations, promotions, arrests, migrations, and revolutions.

In general, an event may be defined as a qualitative change that occurs at some point in time. To apply event history methods, you need event history data, which is simply a longitudinal record of when events occurred to some individual or sample of individuals. If you want to do a causal or predictive analysis, you will also want to measure possible explanatory variables.

Cox Regression

The most popular method for analyzing event history data is Cox regression. Rather than directly modeling the length of the interval, the dependent variable in Cox regression is the hazard of instantaneous likelihood of event occurrence. For repeated events, the hazard may be defined as follows. Let $N_i(t)$ be the number of events that have occurred to individual i, by time t. The hazard for individual I, at time t is given by:

$$h_i(t) = \lim_{\Delta t \to 0} \frac{\Pr\left[N_i(t + \Delta t) - N_i(t) = 1\right]}{\Delta t}$$

In words, this equation says that we should consider the probability of one additional event in some small interval of time Δt. Then form the ratio of this probability to Δt, and take the limit of this ratio as Δt goes to 0. For repeated events, the hazard function is also known as the intensity function.

Next, we model the hazard as a function of the predictor variables. Letting $h_{ik}(t)$ be the hazard for the kth event for individual i, a proportional hazards model is given by:

$$\log h_{ik}(t) = \mu\left(t - t_{i(k-1)}\right) + \beta x_{ik}$$

Where x_{ik} is a column vector of predictor variables that may vary across individuals and across events, β is a row vector of coefficients, $t_{i(k-1)}$ is the time of the (k-1)th event, and $\mu(.)$ is an unspecified function of the length of time since the most recent event. In this model, we assume that $\mu(.)$ is the same function for all individuals in the sample.

A remarkable thing about partial likelihood is that it can estimate β without making any assumptions about the function μ. In Stata, Cox regression is implemented with the *stcox* command.

Cox Regression With Fixed Effects

The first version of our fixed effects regression model is:

$$logh_{ik}(t) = \mu\left(t - t_{i(k-1)}\right) + \beta x_{ik} + a_i$$

The possibility to use dummy variables raises the issue of bias- (Allison 2002) has shown that the Cox regression is more like the logistic regression in this respect. When the average number of intervals per person is less than three, using dummy variables to implement fixed effects produces regression coefficients that are biase (away from zero) by approximately 30% to 90%, depending on the level of censoring (a higher proportion of censored cases produces greater inflation).

There is an alternative method that is easily implemented and very effective. By modifying the previous equation to:

$$\mu_i\left(t - t_{i(k-1)}\right) = \mu\left(t - t_{i(k-1)}\right) + a_i$$

Which then yields:

$$logh_{ik}(t) = \mu_i\left(t - t_{i(k-1)}\right) + \beta x_{ik}$$

In this equation, the fixed effect a_i has been absorbed into the unspecified function of time, which is now allowed to vary from one individual to another. The only difference between this equation and the conventional Cox model is the i, subscript on μ. Thus, each individual has their own hazard function, which is considerably less restrictive than allowing each individual to have their own constant.

Stratification allows different subgroups to have different baseline hazard functions, while constraining the coefficients to be the same across subgroups. It is accomplished by constructing a partial likelihood function for each subgroup, multiplying those likelihood functions together, and then maximizing the resulting likelihood function with respect to the coefficient vector β. With the *stcox* command in Stata, stratification is implemented by specifying the option *strata(caseid)*, which means that each of the cases is treated as a separate stratum.

Stata can also estimate a random effects Cox model, with the assumption that a_i has a gamma distribution and is independent of x_i. Models of this sort are often called ''shared frailty'' models with a_i described as the frailty term. The idea is that some individuals are more frail than others and, hence, are more likely to experience the event. The *stcox* option for estimating such models is *shared(caseid).*

Some Caveats

Despite the attractions of fixed effects Cox regression, it also has unusual disadvantages. There may be a substantial loss of power compared with the conventional analysis. Even for those observations that are retained, the fixed effects method essentially discards information about variation across and only uses variation within. So, if a particular covariate varies a great deal across but shows little variation over time for each individual, the coefficient for that variable will be unreliably estimated.

Fixed effects Cox regression is also susceptible to bias for certain kinds of variables. These problems are most likely to occur with data structures whereby individuals are observed for a fixed period of time and may have multiple events during that period, but only the last interval is censored. Chamberlain (1985) argued that this structure violates a basic condition of likelihood-based estimation because the probability that an interval is censored depends on the length of the previous intervals.

In a simulation study (Allison 1996) showed that this violation does not create a serious problem for most predictor variables but could lead to biases in estimating the effects of variables that describe the previous event history. In particular, fixed effects partial likelihood tends to find negative effects on the hazard for the number of previous events and the length of the previous interval, even when those variables do not have true effects. This problem tends to be the most severe when the average number of events per individual is low, and the proportion of intervals that are censored is high.

The Hybrid Method for Cox Regression

The Hybrid method for Cox regression does not seem to work following Allison's (2007) simulation.

Fixed Effects Event History Methods for Non repeated Events

Fixed effects Cox regression requires that at least some of the individual's in a sample experience more than one event so that within-individual comparisons are possible. The

method cannot be applied to a nonrepeatable event such as death. Under certain conditions, it may be possible to do a fixed effets analysis for nonrepeatable events by treating time as discrete and applying conditional logistic regression. This type of analysis is called a case-crossover study (Maclure 1991).

We regard time as consisting of discrete units, which we can enumerate t=1,2,3,… (using example from Allison 2007 pp.80) Let p_{it} be the probability that husband I, dies on day t, given that he was still alive on the preceding day, and let $W_{it} = 1$ if the wife I, was alive on date t, otherwise 0. We will represent the effect of wife's vital status on the probability of husband's death by a logistic regression model:

$$log\left(\frac{p_{it}}{1-p_{it}}\right) = a_i + \gamma_t + \beta W_{it}$$

Where γ_t represents a linear effect of time on the log-odds of death and a_i represents the fixed effects of all unmeasured variables that are constant over time. No time-invariant predictors are included in the model because their effects are absorbed into the a_i term.

In cases where the algorithm used to maximize the likelihood function does not converge- the log-likelihood quickly goes to 0 and the iteration sequence continues without end. The reason for this failure could be that the sequence of observations consists of a string of 0s on the dependent variable followed by a 1. In other words, the event always occurs at the last observation unit. As a consequence, time or any monotonically increase function of time will perfectly predict the outcome making it impossible to get maximum likelihood estimates for that covariate or ant other covariate in the model. In logistic regression literature this problem is known as complete separation (Allison 2004).

In the above example even removing time, the issue persists as W_{it}, the dummy variable for wife's death may increase with time but never decrease, it perfectly predicts the occurrence of a death on the last dat. Consequently, its coefficient gets larger at each iteration of the algorithm.

One way to circumvent this problem is to redefine W_{it} to be an indicator of whether the wife died within the previous 60 days. This covariate changes from 0 to 1 when the wife dies, but then goes back to 0 after 60 days.

This would however provide no control for time. This can seriously compromise any conclusions drawn from a case-crossover study (Greenland 1996).

Another alternative is to use what Suissa (1995) called the ''case-time-control'' design, the key innovation in this approach is the computational device of reversing the dependent and independent variables in the estimation of the conditional logit model. This makes it possible to introduce a control for time, something that cannot be done with the case-crossover method.

When both the dependent and the independent variables are dichotomous, the odds-ratio is symmetric. In the case-time-control method, the working dependent variable is the dichotomous covariate. Independent variables are the dummy variable for the occurrence of an event on a given day and some appropriate representation of time, for example, a linear function. Again, a conditional logistic regression is estimated with each couple treated as a separate stratum. Under this formulation there is no problem including time as a covariate because the working dependent variable is not a monotonic function of time.

In Suissa's formulation it is essential to include data from all individuals, both those who experienced the event and those who are censored. Censored individuals provide information about the dependence of the covariate on time, information that is not confounded with the occurrence of the event.

The case-time-control method has been critiques for assuming that the dependence of the covariate on time is the same among those who did and did not experience the event (Greenland 1996). This criticism has no force is the data are limited to individuals who experience the events.

The working model is defined as follows (Using example from Allison 2007 pp.83), let H_{it} be a dummy variable for the death of the husband I, on date t, and let P_{it} be the probability that the wife's death occurred within a specific number of days prior to day t. The logistic regression model is as follows:

$$log\left(\frac{p_{it}}{1-p_{it}}\right) = a_i + \beta_1 H_{it} + \beta_2 t + \beta_3 t^2$$

This model allows for a quadratic dependence on time, although other functions could be used instead.

Structural Equation Models With Fixed Effects

By putting a model into a structural equation framework, we can accomplish several things that are difficult or impossible with conventional computational methods. In particular we can:

- Estimate models that are a compromise between fixed and random effects
- Construct a likelihood ratio test for fixed versus random effects
- Estimate fixed effects models with reciprocal effects between the two response variables
- Estimate fixed effects models with lagged values of the response variable
- Estimate models with multiple indicators of latent variables

Random Effects Models

The random effects model:

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + a_i + \varepsilon_{it}$$

Where y_{it} is the value of the response variable for individual i, at time t, x_{it} is a vector of time-varying predictors, z_i is a vector of time-invariant predictors, a_i denotes the random effects, and ε_{it} is a random disturbance term. We assume that a_i and ε_{it} represent independent normally distributed variables with a mean of 0 and each having a constant variance. We also assume, at that these random components are independent of both x_{it} and z_i.

Random effects are models for panel data that utilize those models between and within person variance in Y as a function of Xs and a random error term (a_i). Unlike fixed effects, which assume all between-unit differences are time stable, random effects assumes all unmeasured between unit differences are random (uncorrelated with the error term).

The benefits of a random effects mode are: it allows us to estimate the effect of time stable characteristics on Y, it is more efficient than fixed effects as it utilizes between and within person variation, and is also preferable to pooled OLS because it explicitly models non-independence in the data. However, some of the drawbacks are: unobserved heterogeneity is an issue (the same requirements with OLS remain), and there must be variation in Y over time across unites, if not there's no reason to study this change.

If your goal is to study time-constant independent variables, then random effects models are for you. Also, if you care more about model efficiency and statistical power than unobserved heterogeneity random effects models are a good choice.

Multilevel Modelling and Hierarchical Linear Models (HLM)

Multilevel Modelling

One form of complex data is 'multilevel' or 'clustered' data. It arises when individual records in the data can be located in one or more subgroups that involve one or more 'hierarchies'. Cluster coding information must exist in the data. Multiple hierarchies need not overlap. Most commonly, analysis proceeds at the lower level, but it would seem wrong to ignore the 'clustering' or 'nesting' into higher level units. We might be really interested in the higher level units, deliberately designing a study to assess them, or we might just want to control for them. 'Multilevel modelling' can be thought of as any adjustment to a model that is designed to take appropriate account of clustering.

Cluster coding information must exist in the data. Multiple hierarchies need not overlap. Most commonly, analysis proceeds at the lower level, but it would seem wrong to ignore the 'clustering' or 'nesting' into higher level units. We might be really interested in the higher level units, deliberately designing a study to assess them, or we might just want to control for them. 'Multilevel modelling' can be thought of as any adjustment to a model that is designed to take appropriate account of clustering. Typical examples (N=cases; k = clusters; kb= cases per cluster): Respondents in households (e.g. N = 5000, k=3000, kb=1.7); Resps. in PSUs or interviewer groups (e.g. N = 10000, k = 200, kb=50); Students in classes or schools (e.g. N = 2000, k = 100, kb=20); Subjects in companies / institutions (e.g. N = 500, k = 50, kb=10); Respondents in countries in cross-national studies (e.g. N=16000, k=16, kb=1000). Random effects models generally productive when both N and k are large (often design studies to maximise k, not kb). Random effects not inappropriate, but may be suboptimal, when k or kb is small.

Multilevel structures observe that certain data has hierarchical or clustered structures. Multilevel models recognize the existence of such data hierarchies by allowing for residual components at each level in the hierarchy. For example, a two-level model which allows for grouping of child outcomes within schools would include residuals at the child and school

level- this the residual variance is partitioned into a between-school component and a within-school component.

To aggregate all data to a higher level would result in an ecological fallacy. To pretend that level 2 variables are level 1 variables would result in magic multiplication.

Traditional multiple regression techniques treat the unites of analysis as independent observations. One consequence of failing to recognize hierarchical structures is that standard errors of regression coefficients of higher-level predictor variables will be the most affected by ignoring grouping.

In many situations a key research question concerns the extent of grouping in individual outcomes, and the identification of 'outlying' groups. In evaluations of school performance, for example, interest centers on obtaining 'value-added' school effects on pupil attainment. Such effects correspond to school-level residuals in a multilevel model which adjusts for prior attainment.

An alternative way to allow for group effects is to include dummy variables for groups in a traditional OLS regression model. Such a model is called an analysis of variance or fixed effects model. In a multilevel model, the effects of both types of variables can be estimated.

In a multilevel model the groups in the sample are treated as a random sample from a population of groups. Using a fixed effects model, inferences cannot be made beyond the groups in the sample.

Multilevel modeling becomes relevant if we know of some connections between some of the cases in the data. This might have been deliberately intended as a focus of analysis (e.g. educational research samples including multiple children from the same school). Equally, many social surveys feature clustering of cases which are not central to analysis but should be controlled for. Multilevel models try to find ways to take account of the clustering within their data, in terms of the statistical routines for the individual level analysis.

'Random intercepts' models allow the error term to feature a set difference for all the cases within each higher level unit (i.e. uj adds or subtracts systematically for each higher level unit j; the effect is basically to shift the intercept term up or down by a certain amount for each cluster, hence 'random intercepts'). The clusterspecific adjustments to the intercepts are 'random effects' as they are modelled as a distribution of random values, to be characterised by a dispersion parameter (variance or standard deviation).

