Journal Title
XX(X):1–14

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

[1]Fors Marsh Group LLC

such 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 variable models*/(*CDMs*) 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 CDMs 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 extend on Blevins, Tsang, and Spain (Blevins, Tsang & Spain, 2015) in extending considerations regarding CDV models to assessing relative importance using DA...

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

As a variant of Shapley Value Decomposition, DA is a generally applicable method for decomposing a fit statistic from a statistical model into components. In discussing how to apply DA to CDV models, there are several key considerations that make

CDV models more complex than linear models with continuous DVs. First, CDVs are generalized linear models and traditionally assume a multiplicative relationship between an IV and the DV such that a one unit change in the IV result in a change in the DV that is the anti-log or exponential function (i.e., e^{β}) of the coefficient. Thus, in a CDV model, predicted counts produced by a change in the IV depends on the location of the DV. For instance, a model coefficient of .5 in a linear model will always add .5 to the DV such that

For a CDV, it is common to interpret the coefficient as an incidence rate ratio 1.65 will produce a change at a count of 3 that is 1.9 higher (i.e., $e^{3+.5*1} - e$)

The multiplicative nature of CDV models complicates fit metric computation in that the standard fit metric, the explained variance R^2 assumes that the underlying model seeks to minimize the sum of the squared residuals from the model. As will be discussed below, CDVs do not meet this criterion.

A second consideration relevant to CDV models is the use of model offsets. An offset is a variable that is intended to reflect *exposure* or differences between observations in the capability for that observation to produce a count. Offset variables are usually factors such as population sizes (cite) or exposure time windows (cite) that will affect the observations' counts and are known about different observations beforehand. Offsets are included into the model with a coefficient of 1 and serve to make the CDV a rate as they adjust the count such that $e^{y_i - offset_i} =$

Finally, a common model applied to CDVs are *zero inflated* models that are recommended for use in modeling CDVs in many situations (see Figure 3 of (Blevins, Tsang & Spain, 2015)). Zero inflated models offer a great deal of flexibility in evaluating the processes

The flexibility of zero inflated models comes at the cost of greater complexity when considering how to evaluate the contributions IVs have to prediction as there are two predictive equations, the count-producing model and the zero-producing model, which need not have the same set of predictors. As such, it may be necessary to examine parameter estimate relative importance (PERI; (Luchman, Lei & Kaplan, 2020)) as opposed to independent variable relative importance (IVRI) when examining

In the sections below, each of the three complications regarding CDVs is discussed in greater detail with a focus on how each can affect the determination of relative importance.

Fit Metrics for Count Dependent Variables

CDMs are inherently multiplicative in how the underlying prediction model relates to the observed count DV. The multiplicative nature of CDMs complicates assessing how well the model fits to the data as the way in which residuals are computed for the LRM does not produce useful. Consider the standard explained variance R^2. The R^2 is the squared Pearson correlation between the DV and the predicted values from the model. Pearson correlations, like the LRM, are based on the sums of squared deviations ...

Consider now the simplest of the CDMs, the Poisson regression model. The Poisson regression minimizes the sum of deviations from the product of the DV and the natural log of the predicted value or $\sum(\mu - y \ln \mu)$ (Cameron & Windmeijer, 1996). The form of the minimizing value for the Poisson makes it is possible that using a metric like

the explained variance R^2 that computes the sum of squared deviations will produce negative contributions to the R^2. Specifically, the explained variance R^2 rewards smaller total or summed squared deviations from the observations. By contrast, a model like the Poisson prefers minimizing deviations with a preference toward shrinking to 0; hence, penalizing more for over-prediction than under-prediction. The differences between these penalty factors mean that a Poisson model might choose a coefficient that shrinks predicted values to such an extent that it appears to reduce fit to the model when applying an explained variance metric. The more complex negative Binomial model shows similar properties and, as is noted by Spain et al., is a direct extension of the Poisson (i.e., a mixture of Poisson and Gamma distributions).

A similar discussion to the one above is provided by Traxel and Azen (Azen & Traxel, 2009) focused on the closely related generalized linear model, Logistic regression. Azen and Traxel show that likelihood-based pseudo-R^2 metrics, in particular the McFadden and Estrella R^2s, are useful for DA with logistic regressions as they always increase with the inclusion of more predictors and show several other desirable properties.

