Journal Title
XX(X):1–5

Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Dominance analysis for count dependent variables

Joseph N. Luchman[1]

Abstract

Determining independent variable relative importance is a highly useful practice in organizational science. Whereas techniques to determine independent variable importance are available for normally distributed and binary dependent variable models, such techniques have not been extended to count dependent variables (CDVs). The current work extends previous research on binary and multi-category dependent variable relative importance analysis to provide a methodology for conducting relative importance analysis on CDV models using dominance analysis (DA). Moreover, the current work provides a set of comprehensive data analytic examples that demonstrate how and when to use CDV models in a DA and the advantages general DA statistics offer in interpreting CDV model results. Moreover, the current work outlines best practices for determining independent variable relative importance for CDVs using replaceable examples on data from the publicly available National Longitudinal Survey of Youth 1979 cohort. The present work then contributes to the literature by using in-depth data analytic examples to outline best practices in conducting relative importance analysis for CDV models and by highlighting unique information DA results provide about CDV models.

Keywords

Dominance Analysis, Relative Importance, Poisson Regression, Negative Binomial Regression, R-square

Introduction

Organizational scientists conduct research on work-related problems that focus on many different specific topics including job performance, employee wellness, and effective task staffing. Quantifying topics such as job performance often requires that researchers use data that are in the form of discrete, sometimes infrequent, events such

[1]Fors Marsh Group LLC

as number of contracts won in a year, number of complaints received in a month, or number of days absent for illness in a business quarter. Discrete, infrequent event data, called *count dependent variables*/(*CDVs*) in this paper, are useful representations of many concepts in organizational science but can present additional complications for analysis. One such complication is that count data can diverge from the statistical assumptions made of the Normal or Gaussian distributed linear regression model—by far one of the most common predictive models applied in organizational science ().

CDVs are often modeled using generalized linear models adapted to the structure of infrequent events. Poisson or negative Binomial regressions () are commonly applied to CDVs as they tend to fit better with positive integer-valued data than do Normal/Gaussian distributions. Although models such as the Poisson regression fit better to CDVs, Poisson and similar regressions are more complex to interpret than linear regression as they are intrinsically non-linear. In addition, CDVs' discrete, event-oriented nature requires additional considerations that tend not to apply to continuous, Normally-distributed data.

Decision makers in industry, government, and non-profit organizations look to organizational scientists to estimate and interpret CDV models when required given the research question. In linear regression models, a commonly used tool to assist in the interpretation of statistical modeling is to evaluate and compare the relative importance of the independent variables (). Comparing independent variables and determining their importance relative to one another is most often accomplished using the dominance analysis/(DA) () approach. Published methodological work on DA has discussed multiple intrinsically non-linear models including binary (), ordered, and multinomial logit () models but has not provided an extensive discussion of how to implement and interpret DA with CDVs. Moreover, CDVs can require adjustments to modeling such as the consideration of *exposure* that can affect a DA's relative importance determination results.

The goal of this work is to review both DA and CDV models and then offer recommended practice for applying DA to CDV models. This paper is organized into ... sections. The first section reviews DA The se

Dominance Analysis

DA is an extension of Shapley value decomposition from Cooperative Game Theory () which seeks to find a solution to the problem of how to subdivide payoffs to players in a cooperative game based on their relative contributions.

The Shapley value decomposition method views the predictive model as a cooperative game where the different independent variables work together to predict the dependent variable. The payoff from the predictive model is the value of the model fit statistic; usually this payoff is an R^2.

This methodology can be applied to predictive modeling in a conceptually straightforward way. Predictive models are, in a sense, a game in which independent variables cooperate to produce a payoff in the form of predicting the dependent variable. The component of the decomposition/the proportion of the payoff ascribed to each independent variables can then be interpreted as the IVs importance in the context of the model as that is the contribution it makes to predicting the dependent variable.

In application, DA determines the relative importance of IVs in a predictive model based on each IV's contribution to an overall model fit statistic—a value that describes the entire model's predictions on a dataset at once. DA's goal extends beyond just the decomposition of the focal model fit statistic. In fact, DA produces three different results that it uses to compare the contribution each IV makes in the predictive model against the contributions attributed to each other IV. The use of these three results to compare IVs is the reason DA is an extension of Shapley value decomposition.

Complete dominance between two IVs is designated by:

$$X_v D X_z \; if \; 2^{p-2} = \Sigma 2 \tag{1}$$

Where X_v and X_z are two IVs, S_j is a distinct set of the other IVs in the model not including X_v and X_z which can include the null set (...) with no other IVs, and F is a model fit statistic. Conceptually, this computation implies that when all 2^{p-2} comparisons show that X_v is greater than X_z, then X_v completely dominates X_z.

Conditional dominance statistics are computed as:

$$C^i_{X_v} = \tag{2}$$

Where S_i is a subset of IVs not including X_v and [p-1i-1] is the number of distinct combinations produced choosing the number of elements in the bottom value (i-1) given the number of elements in the top value (p-1; i.e., the value produced by choose(p-1, i-1)).

In effect, the formula above amounts to an average of the differences between each model containing X_v from the comparable model not containing it by the number of IVs in the model total.