Conventional reasons to move from a single level to a multilevel random effects model: We should get more appropriate standard errors for beta coefficients (especially higher-level variables), we should get some additional parameters that tell us useful things about the overall level of influence of the cluster factor, common position: if the total influence is minimal, evident in a low 'ICC' (see later) -> no need to fit the multilevel model and sensible to stick with the single level model, in some circumstances, the beta coefficients might change as well, Influenced by the balance of 'between' and 'within' cluster effects – e.g. Jones 2008, models that ignore random effects are 'unbiassed but inefficient' (i.e. right betas but wrong standard errors); this only holds, though, if 'within' and 'between' patterns are aligned and/or appropriately decomposed, in some circumstances, it's useful to make inferences about individual higher level units on the basis of their model-based residuals from a random effects model.

To assess whether the random effects error decomposition improves the model we should compare the model deviance between the multilevel model and the equivalent single level model.

Single level regression models (and other introductory statistical techniques) typically ignore clustering in data, so the great attraction of multilevel models is the opportunity to incorporate this important feature into the analysis. The main contribution of random effects multilevel models is to model the presumed 'similarity' shared by different members of the same cluster. In general, only 2 or 3 levels of clustering are easy to analyse through multilevel models: it is often easy to conceive of more complex clustering frameworks, but it's not so easy to model them. Clustering need not be hierarchical: non-hierarchical clusters (such as children in families in schools, where children from the same families may attend different schools) are known as 'cross-classified' clusters; they can be modelled, but the procedures are harder. Random effects multilevel models are not always the best way to recognise clustering: (i) the existence of clusters does not necessarily mean they are empirically important to the process being studied; (ii) some groups that might potentially be thought of as a type of clustering might not have properties that suit a random effects analysis (e.g. ethnic group).

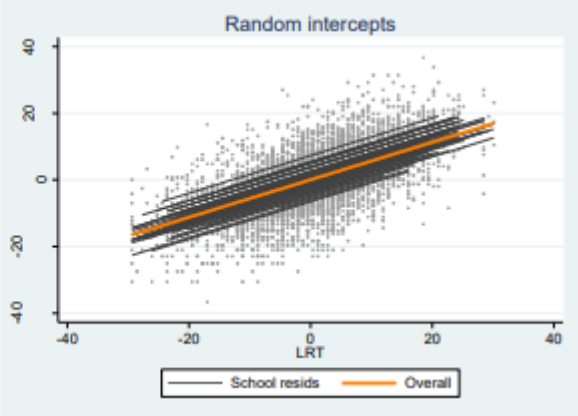
Random intercepts or 'variance components' models can usefully summarise the relative proportion of error variance associated with the different levels Intra-cluster correlation ('rho', ρ) $\rho = \sigma u\ 2\ /\ (\ \sigma u\ 2 + \sigma \varepsilon\ 2\ )$. Total variance $= \sigma u\ 2 + \sigma \varepsilon\ 2$ = (sd_cons)2+ (sdResidual)2 = (1.471)2+ (5.147)2 = 2.16 + 26.49 = 28.65 Intra-cluster correlation $= \rho = \sigma u\ 2\ /\ (\ \sigma u\ 2 + \sigma \varepsilon\ 2\ )$

= 0.076 (ICC=proportion of error variance associated with level 2; a VPC or 'Variance Partition Coefficient' is the proportion of error variance associated with any given level).

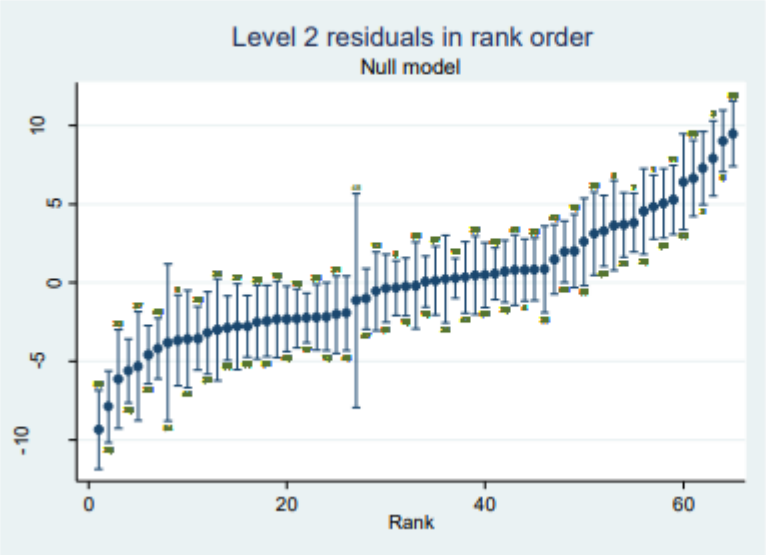
Intra-cluster correlation indicates the breakdown of variance between higher and lower level units. We often calculate the ICC for the null model (tells us about variance pattern in Y). ICC for models with explanatory variables needs careful expression (residual variance in Y). Whether or not a model ICC is non-zero is often used to decide if random effects is needed at all. Often, the ICC will decline when we add fixed part terms related to the higher level (since relatively less remains unexplained about the higher level).

For most people, the main outputs from multilevel models are parameter estimates for the fixed part (beta coefficients) and random parts (error variance partition estimates). Level 1 residuals are the individual error terms and are usually only used to check model assumptions. Level 2 residuals are error patterns at the group level and are substantively informative.

(The group level lines are calculated from the group level residuals; they are not ordinarily generated immediately within the model outputs)



- The **'caterpillar plot'** works out the higher level residuals and their errors, and ranks them by means



Can identify extreme cases (and consider modelling them explicitly, e.g. with a single dummy variable)

This graph is the GCSE attainment by schools dataset, and is generated in lab p3_stata.do

Hierarchical Linear Model

The hierarchical linear model is closely related to OLS regression- it has a fixed and a random part. It also contains interesting information in the residuals. For the case of 2 levels and 2 variables, X and Z the equation looks similar to OLS:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

However, the parameters Beta are different to OLS:

$$\text{Intercept:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + U_{0j}$$
$$\text{Slope:} \quad \beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + U_{1j}$$

The expected value in case X is zero and consists of three parts. Y00 is the constant term; average value of intercept Zj=0. Y01Zj is the influence of a level two variable Zj on the intercept, weighted by Y01. Finally, UOj is the random effect of the jth level two unit on the intercept. This final part is what makes the hierarchical linear model special. It is the random variation in the intercept- has a mean of zero and variance of t2, which indicates how much of the variance we cannot explain.

All together the hierarchical linear model is formulated as:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_j + U_{0j} + (\gamma_{10} + \gamma_{11}Z_j + U_{1j})X_{ij} + r_{ij}$$
$$= \underbrace{\gamma_{00} + \gamma_{01}Z_j + \gamma_{10}X_{ij} + \gamma_{11}Z_jX_{ij}}_{\text{Fixed part}} + \underbrace{U_{0j} + U_{1j}X_{ij} + r_{ij}}_{\text{Random part}}$$

Quasi-Experiments: Instrumental Variables

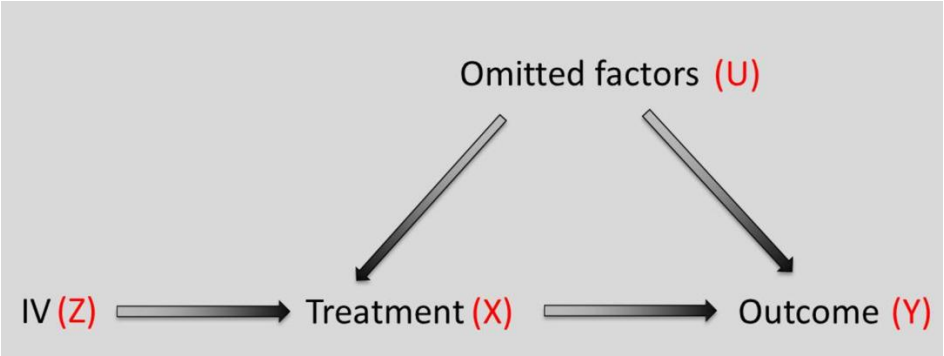
Randomized Control Trials and Quasi-Experimental Methods

With a lot of models there are potential issues that arise from unmeasured confounders as well as reverse causality (whereby we do not know which way the causal relationship occurs). Randomized Control Trials (RCTs) is used to establish causal inference. RCTs work by randomly assigning people to an experiment and a control group and calculating the proportion of people getting X between the two groups. The difference is the average causal effect (ACE) of Y. Whilst RCTs break the link between omitted factors and the treatment there are certain issues. RCTs are often times infeasible (we cannot always manipulate the treatment), costly, and unethical.

Quasi-experimental methods offer an alternative to using RCTs. These quasi-experimental methods are: matching, instrumental variables, difference-in-differences, regression discontinuity design and fixed effects models. Within quasi-experimental studies, the treatment is not randomly assigned by the researcher, so we need to look for ways to approximate a random assignment.

Instrumental Variables

An instrumental variable is a variable that affects the treatment and also does not directly affect the outcome, as depicted below:



Instrumental variables are operationalized via Two Stage Least Squares (2SLS)- using the stata command *ivreg2*. The first stage regresses the treatment (X) on the instrumental variable (Z) and obtains the predicted values of the treatment ($\hat{X}$). The second stage regresses the outcome (Y) on the predicted values of the treatment ($\hat{X}$) to obtain the causal effect of the treatment (X) on the outcome (Y). We then finally obtain the local average

treatment effect- this is different from the average treatment effect as we can divide the population into four groups.

Local Average Treatment Effect

The four groups of the local average treatment effect (LATE) are: the always takers- those that take the treatment regardless of whether they were assigned the treatment or not. The never takers- those that do not take the treatment regardless of whether they were assigned the treatment or not. Compliers- those that take the treatment if they were assigned the treatment and do not take the treatment if they were not assigned the treatment (most important to look at for analysis). Defiers- those that take the treatment if they were not assigned the treatment and do not take the treatment if they were assigned the treatment.

The IV method identifies LATE: the causal effect on the treatment on the outcome for the compliers only.

There are several assumptions for LATE:

Relevance refers to an instrument being strong enough to predict the treatment after controlling for covariates. If it is not strong enough then we will suffer from weak instrument bias. This assumption is empirically testable through an F-statistic in the first stage being greater than 10.

The exclusion restriction states that the instrument should have no direct effect on the outcome. This is not empirically testable and thus requires substantive knowledge.

Independence refers to the idea that there should be unobserved factors affecting both the instrument and the outcome. Once again this is not empirically testable and requires substantive knowledge.

Monotonicity states that there should be no defiers. As with the second and third assumptions, this is not empirically testable and requires substantive knowledge.

Finally, the stable unit treatment value assumption (SUTVA) states that the potential outcome of one individual is not affected by the treatment assignment or instrument assignment of the other individual (no interference). Also, no different forms or versions of each treatment level which lead to different potential outcomes (no hidden variations). This is also not empirically testable and needs substantive knowledge.

Instrumental variables offer the ability to estimate causal effects and eliminates the classical measurement error in the treatment, but this comes at a cost. It is difficult to find instrumental variables that satisfy the assumptions, it only reflects the complier effects- which in turn are vulnerable to small sample bias. There is also an internal/external validity trade-off.

Quasi-Experiments: Regression Discontinuity Design

Sharp and Fuzzy Regression Discontinuity Designs

In quasi-experimental studies, the treatment is not randomly assigned so there is a need to look for ways to approximate random assignment. Regression Discontinuity Designs (RDDs) rely upon a discontinuity or threshold (called a running variable) that determines the probability of treatment assignment.

Sharp RDDs

Within Sharp RDDs all individuals receive treatment on one side of the threshold, and they do not receive treatment on the other side of the threshold. In practice this means that we regress outcome (Y) on the treatment (X) and a polynomial of a running variable (C). The gap between the regression lines constitutes the causal effect at the threshold. The slopes between the two regression lines can be different- they can also be non-linear. It is advisable to only use polynomials between one and two.

An example of a sharp RDD would be the effect of graduating with honours on earnings. Graduating with honours is endogenous (students more motivated etc), we use GPA as a threshold. The reason this is a sharp RDD is because there is perfect compliance, passing the threshold means you graduate with honours and failing means you do not graduate with honours.

Fuzzy RDDs

Fuzzy RDDs state that the probability to receive treatment is larger on one side of the threshold than on the other side of the threshold (Imperfect compliance). Fuzzy RDDs are operationalized via two stage least squares. First, we must construct an instrument Z given a value of one if the running variable passed the threshold and a value of zero if the running variable did not pass the threshold (in a sharp RDD this instrument was equal to the treatment). In the first stage, regress the treatment (X) on the instrumental variable (Z)

and a polynomial of the running variable (C) and obtain the predicted values of the treatment (X). In the second stage, regress the outcome (Y) on the predicted values of the obtained treatment (X) and a polynomial of the running variable (C) to obtain the effect of the treatment (X) on the outcome (Y). Unlike traditional instrumental variables, the fuzzy RDD sees the polynomial of the running variables at both stages. LATE is at the threshold, the causal effect of the treatment on the outcome for the compliers at the threshold only.

Continuity Assumption

To identify a local causal effect at the threshold RDD needs to satisfy the continuity assumptions. The continuity assumption is; all observed and unobserved factors besides the treatment and outcome should be continuous at the threshold- there should be no jump at the threshold. A sufficient condition for the continuity assumption to hold is that individuals cannot perfectly manipulate the threshold. To assess the continuity assumption we can plot observed factors against the threshold and ask if there are any other discontinuities, if there is a placebo discontinuity, as well as include and exclude covariates, and producing a McCray density test (smooth distribution= no manipulation and is the best way to test the continuity assumption).

Strengths and Weaknesses of RDD

Whilst the RDD does produce an unbiased estimate of the causal local treatment effect, and it boasts high internal validity, as well as it being well suited for graphical representation and many checks for the continuity assumption, the RDD does have drawbacks. For one there are certain functional form issues- which polynomial to choose? Nonlinearities are also often mistaken for discontinuities. There is also an internal/external validity tradeoff- causal effect only at the threshold.

