**INTRODUCTION TO GENERALIZED ESTIMATING EQUATIONS**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

**Pre-requisites**

 This document assumes you have developed at least some familiarity with regression analysis. Admittedly, researchers have developed many variants of regression analysis. The following table outlines some of the most common variants.

|  |  |
| --- | --- |
| Variant | Circumstances in which this variant is most suitable |
| Linear regression, often called multiple regression or ordinary least squares | The outcome measures are numerical—and the predictors and outcome measures are likely to be linearly related |
| Logistic regression  | The outcome measures is binary, comprising two categories, such as healthy and unhealthy |
| Ordinal regression  | The outcome measure is ordinal—such as a ranking from 1 to 50 |
| Poisson regression  | The outcome measure is a count, such as number of awards |

If you are also, at least slightly, familiar with multi-level modelling—also called hierarchical linear modelling or random effects models—this document will be even easier to understand. Multi-level models tend to be suitable whenever

* you measure the same people, animals, specimens, or some other unit more than once over time—and hence the design is longitudinal
* the populations from which you recruited these units, such as people, comprise several distinct clusters—but clusters that you are not interested in comparing

**Role of** **generalize estimating equations or GEE**

You might assume that you could utilize one of these regression models or multi-level models to analyse almost any data. But, this assumption is actually incorrect. In particular,

* if your data are longitudinal—that is, you want to measure the same people, animals, specimens, or some other unit more than once over time or the population comprises several categories yet…
* your outcome measures are not numerical but are binary, ordinal, or something else, you need to consider another approach instead

This document outlines an approach that can be utilized in all these circumstances: generalized estimating equations. Indeed, this approach supersedes many other techniques. Once you learn this approach, you may not need to conduct linear regression, logistic regression, ordinal regression, Poisson regression, multi-level modelling, or quite a few other techniques.

Perhaps, you feel slightly aggrieved now. You feel you wasted your time learning other techniques. But do not be too concerned. People who have learned linear regression, logistic regression, or multi-level modelling tend to understand and conduct generalized estimating equations more effectively than people who have not learned any of these techniques.

|  |
| --- |
| **Example with numerical measures** |

**The study**

 To learn about generalized estimating equations, you should observe an example first. Suppose you want to assess whether sleep quality changes across the seasons. Furthermore, you want to ascertain whether an intervention, in which people repeat the word “worried” 40 times before they retire to bed, improves sleep. The premise is that specific words, when repeated many times rapidly, gradually seem funny rather than distressing. For this study

* the sleep quality of 40 participants is gauged four times, once during each season of the year
* to gauge sleep quality, participants simply rate how refreshed they feel when they awoke, on a scale from 1 to 10.
* half the participants are instructed to repeat the word “worried” 40 times rapidly every night
* the other participants do not receive this instruction—and are, hence, the control group

**Enter the data**

First, you need to enter the data. Usually, when you enter data into a spreadsheet, such as SPSS, each row corresponds to one unit, such as one person, one animal, one specimen, and so forth. However, when you conduct generalized estimating equations, you do not utilize this approach. Instead, you utilize an approach called longitudinal data entry—the same approach that is used when researchers conduct multi-level modelling. In particular

* each row represents a specific unit, such as a particular person, and a specific time
* for example, in this instance, participants indicated their sleep quality four times
* therefore, four rows represent each person
* the first column, called ID, differentiates the participants, as the following figure shows



**Analyse the data**

To conduct the analysis, in the *Analyze* menu, choose *Generalized Linear Models* and then *Generalized estimating equations*, to generate the following screen. The term *generalized* implies the technique generalizes to many kinds of measures, rather than only numerical data, for example.



* In the box called *Subjects*, specify *ID*—or whatever you called the column that differentiates the participants—by clicking the relevant arrow
* In the box called *Within-subject variables*, specify the variable that differentiates the time or date the data was collected; in this instance, the variable is *Season*
* In the section called *Working Correlation Matrix*, press the downward arrow to the right of *Structure* and choose *Unstructured*
* On the top, press *Type of Model* to proceed to the next tab



* In this example, the outcome measure, sleep quality, can be regarded as a numerical response
* In these instances, choose *Linear* rather than other options such as ordinal logistic, ordinal probit, Poisson loglinear, and so forth
* On the top, press *Response* to proceed to the next tab



* The dependent variable is simply the outcome measure—in this instance, sleep\_quality
* On the top, press *Predictors* to proceed to the next tab



* In the box called *Factors*, include the categorical predictors such as intervention versus control
* In the box called *Covariates*, include the numerical predictors. If you assume that *season* is linearly related to outcome measure, you would include this variable in this box instead.
* On the top, press *Model* to proceed to the next tab



* In the box called *Model*, you could include the condition—intervention versus control—season, and the interaction between condition and season. A significant interaction would indicate the effect of this intervention varies across the seasons
* To transfer these variables into this box, highlight the variables on the left side, press the downward arrow below *Build Terms*, choose *Factorial*, and then press the arrow under this term
* Press OK.

