**INTRODUCTION TO GENERALIZED LINEAR MODELS**

**by Simon Moss**

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

**Pre-requisites**

This document assumes you are familiar with linear regression, sometimes called multiple regression or ordinary least squares regression. If you are familiar with logistic regression as well, this document will be especially easy to understand.

**Applicability of generalized linear models**

When researchers want to ascertain whether a set of predictors is related to some outcome, linear regression is often, but not invariably, applicable. The following table specifies the circumstances in which linear regression is not applicable.

|  |  |
| --- | --- |
| Circumstance in which linear regression is unsuitable | Illustration or clarification |
| The outcome is binary and, thus, comprises two categories | * The researcher wants to assess whether the grade point average, highest degree, and age of research candidates affects the likelihood they will complete their thesis on time. * In this instance, the outcome measure—whether the individual will complete the thesis on time—is binary |
| The outcome is ordinal, such as a ranking | * The researcher wants to assess whether the grade point average, highest degree, and age of research candidates affects their ranking in a recent competition. * In this instance, the outcome—a ranking from 1 to 100, for example—is called ordinal * Two people ranked 50 and 51 might be more similar than two people ranked 99 and 100. So, the properties of these ranks differ from the properties of typical numbers |
| The outcome is a count | * The researcher wants to assess whether the grade point average, highest degree, and age affects the number of short courses each candidate has completed * Many researchers utilize typical linear regression in these instances. But, other models tend to be more effective |
| The residuals are unlikely to be normally distributed |  |

Instead, in these circumstances, generalized linear models are suitable. These techniques generalize the traditional linear models to other outcome measures—measures that are not numerical, for example. Generalized linear models are similar to traditional linear models except

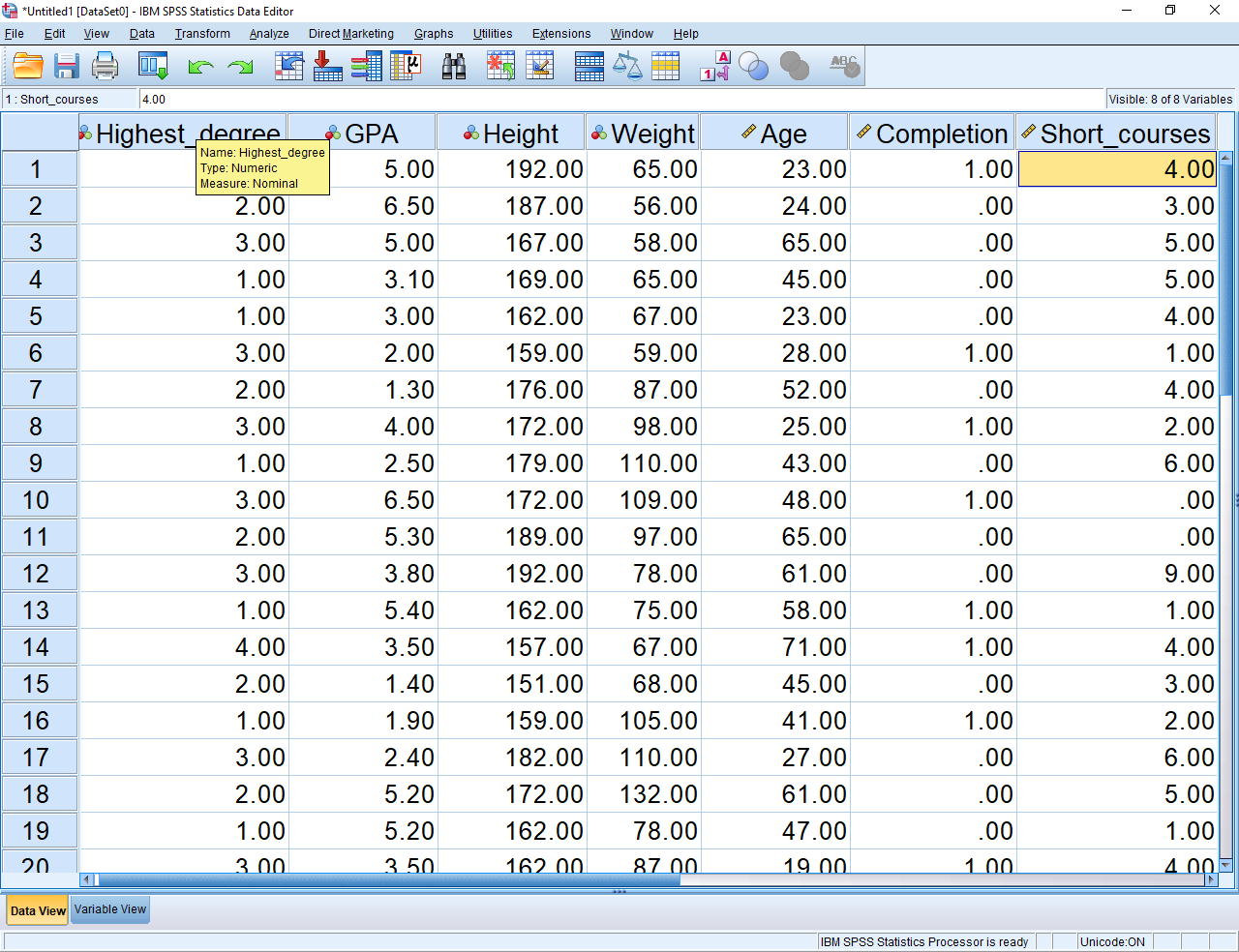
* the error term, in general, is not assumed to be normally distributed—but instead conforms to other distributions or shapes
* some transformation, such as a log, tends to convert the initial set of predicted outcomes to a final set of predicted outcomes
* various iterative procedures are applied to estimate the B coefficients.