Nonparametric RDD

For a nonparametric RDD we must choose a bandwidth across the cutoff point beforehand. The question then becomes how large should the bandwidth actually be? There are luckily some standardized methods to choose a specific bandwidth size.

Stata offers commands for parametric sharp RDD (reg) and fuzzy RDD (ivreg2) as well as graphical representation for non-parametric sharp RDD (rdplot) and fuzzy RDD (rdrobust). Finally stata offers a command for the McCray density test (dcdensity).

Modelling Count Data

Introduction

A count response model is a statistical model for which the dependent variable is a count. A count Is understood as a non-negative discrete integer ranging from zero to some specified greater number. There are typically four types of count data:

- A count or enumeration of events
- A count of items or events occurring in a given spatial area
- A count of items or events occurring within a period of time
- A count of the number of people having a particular disease, adjusted by the size of the population at risk of contracting the disease

Basic Linear Model

For normal linear regression, the errors are Gaussian or normally distributed. μ us used to refer to the predicted value, without a hat. When estimating a parameter, a hat should go over it. The true unknown parameter has no hat.

Models and Probability

All parametric statistical models are based on an underlying probability distribution. The probability distribution function (PDF) can never be truly known. The PDF is assumed to describe the population data, not only the sample from it that we are actually modelling. This way of looking at statistics and data is referred to as frequency-based statistical modelling. Bayesian models look at the relationship of data to probability distributions in a different manner.

Count Models

The majority of count models are based on two probability distributions- the Poisson and negative binomial PDFs. On top of these there are also- the Poisson inverse Gaussian model, or PIG, Greeve's three parameter negative binomial P, or NB-P, and generalized Poisson, or GP models.

The Poisson distribution has a single parameter to be estimated, μ, or the mean, which is also sometimes referred to as the location parameter. The unique feature of the Poisson

distribution is that the mean, and variance are the same. The higher the value of the mean of the distribution, the greater the variance or variability in the data.

The criterion of the Poisson distribution is referred to as the equidispersion criterion. The problem is when modelling real data, the equidispersion criterion is rarely satisfied. Overdispersion is by far the foremost problem facing analysts who use Poisson regression when modelling count data. Overdispersion almost always refers to excess variability or correlation in a Poisson model, but also needs to be considered when modelling other count models as well.

Poisson overdispersion occurs in data where the variability of the data is greater than the mean. A model that fails to adjust for overdispersed data has biased standard errors and cannot be trusted. The most popular method of dealing with apparent Poisson overdispersion is to model the data using a negative binomial model- it has an extra parameter, referred to as the negative binomial dispersion parameter- also known as the heterogeneity or ancillary parameter.

The negative binomial is derived as a Poisson-gamma mixture model, with the dispersion parameter being distributed as gamma shaped. The gamma PDF is pliable and allows for a variety of shapes, as a consequence, most overdispersed count data can be appropriately modelled using a negative binomial regression. The advantage to using a negative binomial regression rests with the fact that when the dispersion parameter is zero, the model is Poisson values of the dispersion parameter greater than zero indicate that the model has adjusted for correspondingly greater amounts of overdispersion. The negative binomial parameter will be symbolised as a(alpha). The negative binomial model cannot be used to model underdispersed Poisson data.

The Poisson inverse Gaussian model is also known as the PIG model. PIG assumes that overdispersion in a Poisson model is best described according to the inverse Gaussian distribution rather than the gamma distribution. The PIG model is used via the 'pigreg' command in Stata. The PIG dispersion parameter is also known as a(alpha) so that a value of a=0 is Poisson.

The three-parameter generalized negative binomial designed by William Greene is also called NB-P. the dispersion parameter, a, of negative binomial and PIG models has the same value for all observations in the model. The third parameter of NB-P, called ρ, or rho

in Greek, allows the dispersion to vary across observations, providing a better opportunity to fit negative binomial data.

The generalized Poisson is also known as the GP model. Similar to the first two models the GP has a second parameter. Also, like the first two models the GP reduces to Poisson when the dispersion is zero. A feature of GP models is that the dispersion parameter can have negative values, which indicate an adjustment for Poisson underdispersion.

Selected count model mean-variance relationship:

Model	Mean	Variance
Poisson	μ	μ
Negative Binomial	μ	$\mu(1+a) = \mu + a\mu$
Negative Binomial 2	μ	$\mu(1+a) = \mu + a\mu^2$
Poisson inverse Gaussian	μ	$\mu(1+a\mu^2) = \mu + a\mu^3$
Negative Binomial-P	μ	$\mu(1+a\mu^p) = \mu + a\mu^p$
Generalized Poisson	μ	$\mu(1+a\mu)^2 = \mu + 2a\mu^3 + a^2\mu^3$

Structure of a Count Model

Nearly all count models follow the basic structure:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n$$

To isolate the predicted mean count on the left side of the equation, both sides are exponentiated, giving:

$$\mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n}$$

There is not a linear relationship between μ and the predictors as there is in a linear model. The linear relationship is between the natural log of μ and the predictors. The linear predictor is the same as the predicted or expected value. Typically, we symbolize the summation of the terms of the linear predictor for each observation in a model as:

$$(x\beta)_i = \sum_{i=1}^{n} \beta_0 + \beta_1 X_{1i} + \cdots \beta_j X_{ji}$$

With, i, indicating the observation number in the model data and j the number of predictors in the model. Notice that I used the standard mathematical ∑ (sigma) sybol for summation in the above equation. The summation starts at the quantity indicated below sigma and ends with the value at its top. Here we have observation number I, starting at 1 representing the first observation in the data, and finishing with n, indirectly the last observation In the data being modeled.

The relationship of the predicted or fitted statistic, μ, and the linear predicted, xb, is the same for Poisson, negative binomial, and PIG regressions. The term log(μ) is called the link function since it links the linear predictor and predicted value:

$$\log(\mu i) = \sum_{i=1}^{n} \beta_0 + \beta_1 X_{1i} + \cdots \beta_j X_{ji}$$

An important feature of having the natural log link for count models is that it guarantees that the predicted values will always be positive (i.e μ>0). Using a linear regression when modelling counts cannot make such a guarantee.

Varieties of Count Models

If zero counts are not a possibility for the data being modelled, then the underlying PDF may need to be amended to adjust for the excluded zero counts. Zero-truncated (ZT) models are constructed for this purpose.

Having data with excess numbers of zeros is also a problem. The expected percentage of zero counts on the basis of the Poisson PDF is well under 1%. Typically analysts use either a two-part hurdle model or a mixture model, such as a zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB).

Hurdle models are nearly always constructed as a two-part 0,1 response logistic or probit regression and a zero-truncated count model. The logistic component models the

probability of obtaining a non-zero count. After separating the data into two components a binary variable is created where all counts greater than zero are assigned the value of one.

Zeros in the count model are zeros in the logit component. The count component truncates or drops observations with zero values for the original counts and, for example, models the data as a zero-truncated Poisson. The model described here is a Poisson-logit hurdle model.

Zero-inflated models are mixture models. They use logistic or probit regression for the binary component, but both components- the binary and count- include the same zero counts when being estimated. The overlap of zero counts means that the mixture of Bernoulli (distribution used in binary log regression) and Poisson distributions must be adjusted so that the resulting PDF sums to one. Zero-inflated models structure the binary component, so it models zeros not ones.

The generalized NBP negative binomial model is one of many three parameter count models that can be used on count data that fail to fit any of the standard count probability distributions, including mixtures of distributions. The NBP model parametrizes the exponent on the second term of the negative binomial variance. The negative binomial variance function is $\mu + a\mu^2$, we may symbolize the parameter as ρ(rho), representing the power $-\mu + a\mu^2$. μ,a,ρ are all parameters to be estimated.

The NB1 model has a variance function of $\mu + a\mu^1$. Since it has a linear negative binomial; the traditional negative binomial is sometimes referred to as the quadratic negative binomial because of the square exponent. The foremost use of the NBP model is to have it determine whether the data prefer NB1 or NB2. If ρ is close to 2, the analyst should use NB2 over NB1. If alpha is 0.5 and ρ is 1.8, then that should be the reported model.

It is possible that data came from population data that cannot have values below 3, or perhaps above 10 etc. if values are truncated at the low end of the counts the model is said to be left truncated; if they cannot exist higher than some cut point, the model is right truncated. Interval truncation exists when counts only exist between specific count values.

Non-parametric models like the generalized additive models (GAMs) are used to assess the linearity of continuous predictors with respect to the response and provide information concerning what type of transform is needed to effect linearity.

Quantile count models are also non-parametric but are used to describe the empirical distribution underlying one's data. Quantile count models are used when a parametric distribution cannot be identified.

Bayesian modelling is appropriate when you wish to have constraints on a predictor or to provide information about a predictor or predictors in a model in addition to the information already available given the predictor. It I also useful when there does not appear to be a PDF underlying the data to be modelled. Using a Markov chain Monte Carolo (MCMC) sampling algorithm, a well-fitted empirical distribution can usually be found for which the user can obtain a mean and standard deviation and 95% quantiles. These translate to a predictor coefficient, standard error, and what is termed a credible interval.

Estimation- the modelling process

All but a very few of the count models discussed are estimated using either: IRLS or maximum likelihood estimation (MLE). IRSL is an acronym meaning ''iterative reweighted least squares', which is the traditional method used to estimate models from the generalized linear model (GLM) family. IRLS is based on a simplification of MLEs that can occur when the models to be estimated are members of the one-parameter exponential family of probability distributions- this includes Poisson and negative binomial regressions.

Mixed effects models use neither IRLS or MLE, instead most use quadrature although a number of analysts are moving to use Bayesian modelling techniques. Mixed effects models structure data to be modelled in panels. A number of models exist for dealing with the independence violation incurred by longitudinal panel models; for example, generalised estimating equations (GEEs), which are estimated suing a variety of IRLS algorithms.

Bayesian models use a sampling algorithm known as Markov chain Monte Carlo (MCMC) to develop a posterior distribution of the data. The two foremost algorithms being Metropolis-Hastings and Gibbs sampling.

Maximum Likelihood Estimation

A probability distribution is itself defined in terms of the values of its parameter or parameters. We may define a probability distribution function for count models as:

$$f(y|\theta,\phi) \; or \; f(y;\theta,\phi)$$

Where y is the count response, θ is the connocial location parameter or link function, and ϕ is the scale parameter. Count models such as Poisson set the scale to a value of one; other more complex models have a scale parameter, and in some cases more than one.

A probability function generates or describes data on the basis of parameters. In modelling, we seek the value of the probability distribution as defined by specific unknown parameters, that makes the data we have most likely. In order to do this, we in effect invert the relationship of y and the PDF parameters, creating what is called a likelihood function. The likelihood function can be defined as:

$$L(\theta, \phi; y)$$

A probability function generates data on the basis of known parameters. A likelihood function determines parameter values on the basis of known data. When modelling we are asking what parameter values of a given PDF most likely generates the data we have to model.

Because of numerical considerations, statisticians maximise the log of the likelihood rather than the likelihood function itself. Maximum likelihood estimation (MCE) is in fact maximum log-likelihood estimation.

Maximization of the log-likelihood function involves taking the partial derivatives of the function, setting the resulting equation to 0, and solving for parameter values. The first derivative, with respect to the coefficients, is called the score function, U. the second derivative is a matrix called Hessian matrix. The standard errors of the predictors in the model are obtained by taking the square root of the diagonal terms of the negative inverse Hessian, $-H^1$.

In its simplest form, the Poisson probability distribution can be expressed:

$$f(y; \mu) = \frac{e^{-\mu i} \mu_i^{yi}}{y_i!}$$

Where y, represents a variable consistently of count values and μ is the expected or predicted mean of the count variable y. y!, meaning y-factorial, is the product of counts up to a specific count value, y. f(y;μ) indicate the probability of y given or based on the value of the mean. The subscripts indicate that the distribution describes each observation in the data.

Statas 'poisson (mu, y)' function provides a cumulative Poisson probability for a given mean and count term. 'poissonp(mu, y)' gives a specific probability for a mean any y value. Plotting the values allows us to observe the differences in the shapes of the distributions. Mean values under 1 are shaped like negative exponential distributions. The greater the mean, the more normal the shape of its appearance.

The equation for the Poisson log-likelihood, showing summation across the observations, can be expressed as:

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) - u_i - \ln(y!)\}$$

xβ= ln(μ). This entails that μ=exp(xβ). exp(xβ) is called the inverse link function, which defines μ, it also defines μ for the negative binomial model. Due to this relationship, μ=exp(xβ) is also called the exponential mean function.

The first derivative of the preceding log-likelihood function with respect to the coefficients (β), also called parameters when modelling, provided gradient of the Poisson log-likelihood:

$$\frac{\partial(\mathcal{L}(\beta; y))}{\partial \beta} = \sum_{i=1}^{n} (y_i - \exp(x_i'\beta))x_i'$$

The hessian matrix is calculated as the second derivative of the log-likelihood function and is negative definite for β for the Poisson it may be expressed as:

$$\frac{\partial(\mathcal{L}(\beta; y))}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} (\exp(x_i'\beta))x_i x_i'$$

Estimation of the maximum likelihood variance-covariance matrix is based on the negative inverse of the Hessian matrix, oft represented as ∑ (not to be confused as a sum symbol), given as:

$$\sum = -H^{-1} = [\sum_{i=1}^{n} (\exp(x_i'\beta))x_i x_j']^{-1}$$

The square roots of the respective terms on the diagonal of the negative inverse Hessian are the values of predictor standard errors. A Newton-Raphson type algorithm can be used for the maximum likelihood estimation of the parameters:

$$\beta_r + 1 = \beta_r - H^{-1}g$$

Which is the standard form of the maximum likelihood estimating equation. The algorithm typically updates estimates based on the value of the log-likelihood function. When the difference between old and updated values is less than a specified tolerance level- usually 10^{-6} – iteration stops and the values of the various statistics are at their maximum likelihood estimated values.