R^2 recommendations specific to CDMs have also been proposed in the literature. Cameron and Windmeijer (Cameron & Windmeijer, 1996) provide perhaps the most comprehensive review and critique of different fit metrics for CDMs. Cameron and Windmeijer, similar to Azen and Traxel, focus on desirable properties of R^2s but conclude that the R^2 that best reflects fit to the data for CRMs is the deviance R^2 or R^2_{DEV}. The deviance R^2 is computed as:

$$R^2_{DEV} = \frac{D_{model}}{D_{null}}$$

Where D is the GLM deviance residual computation. In fact, the McFadden R^2 recommended by Axen and Traxel is tantamount to the deviance R^2 for the logistic model given the way the model is scaled, but differs for CRMs by a factor of ...

In sum, the recommended R^2 metric is the deviance R^2 which has a number of desirable properties that make it similar to the explained variance R^2 across a number of the models discussed by Blevins, Tsang, and Spain (Blevins, Tsang & Spain, 2015). These desirable properties make R^2_{DEV} the fit metric that is recommended for use in DA when using CRMs. Further discussion of fit metrics for modeling and DA in CRMs is continued below as applied to the data analytic examples.

Exposure and Offset Terms

CDMs, as models of counts of discrete events, implicitly assume that the events being counted are comparable. An observation with a count of 10 is assumed to have had the same chance to get that count of 10 events as any other observation with 10 events. That is, that the probability of observing events for both observations is equal. By contrast, it is possible two observations with an observed count of 10 arrived at their count with different underlying probabilities of observing events and number of times the events could be realized into a count. For example, two 10 counts could have resulted from an observation with an event probability of .1 over 100 'trials'. Another could have arisen from an observation with an event probability of .01 over 1000 'trials'. In

data, such equifinality in counts from different underlying probabilities can arise from observations that are units reflecting populations of different sizes or units reflecting different time periods in which events could occur (i.e., unequal employment tenures).

CDMs can control for such unequal *exposure* to the count generating conceptual phenomenon using what is known as an *offset* term. The offset term is usually assumed to be natural log transformed for a CDM and enters into the CDM with a coefficient of 1. The coefficient of 1 results in the count DV being transformed into a rate out of the offset variable. How the count DV is transformed into a rate with a coefficient of 1 follows from the following algebraic manipulations. First, consider a simple case where an intercept only model is fit with an offset such as $y = e^{offset+\beta}$. Recall that the offset term is the natural log of a variable, say, o. Applying the natural log, the prior equation can be written as $\ln y = \ln o + \beta$. Re-arranging the $\ln o$ term results in $\ln y - \ln o = \beta$ which, given the properties of logarithms, is tantamount to $\ln \frac{y}{o} = \beta$. Therefore, the inclusion of a natural log transformed offset variable results in the count DV transforming into a rate. This offset transformation corrects for the issue discussed above as the 10 counts result in the correct underlying rate of .1 versus .01 for each observation.

In large part, an offset adjustment serves to re-scale specific observations' predictions and has the biggest effect on CDM's intercept value; adjusting it back to reflect the correct average rate across observations. The offset can, however, affect the magnitude of estimated coefficients if the offset is correlated with an IV. When the offset is correlated with an IV, then the IV may be both related to exposure to the count generating phenomenon as well as the rate of count generation broadly. Not including the offset can then bias the CDM and can have a notable impact on IV relative importance. How offsets affect CRM modeling and importance determination will also be explored further in the empirical examples below.

Multiple Equations

... Poisson vs alternatives ...

Data Analytic Examples

The purpose of this manuscript is to recommend approaches for addressing three key considerations that I argue are impactful for determining the relative importance of IVs in CRMs using DA. This section extends on the conceptual discussion of each consideration above by demonstrating exactly how each can affect relative importance determination using a series of data analytic examples. The data analytic examples outlined below are based on computer simulated data and have been designed to illustrate the impact of the selected considerations discussed above as applied to both parameter estimation and importance determination. The section below describes the data simulation process and provides rationale for how the data were generated.

Data Generation

All the simulated data for the data analytic examples in this manuscript were generated from series of four underlying data structures. The four data structures designed for

this manuscript were intended to highlight the impact of IV variance as well as inter-IV covariance on both parameter estimation as well as relative importance determination. I expected that these four designs would also produce useful comparisons across CRMs in terms of how each is affected by IV (co-)variances.

The first data structure designed generated all IVs with equal variances and no/zero covariances. This data structure, referred to as EU (i.e., *E*qual and *U*ncorrelated), does not reflect "real" data but is a commonly simulated and useful data structure to show the effect of a model's parameter estimation and relative importance determination results under what could be considered "ideal" estimation conditions when the causal effects of IVs are easiest to identify given they do not overlap. The EU data structure is then most useful as a comparison point for the other, more realistic data structures to more clearly assess the impact of differences in (co-)variances.