|  |
| --- |
| **Overview** |

Most of this document will demonstrate generalized linear models in one software program: SPSS. Even if you do not use SPSS, read these sections to learn about the technique. The end of this document then clarifies how to conduct generalized linear models in R.

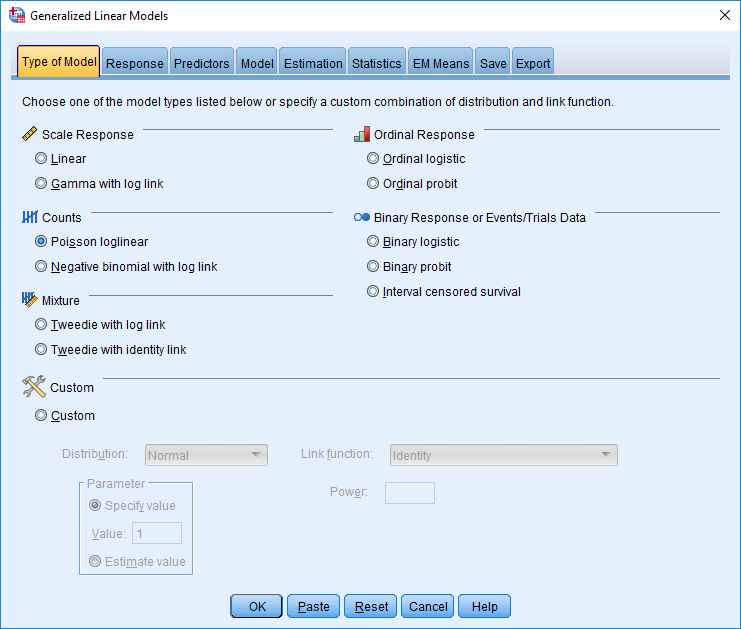
**Illustration**

To learn about generalized linear models, suppose a researcher wants to ascertain which characteristics and experiences foster success and satisfaction in research candidates. The predictors include highest degree, grade point average, height, weight, and age, for example. The measures of success and satisfaction include time needed to complete the thesis, number of short courses completed, and level of enjoyment. The following screen demonstrates how you would enter the data—in which each row corresponds to one candidate.



**Open the dialogue box**

To conduct the analysis in SPSS, in the *Analyze* menu, choose *Generalized Linear Models* and then *Generalized linear models* again, to generate the following screen. At the top of this screen is a series of tabs, such as model type, response, predictors, model, estimation, and statistics. To conduct the analysis, you need to select each tab in turn and then choose the relevant options.