$\beta_0$indicates estimationa t the old value of the parameters (coefficients) $\beta_n$is the new, updated value. Likewise for likelihood, L. tol indicates the level of tolerance needed for convergence, set at 10^{-6}.

Iterative Reweighted Least Squares Algorithms

Poisson and negative binomial regression are often estimated using a generalized linear model. Stata uses 'glm' for an IRLS algorithm.

A standard formulation for one-parameter models is expressed as:

$$f(y;\theta) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\}$$

$\theta_i$is the canonical parameter or link function, defined as ln(mu) for the Poisson and NB models. b(θ_i) is the cumulant. The first and second derivatives and b(θ_i) define the mean and variance. C(y) is the nomralisation term, as given in the first equation.

Derivation of the IRLS algorithm is based on a modification of a two-term Taylor expansion of the log-likelihood function where for count models y is a vector of count values and θ is the parameter or parameters of the probability function generating y. The logic, where:

- g(μ) is the link function: for Poisson and NB, ln(mu)
- $g'(\mu)$ is the derivative of the link, or 1/mu
- $g^{-1}(\mu)$ is the inverse link function, exp(xβ)
- $b'(\theta)$ is mu, the mean
- $V\ is\ b"(\theta)$, the variance: Poisson: mu; NB:mu+a*mu2; PIG: mu+a*mu3
- Deviance: D=2{LLs-LLm}

Poisson Regression

Using a Poisson model on real study data is usually unsatisfactory due to the assumption of equidispersion.

All other count models are adjustments or variations from the basic Poisson model. Each of the following assumptions should be tested:

1. The distribution is discrete with a single parameter, the mean, which is usually symbolised as either λ(lambda) or μ(mu). The mean is also understood as the rate parameter.
2. The response terms, or y values, are non-negative integers, $Y \geq 0$.
3. Observations are independent of one another
4. No cell of observed counts has substantially more or less than what is expected based on the mean of the empirical distribution. As the value of μ increases, the probability of zero(0) counts is reduced.
5. The mean is the variance
6. The Pearson chi2 dispersion statistic has a value approximating 1.0. a value of 1.0 results when the observed and predicted variances of the response are the same

Assumption Three

There are two ways of testing for independence:

Check to determine whether the data are structures in panels. When we model panels as though they are independence, we say that the data is pooled.

Check the difference between the model SEs and the SEs adjusted by; employing a robust sandwich estimator to the SEs, bootstrapping the SEs, checking the SEs sealed by the dispersion statistic (model SEs multiplied by the square root of the Pearson chi2).

Assumption Four

Calculate the percentage of zeros in the empirical distribution and compare them with the frequency of zero counts expected based on a Poisson PDF with the mean determined for the observed distribution. If the frequencies substantially differ, a violation has occurred. A Chi2 test can be used for this assessment.

Assumption Five

Calculate the mean and the variance of the empirical count response. If they differ then the assumption has been violated.

Assumption Six

Estimate the full model. If the dispersion varies from 1.0, the model is ''poisson extradispersed''. A boundary likelihood ratio test can be used to assess overdispersion. A generalised Poisson model can test for the statistical significance of either under or over-dispersion.

Assumption Conclusions

If a model fails to violate, we model using a standard Poisson model. If it does violate we either employ an alternative count model or if no clear distributional violation is apparent and the dispersion still differs from one(1), the model may in fact be only apparently extradispersed.

Construct a ''True'' Poisson Model

To determine whether the single Poisson models are providing values that approximate the parameter values we assigned to the algorithm, we use a monte carlo programme algorithm.

Poisson Regression: Modelling Real Data

Output your dependent variable, summarise the variable to get the mean and variance and the % of zero counts (remember that the variance = the standard deviation squared). Calculate the expected number of zero counts, based on a Poisson distribution mean.

Most statisticians 'centre' a continuous predictor when it starts far from zero. Centring is the process where the mean of the variable is subtracted from every value of the variable. Centring changes only the value of the intercept in the model.

The bias resulting from overdispersion means that the p-values tell us nothing about the relationship of the predictor and response. If two models have the same explanatory power, the simpler model is preferred 'Occam's dictum'.

How to Interpret a Poisson Coefficient and Associated Statistics

The coefficient β_j in general is the change in the log-count of the response for a one-unit change in the predictor.

The standard errors of the model parameter estimates are obtained as the square root of the diagonal terms of the inverse negative of the Hessian.

The values calculated match the standard errors displayed in the coefficient table. These are model standard errors. The 95% CI= di coef +- 1.96*std.err.

Many prefer to model standard errors using a likelihood method because many coefficients are not distributed normally. In general, standard errors based on profile likelihood are preferable to traditional model-based standard errors.

Rate Ratios and Probability

In order to have a change in predictor value reflect a change in X, we must exponentiate the coefficient $-e^{\beta j}$. Using stata's 'glm' command, the 'eform' option exponentiates the coefficents and confidence intervals of the coefficients. The standard error is calculated using the delta method exp(β)*βSE. calculations can be verified by displaying the table of incidence rate ratios. The IRR indicates the ratio of the rate of counts between two ascending contiguous levels of the response. Most prefer to exponentiate coefficients and interpret parameter estimates as rate ratios.

Exposure: Modelling Over Time, Area, and Space

The rate of counts, μ, is calculated as the number of events counted divided by the period of time that counting occurs, and likewise for counts per area.

Statisticians use an offset with a model to adjust for counts of events over time periods, areas, and volumes. The model is sometimes referred to as a proportional intensity model.

μ is sometimes said to be an intensity or rate parameter, it is such only when thought of in conjunction with a constant coefficient, t. the rate parameterisation of the Poisson PDF can be expressed as:

$$f(y;\mu) = \frac{e^{-t\mu}(t\mu)^y}{y!}$$

Where t represents the length of time, or exposure, during which events or counts uniformly occur. When t=1, the model is understood to apply individual counts without a consideration of size. Where unequal periods of time, area, or volume, (TAV) occur in the model, an offset must be given.

Prediction

Predicted counts and their 95% CI may be obtained from:

Poisson…..;predict mu; predict eta, xb; predict se_eta, stdp; gen low=eta-invnormal(0.975)*se_eta; gen up=eta+invnormal(0.975)*se_eta; gen le:=exp(low); gen uc:=exp(up); sort mu

Twoway(line le:nov uc:eta, lpattern(dash1dash1)), ytitle("predictied count and 95%CI"); #delimit cr

Poisson Marginal Effects

Marginal effects pertain only to be continuous predictors. Discrete change or partial effects are used for binary and categorical predictors.

A marginal effect relates a continuous predictor to the predicted probability of the response variable. Other predictors are held at their mean, or median values. Basic interpretation of marginal effects: how the probability of the count response changes with a one-unit change in the value of the continuous predictor.

For count models, the marginal effect at the mean is defined as:

$$MEmean = \exp\,(x_i'\beta_k)\beta_k$$

Average marginal effects are defined as:

$$\beta_k\hat{y}$$

Discrete change is used to evaluate the change in predicted probability of the response when a binary predictor changes values from 0 to 1. To determine the partial effects, the predictor in Stata must specifically be made a factor variable. Average partial effects use the same method as for average marginal effects, except for the factoring.

There are several other ways to relate predictors and the response; elasticities and semi-elasticities (Hilbe 2011).

Testing Overdispersion

Basics of Count Model Fit Statistics

If we can determine the cause of overdispersion, we can employ the appropriate model to use on the data. The earliest fit test used with Poisson regression is called the deviance goodness-of-fit test. The test is based on the deviance statistic.

The deviance is defined as the difference between a saturated log-likelihood and full model log-likelihood. The saturated log-likelihood is calculated by changing every μ(mu) in the function to a y. represents a situation in which there is a parameter for every observation in the model. It indiciates a model with a perfect, but uninformative fit:

$$D = z\sum_{i=1}^{n}\{\mathcal{L}(y_i; y_i) - \mathcal{L}(\mu_i; y_i)\}$$

The Poisson log-likelihood function:

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n}\{y_i \log(\mu) - \mu_i - \log(y_i!)\}$$

The saturated log-likelihood function is ylog(y)-y-log(y!). subtracting this function by the last equation:

$$D = z\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)$$

Log(y!) is the normalisation term that provides for the function to sum to 1, cancels.

The deviance goodness-of-fit (GOF) test is based on the view that the deviance is distributed as Chi2. Chi2 has both the mean and the scale. For the deviance GOF, this is the deviance statistic and residual degrees of freedom. If the Chi2 p-value <0.05, the model is considered well fit.

With a p-value <0.05, the deviance GOF test indicates that we can reject the hypothesis that the model is not well fitted.

Statisticians have discovered that many models appearing to be well fitted on the basis of the deviance test in fact poorly fit the data. If the value of D is very large, then we can generally be safe in rejecting the goodness of the model fit.

''deviance is in effect a measure of the distance between the most full or complete (saturated) model we can fit and the proposed model we are testing for fit''. The smaller the distance, or deviance between them, the better the fit.

Many use Pearson Chi2 instead of a deviance GOF test. Pearson Chi2 defines overdispersion, as it is the squares residuals weighted by the model variance, and summed across all observations in the model:

$$x^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

The sum of squared residuals is an absolutely raw measure of the difference in observed versus predicted model counts, adjusted by both the variance and size of the model. Adjustment is made by dividing the squared residuals by the product of the variance and residual degrees of freedom. The result is the dispersion statistic, if >1 shows overdispersion, <1 shows under dispersion. In Stata following 'poisson' run 'estatgof' to produce both GOF tests.

Overdispersion

Apparent overdispersion can sometimes be identified and the model amended to eliminate it. Equidispersion can this sometimes occur from: adding appropriate predictors, constructing interactions, transforming predictors, transforming response, adjust for outliers, use correct link function.

Real overdispersion is a problem affecting the reliability of both the model parameter estimates and fit in general.

The Score and Lagrance multiplier tests, test for real overdispersion. A Score test is defined as:

$$z = \frac{(y - \mu)^2 - y}{\mu\sqrt{2}}$$

The test is post hoc. We model, then we predict μ, calculate z from the preceding formula, and regress z using linear regression. The test is based on two assumptions: the data set on which the test is used is large and z is t-distributed.

The Lagrange multiplier test is a Chi2 test, defined as:

$$x^2 = \frac{(\sum_{i=1}^{n} \mu_i^2 - n\hat{y}_i)^2}{2\sum_{i=1}^{n} \mu_i^2}$$

With one degree of freedom.

A majority consider the most important test of fit for a count model to be an analysis of the difference between observed and expected counts across the full range of counts in the data.

An excess of zero counts is a common reason for overdispersion. Using Stata, we can determine the observed and expected values for 0 counts as follows. (pg. 92 Hilbe).

Scaling Standard Errors: Quasi-Count Models

Scaling of standard errors was the first method used to deal with overdispersion. The method replaces the w, or model weight, in the IRLS algorithm when $\beta_r$are calculated:

$$\beta = (X'WX)^{-1}X'Wz$$

With the inverse square root of the dispersion statistic.

Scaling by the Pearson dispersion statistic entails estimating the model abstracting the dispersion statistic, and multiplying the model standard errors by the square root of the dispersion, then running one additional iteration of the algorithm:

$$\beta = (X'W_dX)^{-1}X'W_dz$$

Pearson based dispersion should always be used to assess could model over dispersion.

The R quasi-poisson family option is aimed to adjust for overdispersion in Poisson models, but it is simply scaling the standard errors. A table of IRR statistics can be formed using the 'eform' option or 'irr' option with 'poisson'.

Quasi-likelihood methods were first developed by Webberborn (1974). They are based on GLM principles but allow parameter estimates to be calculated based only on a specification of the mean and variance of the model observations without regard to those

specifications originating from a member of the single-parameter exponential family of distributions.

Quasi-likelihood models allow us to model data without explicit specification of an underlying log-likelihood function. The quasi-likelihood or the derived quasi-deviance function is then used in an IRLS algorithm to estimate parameters just as for GLM. When the mean and variance functions are those from a specific member of the exponential family. The quasi-likelihood is defined as:

$$Q(y;\mu) = \int_{y}^{\mu} \frac{y-\mu}{\Phi V(\mu)}$$

By talking the integral (y-μ)/μ from μ to y with respect to μ, the resultant equation is the Poisson log-likelihood but without the final ln(y!) normalising term.

The fact that the variance function is multiple by a constant changed the likelihood, or deviance function by dividing it by the scale(ψ). It is the next stage in amending the Poisson variance function to adjust for overdispersion.

We enter the Pearson dispersion statistic from the base model as the variance multiplier. This is the same as dividing the model standard errors by the square root of the dispersion. Compare the summary statistics of the two models- lower deviance indicates a better model fit. The quasi-likelihood model is not a true likelihood model and thus the standard errors are not based on a correct model-based Hessian matrix. The Stata command 'irls' is required for this option.

Unlike the standard variance estimator, $-H(\beta)^{-1}$, a robust variance estimator adjusts standard errors for correlation in the data. Robust standard errors should be used when the data are not independent, perhaps gathered over different household, hospitals, schools and so forth.

Modified sandwich variance estimators or robust cluster variance estimators provide standard errors that allow inference that is robust to within-group correlation but assume that clusters of groups are independent. Robust estimator's may be used with any maximum likelihood algorithm, not only GLM-based algorithms.

Statistics use empirical standard errors with Poisson regression as a catch-all adjustment for extra dispersion. If the model is in fact equidispersion. If the model is in fact equidispersed, the mode and SEs will be nearly identical. If the model is over- or underdispersed, using

robust SEs will provide more accurate information regarding the significance of the predictors in explaining the count response.

Non-parametric bootstrapping makes no assumptions about the underlying distribution of the model. SEs are calculated based on the data at hand. Samples are repeatedly taken with each sample providing model estimates. The collection of vector estimates for all samples is used to calculate a variance matrix from which reported SEs are calculated and used to determine Cis. Such Cis can be constructed from percentiles in the collection of point estimates, or from large sample theory arguments.