**Interpret the data**

The following figure presents an extract of the data. In essence, as the table called *Tests of Model Effects*, indicates

* the condition is significant, indicating that sleep quality differed between participants who received the intervention and participants who did not receive the intervention
* neither season nor the interaction are significant.

|  |
| --- |
|  |

 As the table called *Parameter Estimates* reveals

* The B value associated with intervention condition 0 is negative.
* This finding implies that participants in intervention condition 0 generated lower scores on sleep quality than participants in intervention condition 1.
* If you defined the control condition as 0 and the intervention condition as 1, this finding suggests the intervention condition was effective
* If the effect of season had been significant, the researchers would then interpret the B values in this table
* For example, they might note the B value associated with Season 3 is positive—indicating that sleep quality during Season 3 is higher than sleep quality during Season 4, the reference condition. If you do not understand this comparison, you might need to learn more about dummy variables in linear regression.

|  |
| --- |
| **Rationale** |

 The previous example merely demonstrated the procedure, but did not explain or justify the choices. This section offers some insight into the underlying rationale and decisions.

**Covariance or correlation matrix**

In the previous example, the researcher had to choose the *Working Correlation Matrix*. Several options are available, such as unstructured, independent, AR(1), exchangeable, and M-dependent. In this example, these correlations refer to the correlations of the outcome measure—in this instance, sleep quality—between consecutive seasons. To illustrate, consider the extract of data below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Season 1 | Season 2 | Season 3 |  Season 4 |
| Adam | 3 | 5 | 4 | 5 |
| Betty | 7 | 9 | 8 | 9 |
| Chris | 4 | 6 | 5 | 7 |
| Donna | 1 | 3 | 2 | 2 |
| Eve | 6 | 4 | 5 | 6 |
| Fred | 7 | 4 | 7 | 5 |
| Georgia | 8 | 6 | 7 | 9 |

The columns are obviously, and unsurprisingly, correlated with each other. For example, the people who do not sleep as well in Season 1—such as Adam and Donna—also do not sleep as well in Season 2. The people who sleep well in Season 1—including Betty and Georgia—also sleep well in Season 2. Indeed, the following table estimates the correlation between each pair of seasons.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Season | Season 1 | Season 2 | Season 3 |  Season 4 |
| Season 1 | 1 |   |   |   |
| Season 2 | 0.5 | 1 |   |   |
| Season 3 | 0.24 | 0.46 | 1 |   |
| Season 4 | 0.16 | 0.32 | 0.38 | 1 |

 If you choose *unstructured*, the statistical package assumes all the correlations or covariances are different. This assumption is probably accurate. However, to apply this model, the statistical package needs to estimate each of these correlations or covariances—and the need to estimate many parameters diminishes power and thus reduces the likelihood of a significant result.

**Exchangeable**

 Instead, you could choose other alternatives instead. For example, you could instruct the statistical package to assume that all the correlations between two different times or seasons are the same—as the following table illustrates. This alternative is called *exchangeable­* and is equivalent to *compound symmetry* in multi-level modelling*.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Season | Season 1 | Season 2 | Season 3 |  Season 4 |
| Season 1 | 1 |   |   |   |
| Season 2 | 0.19 | 1 |   |   |
| Season 3 | 0.19 | 0.19 | 1 |   |
| Season 4 | 0.19 | 0.19 | 0.19 | 1 |

To apply this model, the statistical package needs to estimate only one correlation or covariance, increasing power. Unfortunately, in practice, this assumption is not especially plausible. Typically, the correlation between times or seasons closer in time should be higher than times or seasons farther in time.

**Auto-regression**

Consequently, you might want to choose more realistic assumptions. One example is called Autoregressive 1. This alternative assumes that two consecutive times or seasons, such as Seasons 1 and 2, might be correlated to a certain degree—a degree we refer to as r. Two times or seasons that are two intervals apart, such as Seasons 1 and 3, would be correlated r2. Two times or seasons that are three intervals apart, such as Seasons 1 and 4, would be correlated r3, and so forth. The following table illustrates this pattern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Season | Season 1 | Season 2 | Season 3 |  Season 4 |
| Season 1 | 1 |   |   |   |
| Season 2 | 0.20 | 1 |   |   |
| Season 3 | 0.04 | 0.20 | 1 |   |
| Season 4 | 0.008 | 0.04 | 0.20 | 1 |

 To apply this model, again the statistical package needs to estimate only one correlation or covariance, increasing power. Yet, this pattern is slightly more realistic than is exchangeable, but still not entirely plausible.