**Type of Model**

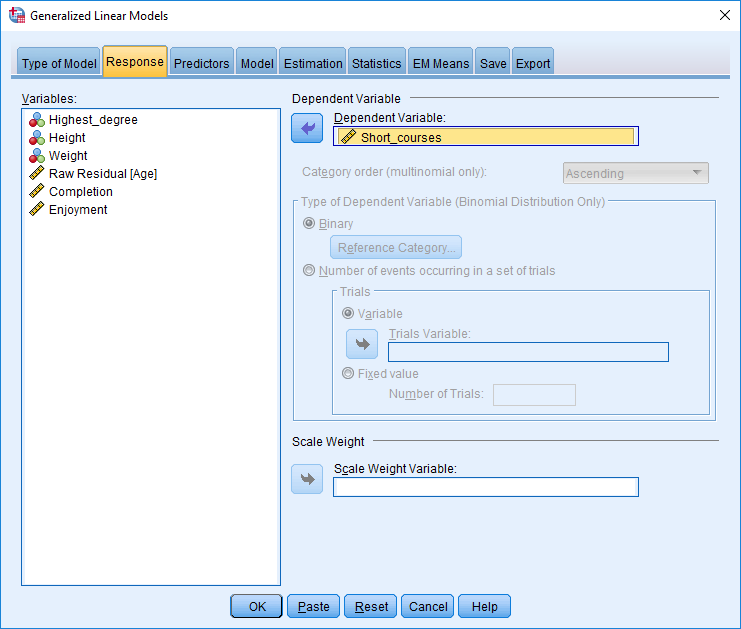
After you click the type of model tab, you need to indicate which family of models you want to examine. Although this document will differentiate these varieties in more details later, the following table offers some guidelines to help you reach this decision.

|  |  |
| --- | --- |
| Characteristics of your outcome measure | Choices |
| The outcome is **numerical**, such as duration before you completed your thesis | * Linear * Gamma with log link |
| The outcome measure is a **count**—such as the number of short courses a candidate has completed | * Poisson loglinear * Negative binomial with log link |
| The outcome measure is **ordinal**, such as a ranking in a competition | * Ordinal logistic * Ordinal probit |
| The outcome measure is **binary** and, thus, comprises two categories—such as whether the individual completed the thesis | * Binary logistic * Binary probit * Interval censored survival model |

**Response**

After you choose the “Response” tab, specify the outcome measure, such as number of short courses. Typically, you merely need to transfer the outcome measure, also called the dependent variable, into the box labelled dependent variable, as shown below. The default settings are usually adequate. However

* when the outcome measure is binary or ordinal, you can choose other options. Particular options might facilitate the interpretations but do not change the model fundamentally.
* one exception revolves around outcome measures that are counts.
* in these circumstances, you can also specify a variable or column that indicates the maximum possible count for this individual, often representing the the total number of trials or attempts on some task.



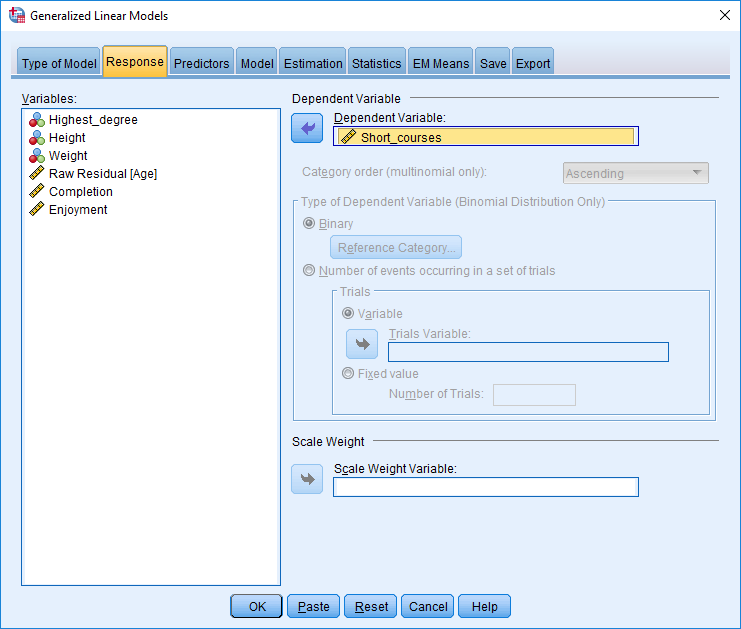
**Predictors**

After you choose the “Predictors” tab, specify all thepredictors you might include in your model—the variables that could be related to the outcome measure. Specifically

* in the box called factors, specify the categorical predictors, such as gender
* in the box called covariates, specify the numeric predictors, such as height. These predictors must be designated as “scale” in Variable view.