Bootstrapping has become a popular way of attempting to discover optimal SEs for model coefficients.

If the values of bootstrapped or robust standard errors differ substantially from model SEs, this is evidence that the count model is extradispersed. Use the bootstrapped or robust standard errors for reporting your model, but check for reasons why the data are overdispersed and identify an appropriate model to estimate parameters.

The likelihood ratio test for determining the inclusion or exclusion of predictors in a model is preferred over the standard Wald test, which is another way of saying regular predictor p-values.

Assessment of Fit

When modelling using either full Newton-Raphson maximum likelihood or IRLS, it is simple to calculate the linear predictor as:

$$x\beta = n = a + \beta_1 + \beta_1 + \cdots + \beta_n$$

Or

$$x\beta = n = \beta_0 + \beta_1 + \beta_2 + \cdots + \beta_n$$

Where a or β_0 is the intercept, defined as the value of the linear predictor of model observations when the value of each predictor is 0.

For members of the glm family, a link function, converts a linear predictor to a fitted or predictor value. The linear predictor and fit are essential components of all residuals.

The basic or raw residual is defined as the difference between the observed response and the predicted or fitted response. When y is used to identify the response, ŷ or μ is commonly used to characterise fit:

$$Raw\ Residual = y - \hat{y} \text{ or } y - \mu \text{ or } y - E(y)$$

In the preceding formula, the model variance function is V, the hat matrix diagonal as hat or h, and the SE of the prediction as stdp. A scale value ϕ, is user defined and is employed based on the type of data being modelled.

There are times when the Anscombe residual performs better than the standardised deviance (R^d). Anscombe residuals attempt to normalise the residual so that heterogeneity in the data, as well as outliers, become easily identifiable.

Anscombe defined the residual as:

$$R^A = \frac{A(y) - A(\mu)}{A'(\mu)\sqrt{V}}$$

Here $A(.) = \int V^{\frac{1}{3}}$. The calculated Anscombe residuals for the Poisson model are:

$$Poisson: 3(y^{2/3} - \mu^{2/3})/(2\mu^{1/6})$$

And for the negative binomial:

$$\frac{\{\frac{3}{a}[(1+ay)^{2/3} - \left(1+a\mu\right)^{2/3}] + 3\left(y^{2/3} - \mu^{2/3}\right)\}}{2(a\mu^2 + \mu)^{1/6}}$$

Analysists generally prefer to graph the standardised Pearson or Anscombe residuals by μ

We look for evidence of poor fit and non-random patterns by graphing residuals. Patterns typically mean that observations are not independent, it may also indicate that a predictor needs to be converted to another scale.

The test evaluates whether the predictors with drawn from a model should in fact have been retained. One may use a likelihood ratio test to determine whether data should be modelled using a Poisson or negative binomial regression.

The traditional likelihood ratio test is defined as:

$$LR = -2(\mathcal{L}_R - \mathcal{L}_F)$$

Where $\mathcal{L}_F$ is log-likelihood for.a full or more compelte model and $\mathcal{L}_R$ is the log-likelihood for a reduced model.

A Stata command 'lrdrop1' can be used for logit, logistic, and poisson to find predictors that together best fit the model.

The boundary likelihood ratio test (BLR) is a test used on negative binomial models to determine whether the value of the dispersion parameter, a, is significantly different from 0. The BLR equation is given as:

$$-2(\mathcal{L}_P - \mathcal{L}_{NB})$$

With $\mathcal{L}$ symbolising the log-likelihood function. The resulting value is measured by an upper tail Chi2 distribution with (1) degree of freedom. Only one half of the full distribution is used therefore the Chi2 test is divided by two.

As a key to remember, given the value of -2 times the difference in the log-likelihood values, if the difference in Poisson and negative binomial log-likelihoods is less thean 1.352, the model is Poisson; it is greater the data needs to be modelled other than Poisson.

Model Selection Criteria

The Akaike Information Criterion (AIC) is found in two forms traditional:

$$AIC = -2\mathcal{L} + 2k = -2(\mathcal{L} - k)$$

And the version witht eh main AIC terms divided by n, the number of observations in the model:

$$AIC = \frac{-2\mathcal{L} + 2k}{n} = -2(\mathcal{L} - k)/n$$

Where $\mathcal{L}$ is the model log-likelihood, k is the number of predictors and n the number of observations in the model. 2k is reffered to as a penalty term, which adjusts for the dimension of the model. As we increase the number of predictors, $-2\mathcal{L}$ becomes smaller. The penalty, 2k, is added to the log-likelihood to adjust for this possible bias.

AIC is also used by some to compare models of different sample size.

The Bayesian Information Criterion (BIC) is formulated as:

$$BIC = -2\mathcal{L} + klog(n)$$

With k indicating the number of predictors, including the intercept, and n the number of observations in the model. Statisticians prefer the Schwarz BIC over the traditional BIC.

The stata command 'abic' display the values of most used parameterisations of the AIC and BIC, 'estat ic' is an alternative.

Negative Binomial Regression

The traditional parameterisation of he negative binomial is also known as NB2 negative binomial model, based on the value of the exponent in its second term. An NB1 model has also been formulated for which the second terms exponent has a value of 1. Both NB1 and NB2 use a maximum likelihood algorithm for estimating parameters but NB2 may also be estimated using an IRLS algorithm within the scope of generalised linear models.

The negative binomial is a two-parameter model- with mean(μ) and dispersion (a) parameters.

Zero-truncated and zero-inflated models have been designed for this purpose of dealing with data with more zeros than are allowed by Poisson or negative binomial distributional assumptions.

The NB2 has the same distributional assumptions as the poisson distribution with the exception that it has a second parameter- which provides for a wider shape to the distribution of counts than is allowed under Poisson assumptions. The negative binomial allows us to model a far wider range of variability than the Poisson.

The negative binomial model is nearly always used to estimate the parameters of overdispersed Poisson data.

The probability distribution can be expressed in a variety of ways, a common parameterisation appearing as:

$$f(y;\mu,a) = \left(\frac{y_i + \frac{1}{a} - 1}{\frac{1}{a} - 1}\right)\left(\frac{1}{1+a\mu_i}\right)^{\frac{1}{a}}\left(\frac{a\mu_i}{1+a\mu_i}\right)^{y_i}$$

When the negative binomial is estimated using a full maximum likelihood algorithm, both μ and the dispersion parameter a are estimated. When estimated using glm, only μ is estimated; a must be inserted into the algorithm as a constant.

When the value of a approaches 0, the model is Poisson. For θ, when θ approaches infinity, the model is Poisson. When a Poisson model is overdispersed, the Poisson dispersion statistic, Pearson Chi2/(n-r), is greater than 1, and the negative binomial value of a is greater than 0. A true Poisson model has a Poisson dispersion statistic of 1 and negative binomial dispersion parameter of 0.

Unless your Poisson or negative binomial model is well fitted and meets its respective distributional assumptions, use robust or empirical standard errors as a default.

The Stata command 'count fit' obtains an overview of the differences between Poisson and Negative binomial models. The IRR + SEs are listed for all predictors. It helps to graphically view comparisons between observed and predictors counts for dependent variable.

There are times when the NB1 fits better than the NB2. The NB-P has a second dispersion parameter that allows the dispersion to vary across observations. The variance functions of NB1 and NB2 are:

$$NB1 = \mu + a\mu \ or \ \mu + a\mu^1$$

$$NB2 = \ \mu + a\mu^2$$

The difference rests in the value of the exponent. To close between them, Greene let the exponent become a parameter to be estimated, called ρ for power(rho). NB-P:

$$NB - P = \ \mu + a\mu^p$$

Another test statistic allows us to determine whether NB2 is preferable to NB1, it is called a reverse cumulative upper tail students t-test with 1 degree of freedom:

$$\frac{nbp - nb2}{nbpse}$$

In order to determine the source of over- or underdispersion it may help to discover how the predictors contribute to the dispersion parameter. Knowing a specific predictor significantly contributes to the dispersion parameter allows the analyst to explore the predictor in more detail.

Some major points regarding heterogenous negative binomials are:

- Dispersion predictors (ln alpha) specify which predictors most influence the value of the dispersion parameter

- Can use NB-H to verify preconceived sources of overdispersion in the data
- Robust SEs must be used as default

Poisson Inverse Gaussian Regression

The Poisson Inverse Gaussian (PIG) is similar to the negative binomial in that both are mixture models. The negative binomial is a mixture of Poisson and gamma distributions, whereas the PIG is a mixture of Poisson and inverse Gaussian distributions, with an inverse Gaussian variance of $\frac{\mu^3}{\phi}$.

The dispersion for PIG will be given the name alpha, a. The dispersion is often given a nu,v,or phi ϕ . Use ϕ to refer to the dispersion apramteterisation indirectly with the mean such that $\phi = 1/a$. GLM theory symbolised the exponential family link function as θ.

Greater values of the mean of the response variable in a PIG regression provide for adjustment of greater amounts of overdispersion than does the negative binomial model.

The key difference in assumption for PIG over negative binomial is PIG regression is sued to model count data that have a high initial peak and that may be skewer to the far right as well as data that are highly Poisson overdispersed.

The PIG probability distribution, as a variety of Sickol distribution can be given as:

$$f(y;\mu,a) = \sqrt{\frac{\phi}{2\pi y^3}\exp}\ (\frac{-\phi(y-\mu)^2}{2\mu^2 y})$$

With {y,μ, ϕ}>0.

Constructing and Interpreting the PIG Model

Stata commands; 'pigreg', 'zipig', and 'ztpig' offer ability to model PIG regression, zero-inflated PIG, and zero-truncated PIG.

The 'linktest' evaluates whether the assumption of linearity has been violated. If the square of the hat matrix diagonal is significant then the assumption has been violated.

Problems with Zeros

When there is a substantial disparity between the expected and observed zero counts in the data, given the mean of the response variable and number of observations in the model, a Poisson distribution should likely not be used to model the data.

Two-part Hurdle Models

The foremost use of a hurdle model is to deal with count response variables that have more-or fewer- zero counts than allowed by the distributional assumptions of the count data. Two types are used to handle excess zeros: hurdle models and zero-inflated models.

The idea of a hurdle model is to partition the model into two parts- first, a binary process generating only positive counts. The binary process is typically modelled using a binary model, and the positive count is modelled using a zero-truncated model. Most commonly used hurdle models are; Poisson-logit, NB2 logit, and NB2-probit models.

The two processes are conjoined using the log-likelihood:

$$\mathcal{L} = \ln\big(f(0)\big) + \{\ln[1 - f(0)] + lnP(t)\}$$

Where f(0) represents the probability of a zero count, and P(t) represents the probability of a positive count. The hurdle model log-likelihood is the log of the probability of y=0 plus the log of y=1 plus the log of y being a positive count.

In case of a logit model, the probability of zero:

$$f(0) = P(y = 0; x) = \frac{1}{1 + \exp\,(x_i'\beta_b)}$$

And 1-f(0) is:

$$\frac{\exp\,(x_i'\beta_b)}{1 + \exp\,(x_i'\beta_b)}$$

Which is the probability of y=1

To conduct the models separately to test 'hplogit', create a variable equal to 1 dependent variable >0 and 0 otherwise. Model x on predictors- the binary component of the mode. Model dependent variable >0 on predictors using a zero-truncated Poisson. Compare with hurdle model.

Each predictor is evaluated in terms of the contribution it makes to each respective model.

The best action regarding predicting is to predict for a specified component in a multi-component model.

Marginal effects can also be obtained. It is easiest to obtain separate marginal effects for each component of the model.

The PIG-logit hurdle model will likely be of use to those who have a substantial amount of variability in the model, more than is accounted for by simply adjusting the zero components in the model.

Zero-inflated Mixture Models

It is not always desirable to use zero-inflated models if a negative binomial or hurdle model will be all that is needed. The analyst should have a theory as to why there are a class of observations having both observed and zero-counts. Zero-inflated models can be thought of as finite-mixture models.

There are two different types of 0's in the data- one generated by a binary component modelling 0's and one in the count model component fo the mixture model. Binary 0's ''bad zeros'', count model 0's ''good zeros''.

To evaluate zero-inflated models we use boundary likelihood tests- a test of one ZI model against another that is presumably nested in it. A Vuong test- non-nested test of a zero-inflated model adjusts a non-inflated model. And AIC/BIC tests to be certain to check whether a standard non-inflated model might fit better than a zero-inflated model.

The first component is the binary, usually modelled as a logit or probit. The binary component has a value of 1 for all 0's int eh data and 0 for all other counts greater than zero. The count component simply models all of the counts from zero to greater than zero. The logic of a zero-inflated model is that counts are estimates as:

$$\begin{aligned}\Pr(Y = 0) &= \Pr(Bin = 0) + (1 - \Pr(Bin = 0)) * \Pr(Count = 0)\Pr(Y > 0) \\ &= (1 - \Pr(Bin = 0)) + PDFCount\end{aligned}$$

Since we will use the logit models for the binary component of zero-inflated models, the logit equation for the probability of 0's as $1/(1+\exp(x\beta)$:

$$\Pr(Bin = 0) = \frac{1}{1 + \exp(x\beta)}$$

For a zero-inflated Poisson model with a logit binary component we have the equation:

$$(y = 0) = \log\left(\frac{1}{1 + \exp(-x\beta_b)} + \frac{\exp(-\exp(x\beta))}{1 + \exp(x\beta_b)}\right)$$

$$(y > 0) = \log\left(\frac{1}{1 + \exp(-x\beta_b)}\right) - \exp(x\beta) + y(x\beta) - log\tau(y + 1)$$

Where b is the subscript indicates that x is a binary model component and without subscript b, that xβ is from the count component, in this case Poisson.

The Vuong statistic is biased toward the zero-inflated model because the same data are used to estimate both the binary and count component parameters.