General dominance is computed as:

$$C_{X_v} = \frac{\sum_p^i C^i_{X_v}}{p} \tag{3}$$

Where, $C^i_{X_v}$ are the conditional dominance statistics for X_v with i IVs. Hence, the general dominance statistics are the arithmetic average of all the conditional dominance statistics for an IV.

In the section below, I transition to discussing some of the nuances of CDVs for the application of DA.

Applying Dominance Analysis to Count Dependent Variable Models

CDV models are complex, inherently multiplicative (note that it is possible to estimate them as non-multiplicative - but this is rarely done) models that require the use of estimation techniques such as maximum likelihood to obtain parameter estimates and sampling variances. It is helpful to discuss some of the nuances of how these models are estimated to understanding the implications of these complexities for DA. Hence, in the sections to come, I discuss aspects

Poisson Regression: The Most Basic Case

The most conceptually simple CDV model is the Poisson regression. The Poisson regression's log likelihood is a sum of three components as is shown below in (can I reference this equation?).

$$\ln L = \sum_{i=1}^{N} xb_i y_j - (\ln(y_i!) + e^{xb_i}) \tag{4}$$

Where xb_i refers to a respondent's untransformed/linear predicted value from the Poisson regression. Of note with this log-likelihood is the division into two separate components. On the one hand, the log-likelihood increases, with the $xb_i y_j$ term. Hence, the product of the observed CDV value and the predicted value for a respondent contributes directly to the log-likelihood and indicates better fit to the data. On the other hand, the second set of terms, $\ln(y_i!) + e^{xb_i}$ decrease the log-likelihood. Thus, has the log factorial of the CDV increases and as the exponential function of the predicted values increase without a concomitant increase in the $xb_i y_j$.

For example consider a respondent with a CDV value of 5. The best fit to this value would be a transformed predicted value of 5 or an untransformed/linear predicted value of $e^5 = -1.740$. When applied to the log-likelihood function, the value obtained is $5 * -1.740 - (\ln(5) + e^{-1.740} = -1.740$ or untransformed/linear predicted value. Given the Poisson log-likelihood, the best possible fit to this observaton would produce a log-likelihood value of The balance between these two quantities across all respondents results in the final set of estimates.

...negative binomial...

$$\ln L = \sum_{i=1}^{N} xb_i y_j - (\ln(y_i!) + e^{xb_i}) \tag{5}$$

In applying DA to CDVs, many readers might question whether there is a need to use methods other than the standard linear regression with the explained variance R^2 metric.

Although CDV models share a number of similarities with continuous DVs, CDVs and the models designed to work with them are a distinct subset of generalized linear models and behave differently than linear regression in ways that *(cite Blevins and summarize somewhere around here)*.

One similarity across count and continuous DVs is that, as the mean of a CDV variable increases, it grows increasingly good at being approximated with a Normal distribution *(cite!)*. This points to an important difference when applied to CDVs with rare events in that the distributions for count and continuous DVs diverge notably when the mean is nearer 0 (example here?). Such divergence results in important differences to model-to-data fit that not only affects the model fit metric's value overall but also how specific IVs explain variation in the CDV. In particular, CDVs are discrete, non-negative integers and, accordingly, CDV models are discrete probability distributions that accommodate non-negative integers. Normal distributions are continuous

Applying DA to CDVs

A useful first step toward defining recommended practice for applying DA to CDVs is to understand how CDV models differ from linear models. Differences between CDV and linear models affect model estimation and how relative importance among IVs should be determined. One substantial difference between CDV models such as the Poisson from linear models is in the functional form of the model. In a linear model, the magnitude of the regression coefficient for an IV reflects the expected change in the DV given one unit of change in the IV. By contrast, CDV models most often use a log-linear linking function. In a log-linear link model like the Poisson, the coefficients are estimated using from the data as though they were transformed using a natural logarithm. This implied transformation, or linking function, results in the CDV effectively ranging over all real numbers just like a continuous DV is expected to in linear regression.

$$ll = \frac{1}{1} \tag{6}$$

Although the CDV is implied to range over all real numbers in the estimation algorithm, the observed CDV is not changed and the predicted values from the CDV model are in log-linear units as opposed to those of the CDV (i.e., counts of the event). In order to produce meaningful predicted values, CDV models need to back-translate their predicted values to the metric of the CDV. The back-translation applies an anti-log or exponential function to the predicted values.

the natural logarithm linking function used to translate the predicted values model from a linear model back to the original count metric results in a one unit change in the IV producing a different magnitude of change to the dependent variable depending on where on the continuum of the dependent variable the change is located.

The log-linear nature of the coefficients produced by CDV models make the difficult to interpret directly. Typically, CDV coefficients are translated using an exponential function to produce *Incidence Rate Ratios* or *IRRs*. that naturally produce multiplicative effects across the range of each IV.

The naturally multiplicative functional form of CDV models makes the explained-variance R^2 metric less useful for DA. This is because CDVs are not guaranteed to produce an increase to the explained-variance R^2 as more IVs are added to the model ().

There are pseudo-R^2s that are better able to characterize model fit for CDVs.

$$\ln y = \sum \beta_{x_i} \tag{7}$$

Unique Features of CDV Models