**Model**

After you choose the “Models” tab, you specify how these predictors interact to affect the outcome measure. This tab generates the following screen.



To illustrate

* the model could be merely the sum of these predictors, weighted by some coefficient, such as b1 x age + b2 x height. To specify this model, you would highlight these predictors and then choose “main effects” under “Type” before pressing the right arrow.
* the model could include interactions between these predictors, such as b1 x age x height. To specify this model, you might choose “factorial” instead of “main effect”
* the model could even include other relationships between predictors, such as nesting—a concept that is discussed in another document called multi-level modelling

**The other tabs**

The “Estimation tab” enables you to change the techniques that are applied to estimate the B coefficients. Typically, the defaults are sufficient. Indeed, for the other tabs as well, the defaults tend to be suitable as well.

|  |
| --- |
| **Specifying model types** |

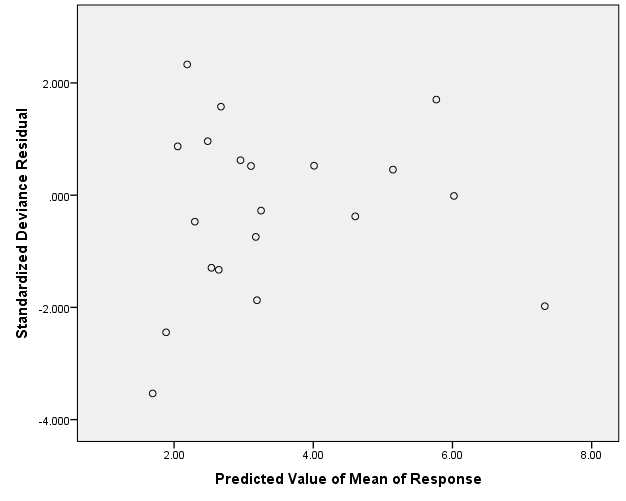
The previous section included some guidelines on how to decide which model types to choose, such as negative binomial or Poisson loglinear. For most classes of outcomes, however, you can select one of several alternatives. This section clarifies how you can differentiate these alternatives. In particular, the following table outlines the guidelines you can apply to differentiate these alternatives.

|  |  |  |
| --- | --- | --- |
| Outcome measure | Choices | Guidelines to decide  which options to choose |
| Numerical | * Linear * Gamma with log link | * When the data generates a positive skew—that is, a small number of very high values—“Gamma with the log-link” may be preferable * The gamma distribution corresponds to a variety of shapes and thus can be useful in many circumstances, such as financial data. * However, whenever a gamma distribution is applied, some of the diagnostic indices, such as Cook’s distance, might be harder to interpret. |
| Count | * Poisson loglinear * Negative binomial with log link | * The Poisson loglinear model is simpler and, thus more powerful if the sample size is limited * Yet, the Poisson loglinear model assumes the variance of some outcome is equal to the mean * For example, suppose that candidates who are 25 years old tend to have completed 5 short courses in their life. The model will assume the variance in people 25 years old is also 5. * In contrast, the negative binomial alternative is more flexible than is the Poisson loglinear model and, therefore, can accommodate deviations from this assumption |
| Ordinal | * Ordinal logistic * Ordinal probit | * These two alternatives generate similar results * The parameters of ordinal logistic might be slightly easier to interpret |
| Binary | * Binary logistic * Binary probit * Interval censored survival model | * In practice, these alternatives tend to generate similar results * Health researchers tend to prefer the binary logistic model, sometimes called logistic regression, because the coefficients can be translated to odds ratios. These odds ratios are simple to interpret—see the document on logistic regression. * In contrast, economists and political scientists tend to prefer the probit model, because this alternative tends to be more resistant to variable error variances or heteroscedasticity |