ZINB models may be compared with ZIP models using the boundary likelihood test, with p>0.05 indicating that the ZINB is preferable to ZIP. ZINB models can be compared to NB2 models with Vuong tests.

If the counts are distributed such that there are many counts in the ower range of numbers, with a long right skew, then a PIG model amy be preferred.

If there is a real separation of mechanisms producing the 0's and the positive counts, a hurdle model appears the best. If there is an overlap of 0 values, then a zero-inflated model be best. Keep in mind that a hurdle model is a two-part model, and a zero-inflated model is a mixed model.

What Model is Probably Best?

Response	Example Models
1: No Zeros	Zero-truncated Models (ZTP; ZTNB)
2: Excessive Zeros	Zero-inflated (ZIP; ZINB; ZAP; ZANB); Hurdle Models
3: Truncated	Truncated Count Models

Response	Example Models
4: Censored	Econometric and Survival Censored Count Models
5: Panel	GEE; Fixed, Random, and Mixed Effects Count Models
6: Separable	Sample selection, Finite Mixture Models
7: Two-Responses	Bivariate Count Models
8: Other	Quantile, Exact, and Bayesian Count Models

Modelling Underdispersed Count Data – Generalised Poisson

An analyst rarely encounters Underdispersed Poisson data when dealing with real data- but it can happen. When you do not pay attention t Underdispersed data the standard errors of the resulting model are overestimated. This leads to thinking that predictors are not significant when in fact they are.

The types of count data that are underdispersed consist of data that are lumped more tightly together than should be expected based on Poisson and negative binomial distributional assumptions.

When $\delta = 0$, the model is equidispersed (i.e, it is Poisson). When $\delta > 0$, the model is overdispersed; if $\delta < 0$, the model is Underdispersed.

Complex Data: More Advanced Models

Small and Unbalanced Data – Exact Poisson Regression

Exact statistics is a highly iterative technique that uses the conditional distributions of the sufficient statistics of the model parameters, assuring that the distribution is completely

determined. Exact statistics is not a maximum likelihood method, which relies on asymptomatic standard errors to determine the significance of predictors.

When dealing with extradispersed data we can use either scaling or robust sandwich estimators. Both methods are typically the same- though statisticians tend to prefer using robust adjustment rather than scaling.

Modelling Truncated and Censored Counts

There are many times when certain data elements are lost, discarded, ignored, or otherwise excluded from analysis. Truncated and censored models have been developed to deal wit these types of data. Both models take three forms: truncation or censoring from below, above, and at the endpoints of an interval of counts. Count model forms take their basic logic from truncated and censored continuous response data, in particular from Tobit (Amemiya 1984) and censored normal regression (Goldberger 1983), respectively.

The essential difference between censored and truncated models relates to how the response values beyond user-defined cut points are handled. Truncated models eliminate the values altogether; censored models revalue them to the value of the cut point. In both cases, the probability function and log-likelihood functions must be adjusted to account for the change in the distribution of the response.

Truncated Count Models

Starting with the basic Poisson probability mass function, defined as:

$Prob(Y = y) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, y=0,1,…

When discussing zero-truncated Poisson models in section 7.1, we adjusted the preceding Poisson distribution to account for the structural absence of zeros. We discovered that the probability of a zero count fo the Poisson distribution is exp $(-\mu)$, for the negative binomial $(1 - a\mu)^{-1/2}$, ad for the PIG exp $(\left(\frac{1}{a}\right) * \left(1 - sqrt\left(1 + \frac{2}{a\mu}\right)\right)$.

When truncating a zero count from the Poisson distribution , the probability of zero is subtracted from 1, and the result is divided into the full Poisson PDF; that is, $PDF/(1 - \exp(-\mu))$. The same is the case for other distributions, except that the formula differs.

Going farther from 0, the Poison probability of 1 Is $\mu * \exp(-\mu)$. If truncation is at 1, then both 0 and 1 must be excluded from the distribution and an adjustment for both must be made in the resulting adjusted PDF. We do this by summing the two probabilities and subtracting from the sum of 1. This value is divided into the Poisson PDF:

$Prob(Y = (y = 0,1)) = \frac{e^{-\mu_i}\mu_i^{y_i}}{(1-(e^{-\mu_i}+\mu_i e^{-\mu_i}))y_i!}$, y=2,3,…

This distribution can be called a left-truncated at 1 Poisson distribution. When establishing a left truncation at point 1, we place a cut point, C, at 1, and the first number to be in the nontruncated distributions is C+1.

The left-truncated Poisson PDF in general is Poisson PDF/Prob(y>C). Numerically this appears as:

$$Prob(Y = y \mid \mathrm{Y} > \mathrm{C}) = \frac{\frac{\exp(-)\mu^y}{y!}}{1 - \sum_{j=0}^{C}\frac{\exp(-\mu)\mu^j}{j!}},\ for\ y = C+1, C+2, \ldots$$

When a cut point is on the right side, it is the higher-value end of the sorted distribution:

$$Prob(Y = y \mid \mathrm{Y} < \mathrm{C}) = \frac{\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}}{Prob(y_i < C)} = \frac{\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}}{\sum_{j=0}^{C-1}\frac{e^{-\mu_i}\mu_i^{j_i}}{j_i!}},\ for\ y = 0,1, \ldots, C-1$$

Employing the user-created Stata *treg* command (Hardin and Hilbe 2014b), you can create left and right cut point-based models.

If for example we wanted to model visits to the doctor, but only for patients who have in fact visited a physician during the calendar year 1984. More-over, for our example, suppose also that visits were not recorded for more than 8 visits. We then model the data with a left truncation at 0 and right truncation at 19. This is called interval truncation.

Censored Count Models

The censoring that is understood when dealing with count models differs from the censoring that occurs in survival models. A survival-type parameterisation for censoring was developed by Hilbe (1998) and is discussed in detail in Hilbe (2011). The difference is

that instead of truncated values being excluded from the truncated distribution, censored values are revalued to the value of the cut point. The distinction is subtle but important:

Left censoring: Left: $(\leq C)$,

If C=3, 3 is smallest value in the model. Values that may have been lower are revalued to C. Any response in the data that is less than 3 is also considered to be less than or equal to 3.

Right censoring: Right: $(Y \geq C)$,

If C=15, 15 is the highest observed value in the model. Values that may have been greater in the data are revalued to the value at C. Any response in the data that is greater than 15 is also considered to be greater than or equal to 5.

Right censoring is a more common application when using censored Poisson or negative binomial regressions, whereas left truncation is more commonly used with truncation models.

cpoissone is a Stata command used to model censored models.

Poisson-Logit Hurdle at 3 Model

An extended type of hurdle model that branches over to finite mixture models, which we discuss next, can be created by setting the hurdle at a higher place in the range of counts than at 0.

Counts with Multiple Components – Finite Mixture Models

Finite mixture models have been developed to model a situation where we suspect that the response variable of our model consists of counts that have been generated from different data-generating mechanisms. That is, when the data to be modelled are generated from more than one source.

Adding Smoothing Terms to a Model – GAM

Generalised additive models (GAMs) are a class of models based on generalised linear models for which the linear form of the model, $\sum x\beta$, is replaced by a sum of smoothed functions, $s(X)$. The method is used to discover non-linear covariate effects that may not be detectable using traditional statistical techniques.

The key concept in GAM modelling is that the partial residuals of continuous predictors in a model are smoothed using a cubic spline, loess smoother, or another type of smoother while being adjusted by the other predictors in the model. The parameters of the smooths are related to the bandwidth that was used for the particular smooth. The relationship that is traditionally given for the GAM distribution is:

$$y = \beta_0 + \sum_{i=1}^{j} f_j(X_j) + \varepsilon$$

The purpose of using GAM is to determine the appropriate transformation needed by a continuous predictor in order to affect linearity. A GAM employs the iteratively reweighted least squares (IRLS) algorithm used in GLM models for estimate. At each iteration, the partial residuals of each relevant continuous predictor in the model are smoothed. Partial residuals are used since they remove the effect of the other model predictors.

When All Else Fails: Quantile Count Models

When we simply cannot obtain a reasonably fitted Poisson, negative binomial, PIG, or some variation of these models based on a PDF or likelihood function, we may try to model the empirical distribution of counts without assuming an underlying probability or likelihood function. That is, we discover a shape to the distribution as we find it and model it using quantile techniques.

The idea is to model the discrete data with a jitter in such a manner that the distribution appears as a continuous variable. The response is structured as:

$$z = y + uniform(0,1)$$

Z is linearized at the conditional mean of each quantile of the distribution as exp(x), which keeps the distribution positive.

The traditional quantile regression is based on the median, or other quantiles, of a Gaussian distribution.

If the coefficients and standard errors are the same when modelling with fewer iterations, then an analyst can be more confident that the model is working correctly. Marginal effects may be calculated using the Stata command *qcount_mfx.* The model should be compared with a negative binomial model on the same data.

A Word About Longitudinal and Clustered Count Models

There are two broad types of models that are commonly used for longitudinal and multilevel modelling. The first is generalised estimating equations (GEEs), which is a population-averaging method of estimation. GEE models are not true likelihood-based models but rather are examples of quasi-likelihood models. The other type of model is referred to as a subject-specific model. Most random- and mixed-effects models are in this class of models.

Generalised Estimating Equations (GEEs)

GEE models are an extension of the generalised linear model (GLM) where the variance function is adjusted using a correlation matrix. Several standard correlation structures are used for GEE analysis:

Independence- an identity matrix, no correlation effect is specified at all

Exchangeable –a common correlation value provided to each panel or cluster in the data

Autoregressive – lag correlation for longitudinal and other time-dependent variables

Unstructured – separate correlation values for each panel or cluster in the data

Stata's *xtgee* command is a full GEE package.

Mixed-Effects and Multilevel Models

Mixed-effects models are panel models that are combinations of fixed and random effects, both of which have models named for them. Fixed effects are generally regarded as emphasising the measurements of the effects themselves. Random effects are generally structured so that they are normally distributed with a mean of 0 and a standard deviation of σ^2. The random effects themselves are not generally estimated directly but are summarised based on their underlying variance-covariance matrices.

Random-effects models are often divided into two categories – random intercept, and random slopes or coefficients. A random-intercept model is the simplest random-effects model, being structured so that only the intercepts are random. They vary in value across panel intercepts if the coefficients themselves vary between panels, the model is a random-slopes model.

Missing Data and Multiple Imputation

Foundations and Introduction

Missing data has meant that the available data for analysis were unbalanced, thus complicating the planned analysis and in some instances rendering it unfeasible. The wider question of the consequences of nontrivial proportions of missing data for inference was neglected until a seminal paper by Rubin (1976). This set out a typology for assumptions about the reasons for missing data and sketched their implications for analysis and inference.

Missing data is an essential component of any longitudinal data analysis – the major concern being that missing data and non-response is bound to affect the inferences made by the analysis of longitudinal studies (Hawkes and Plewis, 2006: 479; Silverwood *et al.*, 2021). The various factors that account for sample attrition in the datasets outlined above have the potential to present real issues as it relates to comprehensive data analysis. For the purposes of analysis those that exit the sample due to death or emigration are considered 'natural' exists from the original sample. Those, however, that either cannot be found, reject continued participation etc. are individuals that we hold partial data on – being able to utilize this partial data within my analysis could be beneficial.

When dealing with missing data there are three primary types of classification. The first is missing completely at random (MCAR), meaning that missingness does not depend on observed or unobserved values. The second, being missing at random (MAR), meaning that given observed values missingness does not depend on the unobserved ones. Finally, missing not at random (MNAR) meaning that missingness depends on unobserved values (Silverwood et al 2021). If data is found to be MAR then approaches like multiple imputation (MI), inverse probability weighting are made available – the former being extensively documented with the NCDS in particular in (Hawkes and Plewis 2006).

When dealing with missing data there are multiple methods to tackle the problem. The first is listwise deletion. Listwise deletion removes all observations from the data which have a missing value in one or more of the variables included in analysis. This is also known as Complete Records Analysis (CRA). The CRA approach is unpredictable, there is no way to know the consequences for this loss of information (Carpenter and Kenward, 2012).

A second method that deals with missing data is the use of survey weights. Survey weights take into account missingness, Inverse Probability Weighting (IPW) creates weighted copies of complete records so as to remove selection bias introduced by missing data. Whilst IPW is a method of dealing with missing data, alternatives such as multiple imputation are regarded as much more efficient (Seaman *et al.*, 2012; Seaman and White, 2013).

A third method involves Multiple Imputation (MI). This method substituted missing data with substituted values. MI is an attractive method because it is practical and widely applicable (Carpenter and Kenward, 2012).

Full-information Maximum Likelihood (FIML) is another method for dealing with missingness. For the regression based analysis including interactions with data from at least two stages of the life course, (Silverwood *et al.*, 2021) as the current analysis is, multiple imputation is plausible and more flexible than FIML. This flexibility stems from the ability to include auxiliary variables more easily within the imputation phase as well as being readily able to after imputing data sets obtain point estimates and standard errors at ease (Carpenter and Kenward, 2012). Recently, there has been some debate surrounding FIML vs MI approaches.

Paul Allison in a series of articles (Allison, 2012a, 2012b, 2015) argues that FIML is 1) simpler to implement, 2) FIML has no incompatibility between an imputation model and an analysis model, 3) FIML produces a deterministic result rather than a different result every time, and 4) FIML is asymptomatically efficient. Firstly, MI does have greater variability than FIML but that increased choice in model selection is not necessarily a negative so long as proper procedures are followed – in fact greater variability of choice has the potential to make MI a more attractive candidate for dealing with missingness over FIML. Secondly, MI models only run into an incompatibility problem when the MI model is inconsistent from the CRA model – something that with appropriate testing and open science practices detailing the model construction, shouldn't happen. Thirdly, MI models are deterministic provided the same seed is used each time you run the imputation. The only time this would not be plausible would be a scenario where open science practices were not followed, and fellow researchers could not access the MI seed. Finally, the argument that FIML is asymptotically efficient only holds true to a certain extent. MI models reach asymptotic efficient by running an infinite number of imputations – though you can reach near to full efficiency with a relatively small number of imputations, Allison (Allison, 2015) argues around 10. Overall, whilst FIML does offer some advantages, there is nothing so considerable as to desire FIML over MI. So long as open science procedures are upheld, most major critiques of MI are dealt with. As such subsequent analysis uses CRA and MI to compare the substantive conclusions between the two and to understand if missingness impacts interpretation.