**M-dependent**

 One compromise is an alternative called M-dependent. This alternative assumes the correlation between times or seasons depends only on the number of intervals apart, as the following table shows. Furthermore, once the number of intervals exceeds some number—a number you specify—the correlation becomes 0. To apply this model, the statistical package needs to estimate a few correlations—equal to the number of times minus one. However, this pattern is fairly plausible.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Season | Season 1 | Season 2 | Season 3 |  Season 4 |
| Season 1 | 1 |   |   |   |
| Season 2 | 0.20 | 1 |   |   |
| Season 3 | 0.16 | 0.20 | 1 |   |
| Season 4 | 0 | 0.16 | 0.20 | 1 |

**Which alternative should you choose?**

So, how can you decide which alternative is preferable. You can apply three approaches. First, you can attempt many of these alternatives, and merely choose the option that generates the lowest Quasi Likelihood under Independence Model Criterion or QIC—a statistic that appears in the output and indicates the accuracy of models.

Second, you can decide which alternative seems most likely from the design. To illustrate

* if the various times are days apart, and spaced evenly, auto-regression or M-dependent are the most plausible
* if the various times are close together—or people are likely to shift quite erratically over time—an unstructured covariance matrix might be necessary.

Finally, if all participants provide data at every time, called a balanced design, the software will actually determine the most suitable alternative. The software will begin with the choice you select, but then try other alternatives, and chose the most effective—one of the best features of generalized estimating equations. This choice, however, is most effective only when the design is balanced.

**Underlying process**

 So how does the software procedure this output? What is the underlying rationale? In essence, the software begins with a general formula. The formula includes parameters, such as B values. Then, the software

* substitutes the B values and other parameters with somewhat random numbers
* the software then determines whether this formula predicts the data well
* if not, the software updates the B values and other parameters with other numbers
* the software continues this process until the formula predicts the data well, called iteration
* in practice, this iteration is not quite as random as this description implies.

Further information is probably harder to explain. But, if you are interested, here are the formulas that underpin generalizing estimating equations:





In this formula

* yi are the outcome measures—such as sleep quality at each time—for person i
* i is the mean outcome measure for person i, such as the mean sleep quality
* Vi is the covariance matrix between the various times, such as an unstructured matrix
* Di is called the matrix of derivatives δμi/δβj and, in essence, represents the extent to which changes in sleep quality are related to changes in the B coefficients
* Ri is the correlation between the various times
* Ai is the variance of the outcome measure—such as the variability of sleep quality over time—for person i
* φ is called an over-dispersion parameter

The over-dispersion parameter equals



In this formula

* N is the number of variables
* p is the number of parameters in the model

|  |
| --- |
| **Variations** |

 Thus far, this document has presented only one example. However, generalizing estimating equations can be modified to accommodate many circumstances. This section discusses some of the options associated with each tab.

**Type of model**

You can choose a variety of options, depending on the outcome measures. The following table clarifies these options:

|  |  |
| --- | --- |
| Variant | Circumstances in which this variant is most suitable |
| Linear | The outcome measures are numerical—and the predictors and outcome measures are likely to be linearly related |
| Gamma with log link | The outcome measure comprises only values that exceed 0 and is skewed towards larger values |
| Binary logistic or binary probit | The outcome measures are binary, comprising two categories, such as healthy and unhealthy |
| Ordinal logistic or ordinal probit | The outcome measure is ordinal—such as a ranking from 1 to 50 |
| Poisson loglinear  | The outcome measure is a count, such as number of awards |

**Statistics**

* If your outcome measures are binary, comprising two categories, in the Statistics tab, tick “include exponential parameter estimates. These estimates are the odds ratios. For researchers who understand logistic regression, these odds ratios are useful

**Clusters**

Sometimes, the populations from which you recruited your participants or specimens comprise several distinct clusters—but clusters that you are not interested in comparing. In these circumstances, implement the same process except

* ID now differentiates the clusters instead of individual participants
* Otherwise, proceed as you would otherwise

|  |
| --- |
| **Limitations** |

 Generalized estimating equations are accurate only when the sample size is large.