**Poisson loglinear model versus negative binomial with log link**

As the previous table indicated, whenever the variability of an outcome measure deviates from the mean, the negative binomial with log link is preferable to the Poisson loglinear model. But, how can researchers test this assumption? Although researchers have suggested a variety of procedures, perhaps the simplest approach is to simply conduct both alternatives: the Poisson model and the negative binomial model. And, when you complete these models

* select the “Save” tab
* tick the options “Predicted value of mean of response” and “Standardized deviance residual”
* execute the analysis
* construct a scatterplot—as illustrated in the following figure—in which the Predicted value of mean of response corresponds to the X axis, and the Standardized Deviance residual corresponds to the Y axis.
* to construct this scatterplot, you could select the “Graphs” menu, “Legacy Dialogs”, and then “Scatter/Dot”



* after you construct these graphs, assess the extent to which the points vary along the Y axis. For example, if most of the points are within the interval -2 to +2, the model seems to represent the data well
* indeed, if a sizeable portion of the points are outside the interval -3 to +3, the model does not at all represent the data well.
* choose the model that seems to represent the data better

**Tweedie models**

Sometimes, the outcome measure comprises only positive numbers but many zeros. For example, in response to the question, what is the approximate time you have spent in Antarctica, none of the answers are negative values, although many answers, if not the majority, are zero. In these circumstances, researchers often utilize a model called Tweedie. They may apply both the Tweedie log-link and Tweedie identity models and then decide which model generates better results.

**The custom option**

Finally, in SPSS, R, and many statistical packages, researchers can customize or design the model themselves. That is, they can choose a distribution function and a link function. But

* what are distribution functions?
* And what are link functions?

Distribution functions explain the diverse patterns of variability in the outcome measure and primarily depend on whether the outcome is binary, categorical, ranks, counts, or numbers. The following table specifies the circumstances in which the various distribution functions are likely to be applicable

|  |  |
| --- | --- |
| Distribution function | Circumstances in which the distribution function is applicable |
| Binomial | * The outcome is binary: Every person or unit is assigned one of two values |
| Gamma | * The outcome includes only values above one. The distribution also includes a few very high values, called a positive skew |
| Inverse Gaussian | * Used in similar circumstances to Gamma distributions |
| Negative binomial | * The outcome is often the number of trials needed to generate a specific number of events—such as the number of attempts before a basketballer can sink 10 baskets |
| Normal | * The outcome is a numerical variable; the error term conforms to a symmetric distribution that resembles a bell |
| Poisson | * The outcome is often the number of times some event transpired within a fixed period of time |
| Tweedie | * The outcome comprises a mixture of Poisson and gamma distributions. Often, many of the cases equal 0. The other cases exceed zero. |
| Multinomial | * The outcome is a rank or ordinal scale |

The link function then transforms the predicted outcome measure—and may affect the range of this outcome measure. To illustrate, consider a subset of link functions in the following table.

|  |  |
| --- | --- |
| Link function | Circumstances in which this link function is applicable |
| Identity | * The outcome measure is not transformed |
| Log | * The new function of y equals log(y) |
| Logit | * The new function of y equals log(y/1-y), suitable only when the distribution is binomial |

|  |
| --- |
| **Conducting generalized linear models in R** |

If you have conducted linear models or multiple regression in R, generalized linear modelling is straightforward. An illustration of typical code as well as an explanation appears in the following table.

|  |  |
| --- | --- |
| Typical code | Explanation |
| model <- glm(outcome ~ predictor1 + predictor2 + predictor3, data=datafile, family = binomial()) | * “model” is merely the name assigned to this model and could be any label * “outcome” is merely the name assigned to the outcome measure * “predictors 1 to 3” are merely the names assigned to the predictors * “datafile” is the name of your data file * the family refers to the distribution function and can be binomial, gaussian, Gamma, inverse.gaussian, poisson, and a variety of other functions. Note that gaussian is synonymous with a normal distribution. |

In the brackets after the distribution is specified, you can change the link function to logit, identity, inverse, log, or a variety of other links. An example is shown in the following box. If the brackets are empty, R uses a default link function—a link function that is often appropriate.

|  |
| --- |
| model <- glm(Outcome ~ Predictor1 + Predictor2 + Predictor3, data=datafile, family = binomial(log)) |