When dealing with MI the subsequent question that naturally follows is how many imputations is sufficient? Silverwood et al (2021) suggest that anything around 50 imputations would be sufficient for reliable estimation of point estimate and estimating p-values with little error. Though sometimes with large samples with sizeable missingness more imputations may be required.

Multiple Imputation by Chained Equations is a tool developed to address missing data on all variables within a given model at the same time. It does this by filling in missing values in multiple variables iteratively by using chained equations Multiple imputation models are estimated using the mi suite in Stata. This suite is compatible with the svy suite and so can also adjust for complex survey design.

Whilst multiple imputation does help when it comes to missingness, it does have some drawbacks. Goodness-of-fit statistics for example are not able to be used – R2 and BIC most prominently. Therefore, it is not possible to assess the more appropriate or parsimonious

model – it is simply possible to compare the substantive effects between a complete records analysis and a multiple imputation model. For multiple imputation models to be compared to a complete records analysis the former needs to be ''congenial'' (White, Royston and Wood, 2011) with the latter. Congeniality or consistency in this respect means that the same variables that are in the complete record analysis are identical to those included in multiple imputation. If the variables between complete records analysis and multiple imputation models differ then the correct variance/covariance matrix will not be estimated and a substantive comparison between the two will become impossible and impracticable due to a loss of statistical power (Von Hippel, 2009; Lynch and Von Hippel, 2013).

Multivariate imputation by chained equations (MICE) is a form of multiple imputation that fills in or imputes missing data within a given dataset through iterative predictive models or *k* imputations. This specification is required when imputing a variable that must only take on specific values such as the categorical nature of the economic activity response variable within the current analytical model. Using MICE, each imputation *k* is drawn from the posterior distribution of the parameters in the given imputation model, then the model itself is imputed (Carpenter and Kenward, 2012). To create the *k*th imputation new parameters are drawn from the posterior distribution. Multiple Imputation following MICE draws from Bayesian influences on the distribution of missing data upon observed data. An important advantage of Multiple Imputation is that it can be applied for data missing at the response variable or its covariates (Carpenter and Kenward, 2012).

Choosing the number of imputations is difficult. Previous literature on the topic suggests that anywhere between 3-5 imputations is sufficient to obtain acceptable properties (Carpenter and Kenward, 2012). Though some modern literature suggest closer to 50 imputations (Silverwood *et al.*, 2021). However, if there is a desire to estimate small p-values or have an MI estimator of the fraction of missing information, greater numbers of imputations are required. Carpenter and Kenward (2012) suggest two routes. If an analysis after imputation is clear-cut after a small number of imputations, there is no need to perform more. If, however after imputation the inference is less clear-cut take K = 100, or 100 imputations. Others promote a slightly different interpretation. White et al (2010) and Bodner (2008) suggest using the Fraction of Missing Information (FMI) as a baseline for the minimum required imputations. If the maximum FMI in a given model is 44 per cent then 44 imputations is at minimum suggested. When following this assumption White et al (2010) found standard errors and p-values were considerably reduced and stabilised.

After Multiple Imputation is performed, four key statistics are relevant to focus upon: variance total, Relative Variance Increase (RVI), Fraction of Missing Information (FMI), and the Relative Efficiency (RE).

The primary usefulness of multiple imputation relies upon its variance estimation. The total variance in multiple imputation is the sum of multiple sources of variance: within imputation variance, between imputation variance and additional sampling variance. The latter of which is calculated by the within imputation variance divided by the number of imputations. The variance total is directly related to how standard errors are calculated.

Unlike simple imputation methods, multiple imputation estimates SEs in such a way that the SEs for each parameter estimate are the square root of their variance totals.

The RVI or Relative Variance Increase is the proportional increase in total sampling variance that is due to missing information. Any variable that has a large amount of missingness or are weakly correlated with other variables in the imputation model with tend to have larger than average RVIs. Weakly correlated auxiliary variables will always trend towards large RVIs.

The FMI is related to the RVI (which in turn is related to the variance total). The FMI is the proportion of the total sampling variance that is due to missing data. It is estimated based upon the percentage of missingness for a particular variable and how correlated this variable is with other variables in the imputation model. When a variable has a high FMI this can be an indicator of a problematic variable which may cause convergence issues.

Finally, the relative efficiency or RE relates to how well the true population parameters are estimates. It is related to both the amount of missingness as well as the number of imputations within an imputation model. The RE is a comparative measure. It compares the relative efficiency of the current model variable to performing an infinite number of imputations. It is relatively easy to achieve a high RE on a given imputation model with a small number of imputations however this does not mean that the standard errors within the given imputation model will be calculated accurately.

Auxiliary variables are variables in the data set that are either correlated with a missing variable or variables but are not a part of the main analytical model of interest. They are included within the imputation model to increase accuracy and statistical power to make the MAR assumption more plausible. Making the MAR assumption more plausible is done by including auxiliary variables – variables that can be used to predict missingness on a given variable. Auxiliary variables are important when there are high levels of missingness upon a given variable (Johnson and Young, 2011; Young and Johnson, 2011). There is no strict threshold for what an auxiliary variable needs to be in order to be included within the imputation however some have recommended an $r > 0.4$ on at least one of the analytical variables within the model (Allison, 2012a). Though this is disputed (Enders, 2010). Others, such as Silverwood et al (2021) instead argue that if an auxiliary variable is predictive of the outcome variable then that makes them suitable for inclusion within the imputation model. An auxiliary variable does not have the requirement that the given variable has to have complete information to be valuable – auxiliary variables can still effective when they have missingness (Enders, 2010).

Prior to imputation it is best to explore the distribution of variables comparative to complete and non-complete cases. In the presence of a MCAR mechanism all distributions should be the same comparatively. If this is not the case, then this is suggestive of a MAR or MNAR mechanism. These imbalances present themselves in every variable within the model except for sex. This is unsurprising considering that sex as a variable presents zero missingness. The distributions of the variables thus far present some indications of a MAR or MNAR mechanism being present.

With all the variables in the model being categorical in nature, convergence issues are a possibility. This risk is increased if a model has many categorical variables. Failure to converge was a consistent problem. Without resorting to re-coding analytical variables, the decision was made to drop one of the auxiliary variables in order to produce an imputed model[1].

After performing the imputation, it is often useful to graph the means and standard deviations saved through the tracing subcommand when using MICE – autocorrelation plots would be useful but are only available for non-MICE related imputations. By graphing variables means and standard deviations through trace plots for example over each imputation, any discrepancy or deviation can easily been found. If this were to be a case this would be problematic for the imputation model and suggest that further imputations would be required (White, Royston and Wood, 2011). The means and standard deviations of imputed values from each iteration[2] were checked to see the distributions of each variable against the imputations. These graphs are seen below. To note, due to the sex variable having zero missingness, no graph was produced as no imputations on that variable were required. As illustrated all analytical variables that were imputed have a relatively stable mean and standard deviation across the iteration numbers.

Multiple imputation (MI) is attractive because it is both practical and widely applicable.

Reasons for Missing Data

All datasets consist of a series of units each of which provides information on a series of items. Within this framework, it is useful to distinguish between units where all the information is missing, termed unit nonresponse, and units who contribute partial information, termed item nonresponse. The statistical issues are the same in both cases, and both can in principle be handled by MI- though much of the focus of MI rests with the latter.

Patterns of Missing Data

It is very important to investigate the patterns of missing data before embarking on a formal analysis. Key questions concern the extent and patterns of missing values, and whether the pattern is monotone, as if it is, this can considerably speed up and simplify the analysis.

Missing data in a set of p variables are said to follow a monotone missingness pattern if the variables can be re-ordered such that, for every unit i, and variable j.

[1] The variable in question was acatnn236, a categorical variable.
[2] Burnin was 20 during imputation.

- If unit i, is observed on variable j, where j=2,…,p, it is observed on all variables j'<j, and
- If unit i, is missing on variable j, where j=2,…,p, it is missing on all variables j'>j

A natural setting for the occurrence of monotone missing data is a longitudinal study, where units are observed either until they are lost to follow-up, or the study concludes. A monotone pattern is thus inconsistent with term missing data, where units are observed for a period, missing for the subsequent period, but then observed.

Consequences of Missing Data

We may think of the missing data mechanism as a second stage in the sampling process, but one that is not under our control. It acts on the data we intended to collect and leaves us with a partially observed dataset. The missing data mechanism cannot usually be definitively identified from the observed data, although the observed data may indicate plausible mechanisms. Thus, we will need to make an assumption about the missingness mechanism in order to draw inference. The process of making this assumption is quite separate from the statistical methods we use for parameter estimation. Further, to the extent that the missing data mechanism cannot be definitively identified from the data, we will often wish to check the robustness of our inferences to a range of missingness mechanisms that are consistent with the observed data.

Due to the mechanisms causing the missing data being rarely able to be definitively established we will often wish to explore the robustness of our inferences to a range of plausible missingness mechanisms- a process we call sensitivity analysis.

From a general standpoint, missing data may cause two problems: loss of efficiency and bias.

First, loss of efficiency, or information, is an inevitable consequence of missing data. Unfortunately, the extent of information loss is not directly linked to the proportion of incomplete records. Instead, it is intrinsically linked to the analysis question.

Faced with an incomplete dataset, most software automatically restricts analysis to complete records. The consequence of this for loss of information is not always easy to predict. Nevertheless, in many settings it will be important to include the information from partially complete records. Not least of the reasons for this is the time and money it has taken to collect even the partially complete records. Second, the subset of complete records

may not be representative of the population under study. Restricting analysis to complete records may then lead to biased inference. The extent of such bias depends on the statistical behaviour of the missing data. A formal framework to describe this behaviour is thus fundamental and such a framework was originally elucidated in a seminal paper by Rubin (1976).

Inferential Framework and Notation

We suppose we have a sample of n units, which will often be individuals, from a population that for practical inferential purposes can be considered infinite. Let $Y_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,p})^T$ denote the p variables we intended to collect from the ith unit, i=1…,n. We wish to use these data to make inferences about a set of p population parameters $\theta = (\theta_1, \dots, \theta_p)^T$.

For each unit i=1,…,n let $Y_{i,O}$ denote the subset of p variables that are observed, and $Y_{i,M}$ denote the subset that are missing. Thus, for different individuals $Y_{i,O}$ and $Y_{i,M}$ may well be different subsets of the p variables. If no data are missing, $Y_{i,M}$ will be empty.

Next, again for each individual i=1,…,n and variable j=1,…,p, let $R_{i,j} = 1\ if\ Y_{i,j}$ is observed and $R_{i,j} = 0\ if\ Y_{i,j}$ is missing. Let $R_i = (R_{i,1}, \dots, R_{i,p})^T$.. Consistent with the definition of monotone missingness patterns, the pattern is monotone if the p variables can be re-ordered so that for each unit i,

$$R_{i,j} = 0 \Rightarrow R_{i,j'} = 0\ for\ j' = j+1, \dots, p$$

The missing value mechanism is then formally defined as:

$$\Pr\ (R_i \mathrm{I} Y_i)$$

The probability of observing unit i's data given their potentially unseen values Y_i. It is important to note that, in what follows, we assume that unit i's data exist. In other words, if it had been possible for us to be in the right place at the right time, we would have been able to observe the complete data. What the above equation describes therefore, is the probability that the data collection we were able to undertake on unit i yielded values on $Y_{i,0}$. Thus, until we consider sensitivity analysis the missing data are not counter-factual.

Missing Completely At Random (MCAR)

We say that data are Missing Completely At Random (MCAR) if the probability of a value being missing is unrelated to the observed and unobserved data on that unit. Algebraically:

$$\Pr(R_i|Y_i) = \Pr(R_i)$$

Since, when data are MCAR, the chance of the data being missing is unrelated to the values, the observed data are therefore representative of the population. However, relative to the data we intended to collect, information has been lsot.

Missing At Random (MAR)

We say data are Missing At Random (MAR) if given, or conditional on, the observed data the probability distirbution of R_i is independent of the unobserved data. Recalling that for individual i we can partition Y_i as $(Y_{i,O}, Y_{i,M})$ we can express this mathematically as:

$$\Pr(R_i|Y_i) = \Pr\left(R_i\middle|Y_{i,O}\right)$$

This does not mean- as is sometimes supposed- that the probability of observing a variable on an individual is independent of the value of that variable. Under the MAR the chance of observing a variable will depend on its value. Crucially though, given the observed data this dependence is broken.

Missing Not At Random (MNAR)

If the mechanism causing missing data is neither MCAR nor MAR, we say it is Missing Not At Random (MNAR). Under a MNAR mechanism, the probability of an observation being missing depends on the underlying value, and this dependence remains even given the observed data. Mathematically:

$$\Pr(R_i|Y_i) \neq \Pr\left(R_i\middle|Y_{i,O}\right)$$

While in some settings MNAR may be more plausible than MAR, analysis under MNAR is considerably harder. This is because under MAR showed that conditional distirbution of partially observed variables given fully observed variables are the same in units who do, and do not, have the data observed. However, MAR does not hold if MNAR holds.

It follows that influence under MNAR involves an explicit specification of either the selection mechanism, or how conditional distributions of partially observed variables given fully observed variables differ between units who do, and do not, have the data observed.

Formally, we can write the joint distribution of unit i's variables, Y_i, and the indictor for observing those variables, R_i as:

$$\Pr(R_i|Y_i)\, Pr\, \Pr(Y_i) = \Pr(R_i, Y_i) = \Pr(Y_i|R_i)\, \Pr(R_i)$$

In the centre is the joint distribution, and this can be written either as:

- A selection model – the LHS of above equation, i.e., a product of (i) the conditional probability of observing the variables, given their values and (ii) the marginal distribution of the data, OR
- A pattern mixture model – the RHS of above equation, i.e., a product of (i) the probability distribution of the data within each missingness pattern and (ii) the marginal probability of the missingness pattern

This we can specify a MNAR mechanism either by specifying the selection model (which implies the pattern mixture model) or by specifying a pattern mixture model (which implies a selection model). Depending on the context, both approaches may be helpful. Unfortunately, even in apparently simple settings, explicitly calculating the selection implication of a pattern mixture model, or vice versa, can be awkward.

Ignitability

If, under the assumption about the missingness mechanism, we can construct a valid analysis that does not require us to explicitly include the model for that missing value mechanism, we term the mechanism, in the context of this analysis, ignorable.

A common example of this is a likelihood-based analysis is assuming MAR. However, as we see below there are other settings, where we do not assume MAR, that do not require us to explicitly include the model for the missingness mechanism yet still result in valid inference. A complete records regression analysis is valid if data are MNAR dependent only on the covariates.

Using Observed Data to Inform Assumptions About the Missingness Mechanism

The observed data can help frame plausible assumptions about the missingness mechanism- in other words assumptions which are consistent with eh observed data. Exploratory analyses of this nature are important for (i) assessing whether a complete records analysis is likely to be biased and (ii) framing appropriate imputation models. Two key tools for this are summaries of fully observed, or near-fully observed variables by missingness pattern and logistic regression of missingness indicators on observed, or near-fully observed variables.

Implications of missing data mechanisms for regression analyses

Usually, we wish to fit some form of regression model to address our substantive questions. Here, we look at the implications, in terms of bias and loss of information, of missing data in the response and/or covariates under different missingness mechanisms. We first focus on linear regression; our findings there hold for most other regression models, including relative risk regression and survival analysis.

Partially observed response

Suppose we wish to fit the model

$$Y_i = \beta_0 + \beta_1 x_1 + e_i$$

But Y is partially observed. Let R_i indicate whether Y_i is observed. For now assume that the x_i are known without error; for example it may be a design variable. Then the contribution to the likelihood for $\beta = (\beta_0, \beta_1)$ from unit i, conditional on x_i, is

$$L_i = \Pr(R_i, Y_i | x_i) = \Pr(R_i | Y_i, x_i) \Pr(Y_i \mid x_i)$$

Assume, as will typically be the case, that the parameters of $\Pr(Y_i|x_i)$, β, are distinct from the parameters of $\Pr(R_i \mid Y_i, x_i)$.

Provided that Y is MAR given the covariates in the model, units with missing response have no information about β.

The controibution to the likelihood for an individual with missing response is obtained by integrating (for discrete variables summing) over all possible values of the missing response variable $Y_{i,}\ given\ x_i$. This is

$$\int \Pr(Y_i|x_i)\, dY_i = 1$$

Because we are integrating over all possible values of $Y_i\ given\ \beta, x_i$ so the total probability is 1. Conditional on x, all idnviiduals with missing Y thus contribute 1 to the likelihood for β, and so have no effect on, or information about the maximum likelihood estimate of β.

For linear regression there is no information on the regression because the parameter space of the conditional distribution of Y given X is sperate from that of the marginal distribution for X. In other words, the mean and the variance of X have no information on, and place no restriction on, the parameters of the distribution of Y | X. Equivalently, the conditional

distribution of Pr(Y | X) has no information on, and places no restriction on, the marginal distribution of X.

Generalised Estimating Equations (GEEs)

Chapter One

All GEE models consider an estimating equation that is written in two parts. The first part estimates the regression parameters, and the second estimates the association parameters or the parameters of the second order variance distribution.

A short review of generalized linear models

Generalised estimating equations (GEE), is traditionally presented as an extension to the standard array of Generalised Linear Models (GLMs) as initially constructed by Wedderburn and Nelder in the mid-1970s. As such, we provide an overview of GLM and discuss the various ways that GLMs are extended to allow the modelling of correlated data.

GLMs as Likelihood-based Models

GLMs are based on the assumption that individual subjects or observations are independent. This assumption is commonly referred to as the iid requirement, i.e., observations are independent and identically distributed. There are, however, many instances in which responses are correlated.

In the late 1970s, John Nelder designed the first commercial software developed exclusively for GLMs called GLIM.

GLMs and correlated data

Later, Nelder introduced capabilities into GLIM that allowed for the adjustment of the variance-covariance or Hessian matrix so that the effects of extra correlation in the data would be taken into account with respect to standard errors. This was accomplished through estimation of the dispersion statistic.

There are two types of dispersion statistic in GLM modelling. The first is based on the deviance statistic; the second one the Pearson X^2. The overall model deviance and Pearson

X^2 statistics are summary measures of model fit that are traditionally included in model output.

The deviance dispersion is derived by dividing the deviance statistic by the model residual degrees of freedom. Likewise, the Pearson X^2 statistic is calculated by dividing the summary Pearson X^2 by the same model degrees of freedom. The residual degrees of freedom is itself defined as (n-p) where n is the number of cases in the model and p refers to the number of model predictors, including a constant if applicable.

Depending on the type of correlation effect, we characterize response data on counts and binomial trials as under or overdispersed.

The earliest method used to adjust standard errors due to perceived correlation effects was to multiply each parameter standard error by the square root of the Pearson X^2 dispersion statistic. This is a process called the scaling of standard errors. It is also a post-estimation technique that has no effect on the fitted regression coefficients.

GLMs and overdispersed data

A dispersion statistic of greater than 1.0 indicates possible extra correlation in the data. Scaling is an attempt to adjust the standard errors to values that would be observed if the data were not overdispersed. That is, scaling provides standard errors that would be obtained if the dispersion statistic were 1.0.

There are occasions when a model may appear to be overdispersed when in fact it is not. For instance, if the deviance based dispersion of a Poisson model is greater than 1.0, this provides prima facie evidence that the model is overdispersed. In practice, analysts typically start terming a model as overdispersed when the dispersion statistic is above 1.5 and the number of cases in the model is large.

A model may also be what we term apparently overdispersed. Apparent overdispersion occurs when a model omits relevant explanatory predictors, or when the data contain influential and possibly mistakenly coded outliers, or when the model has failed to account for needed interaction terms, or when one or more predictors need to be transformed to another scale, or when the assumed linear relationship between the response and predictors is in fact some other relationship. If any of these are true, then this can lead to an inflation of the dispersion statistic.

The manner in which overdispersion is dealt with in large part depends on the perceived source of overdispersion. Standard methods include; scaling, using robust variance estimators, or implementing models that internally adjust for correlated data.

Scaling Standard Errors

Scaling standard errors is a post hoc method of analyzing correlated data and only adjusts standard errors. The major deficiency is that it does not capture, or appropriately adjust for, an identified cluster or correlation effect. The method simply provides an overall adjustment.

The modified sandwich variance estimator

This method is also post-hoc and only affects standard errors, and not the parameters themselves. The scaling matrix adjusts the Hessian matrix at the next algorithm iteration. Each subsequent iteration in the algorithm updates the parameter estimates, the adjusted Hessian matrix, and a matrix of scales.

The basics of GLMs

Many models now integrated into the GLM framework were previously estimated using maximum likelihood methods. Examples include; logistic, Poisson, and probit regression.

Wedderburn and Nelder discovered that the methods used to estimate weighted linear regression could be adjusted to model many data situations that were previously estimated via maximum likelihood, particularly for those likelihood models based on exponential family distributions. They accomplished this by applying Iterative Weighted Least Squares (IWLS). In addition, they employed a link function which linearized such functions as the logistic, probit, and log. The IWLS was later renamed to IRLS, meaning Iterative Re-weighted Least Squares.

The algorithm takes advantage of forms of variance estimates available from Fisher scoring to develop an easy framework from which computer code can be developed. Later, when computer memory and power become more available, GLM algorithm was extended to include Newton-Raphson based estimation.

Generalised Linear Models are based on the exponential family of distributions, of which include Gaussian or normal, binomial, gamma, inverse Gaussian, Poisson, geometric, and negative binomial. The GEE extension for GLM focused on traditional Gaussian, binomial, gamma and Poisson family members.

All members of the traditional class of generalized linear models are based on one of the above probability functions. The likelihood function is simply a re-parameterization of the probability function or density. A probability function estimates a probability based on given location and scale parameters. A likelihood function, on the other hand, estimates the parameters on the basis of given probabilities or means. The idea is that the likelihood estimates parameters that make the observed data most probable or likely. Statisticians use the log transform of the likelihood, however, because it is (usually) more tractable to use in computer estimation. More detailed justification can be found in Gould, Pitblado, and Poi (2010).

Members of the exponential family of distributions have the unique property that their likelihood formulation may be expressed as

$$exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right\}$$

The expected value of the exponential family distribution is related to the outcome variable of interest. There is a natural connection between these two quantities that allows us to introduce covariates into the model in place of the expected value. This connection is the θ parameter. When a particular distribution is written in exponential family form, the θ parameter is represented by some monotonic differentiable function of the expected value μ. This function links the outcome variable y to the expected value μ. The particular function that results from writing a distributional in exponential form is called the canonical link.

For any member of the exponential family of distributions, there is a general link function, called the canonical link, that relates the linear predictor $\eta = X\beta$ to the expected value of μ. These canonical links occur when $\theta = \eta$. For the Poisson model, we see that $\theta = \ln(\mu)$, implying that the canonical link is given by the log-link $\eta = \ln(\mu)$. Since there is no compelling reason that the systematic components of the model should be linear on the scale of the canonical link, we can, choose any monotonic differentiable function.

Link and variance functions

The inverse link function is what converts the linear predictor $X\hat{\beta}$ into an estimate of the expected value μ. Positive outcomes similarly lead analysts to choose inverse link functions that transform the linear predictor $\eta = X\beta$ to positive values. Some standard choices of link and inverse link functions are listed below with variance functions corresponding to member distributions in the exponential family are listed below that.

Link Name	Link Function $\boldsymbol{\eta = g(\mu)}$	Inverse Link $\boldsymbol{\mu = g^{-1}(\eta)}$
Complementary log-log	$\ln\{-\ln(1-\mu)\}$	$1-\exp\{-\exp(\eta)\}$
Identity	μ	η
Inverse square	$1/\mu^2$	$1/\sqrt{\eta}$
Log	$\ln(\mu)$	$\exp(\eta)$
Log-log	$-\ln\{-\ln(\mu)\}$	$\exp\{-\exp(-\eta)\}$
Logit	$\ln(\frac{\mu}{1-\mu})$	$e^{\eta}/(1+e^{\eta})$
Negative binomial(a)	$\ln(\frac{a\mu}{1+a\mu})$	$1/[a(\exp(-\eta)-1)]$
Probit	$\phi^{-1}(\mu)$	$\phi(\eta)$
Reciprocal	$1/\mu$	$1/\eta$

Standard link and inverse link functions. Binomial data replaces μ with μ/k where k is the number of trials for the particular observation.

Distribution	*Variance V($\boldsymbol{\mu}$)*
Bernoulli	$\mu(1-\mu)$
Binomial(k)	$\mu(1-\frac{\mu}{k})$
Gamma	μ^2
Gaussian	1
Inverse Gaussian	μ^3
Negative Binomial	$\mu + k\mu^2$
Poisson	μ

Variance functions for distributions in the exponential family

Model Construction and Estimating Equations

Independent Data

A common introduction to likelihood-based model construction involves serval standard steps which follow:

- Choose a distribution for the outcome variable
- Write the join distribution for the dataset
- Convert the joint distribution to a likelihood
- Generalize the likelihood via introduction of a linear combination of covariates and associated coefficients
- Parameterize the linear combination of covariates to enforce range restrictions on the mean and variance implied by the distribution
- Write the estimating equation for the solution of unknown parameters

Optimization

Once the model is derived, we may choose to estimate the fully specified log-likelihood with any extra parameters, or we may consider those extra parameters ancillary to the analysis. The former is called full information maximum likelihood (FIML); the latter is called limited information maximum likelihood (LIML). Estimation may then be carried out using an optimization method- most common of which being the Newton-Raphson and iteratively reweighted least squares algorithms (IRLS).

The FIML estimating equation for linear regression

Assuming we have a dataset where the outcome variable of interest is effectively continuous with a large range. In this situation the normal Gaussian distribution is typically used as the foundation for estimation. The density for the normal distribution $N(\mu, \sigma^2) is\ given\ by$:

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Where

$$E(y) = \mu\ \epsilon\ R$$

$$V(y) = \sigma^2 > 0$$

And R indicates the range of real numbers. The density for a single outcome is then

$$f(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{(y_i-\mu)^2}{2\sigma^2}\right\}$$

The joint density for n independent outcomes subscripted from 1,…, n is the product of the densities for the individual outcomes

$$\begin{aligned} f(y_1, \dots, y_n|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{(y_i-\mu)^2}{2\sigma^2}\right\} \\ &= \prod_{i=1}^{n} exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

The likelihood is simply a restatement of the join density where we consider the outcomes as given, and model the parameters as unknown.

For More Information:

This package of information regarding statistics and methodology is not a wholly original piece. It is better regarded as an anthology of various books that I have taken from and stuck together to make a comprehensive guide to statistical and advanced statistical theory.

The basic statistical theory is all gathered from Agresti's intro to statistical theory.

Missing data is taken from Carpenter and Kenward's Multiple Imputation and its application.

Count Models is taken from Hilbe's Modelling Count Data

Fixed Effects is taken from Allison's Fixed Effects Regression Models.

Generalised Estimating Equations is taken from Allison's of the same name.

The bulk of the rest is taken from lecture notes from the MSc Sociology advanced quantitative statistics course at the University of Oxford.