The second data structure designed was similar to the first in that all IVs were generated with equal variances but were also generated such that the IVs had non-zero covariances. This data structure, referred to as EC (i.e., *E*qual and *C*orrelated), is intended to show the effect of IV intercorrelations on both parameter estimation and relative importance determination as compared to EU. I expect that the intercorrelations imposed in this data structure would be most impactful for relative importance determination which attempts to isolate variance explained by each IV irrespective of its association with other IVs. The EC data structure is useful as a condition to evaluate the effects of IV intercorrelations on parameter estimates and relative importance independent of the underlying predictor variances.

The third was an uncorrelated with unequal half-life type variance reduction...

The final was a combination of the half-life correlation and variances... In all cases, the strength of the correlations generated between the variables varied inversely with the size of the variance generated for that variable

For each of the four data structures, a single dataset with 100,000 observations and four IVs was generated based on the theoretical data structure. For the EU and EC data structures, all variances were set to 1. For the UU and UC data structures, the variance for *V1* was set to 1, for *V2* was set to 1.5, for *V3* was set to 2, and for *V4* was set to 4. These differences in variance were intended to increasingly enhance the relative importance of IVs with larger numeric labels as balanced with each variables' effect on the DV to be discussed below. The EC and UC data structures shared a single underlying correlation matrix reported below in Table X.

	V_1	V2	V3	V4
V_1	1.0	0.5	0.25	0.125
V2	0.5	1.0	0.125	0.25
V3	0.25	0.125	1.0	0.5
V4	0.125	0.25	0.5	1.0

The correlation matrix in Table X produces a realistic correlations that emphasize the relationship between *V1* and *V2* as well as the relationship between *V3* and *V4*. The balance of the correlations in Table X also does not result in one variable being more redundant than others; all IVs have equal redundancies as others, but are more or less redundant with different other IVs. The means for all variables were set to 0.

Each data structure's dataset was also used to generate all DVs. An initial Gaussian/Normally distributed DV was generated for each dataset using the following equation: $Y_{cont} = V_1 * .2 + V_2 * .15 + V_3 * .125 + V_4 * .1$. A residual residual variance for Y_{cont} was generated for each dataset as $1 - \sigma_{Y_{cont}}$; hence, Y_{cont} had a variance of 1 and a mean of 0. The Normal distributed variable was generated both as a point of comparison for the CRM results as a well as to serve as the basis for generating each count DV discussed further below.

A Poisson distributed version of Y_{cont}, Y_{pois}, was also generated by obtaining the cumulative probability of each observations's Y_{cont} score given the standard Gaussian/Normal distribution (i.e., $N(0, 1)$) and translating that value into a count value using the Poisson quantile function assuming an underlying mean and variance of 1. The mean of 1 value was chosen to make the Poisson distributed result as close to the Normally distributed result as possible as both have variances of 1.

A negative Binomial version, Y_{nb}, was created similarly by translating Y_{cont}'s cumulative probability for each observation into a count variable using the negative Binomial quantile function assuming a Gamma shape parameter of .5 and scale parameter of .33; these factors result in a mean of 1 and variance of 3 for the negative Binomial distribution. The negative Binomial parameters obtaining a mean of 1 and variance of 3 were chosen to maintain similarities to the Poisson version of the variable (i.e., their means are identical) but to emphasize the unique feature of the negative Binomial distribution (i.e., the capability to model over-dispersion).

... offets ...

... ZIP ...

The data for the analytic examples were generated using R (R Core Team, 2022) using a combination of base R's `stats` package as well as the `MASS` package (Venables & Ripley, 2002). Other details of data generation for the reproducibility of the analysis can be found in Appendix X. The section below transitions from discussing data generation processes to data analysis.

Parameter Estimation Results

The data analysis begins where all relative importance analysis should, with parameter estimation using the appropriate statistical model for the DV. Relative importance analysis is an extension of parameter estimation (cite!) that provides additional, useful information about the model and data. As such, relative importance analysis results should be interpreted in light of the parameter estimates produced by the model that is being dominance analyzed. The goal of this section is to highlight similarities and differences between the models' results across the four data structures with attention to each data structure's effect parameter estimate magnitudes.

Base Model Results This subsection discusses the parameter estimates obtained from the Normal, Poisson, and negative Binomial distributed DVs across all four data structures. Note that descriptive statistics for the IVs and DVs are reported in Appendix X. The results from the four models estimated on the Normal DVs are listed below in Table X.

Parameter	Component	Coefficient.Model 1	SE.Model 1	CI.Model 1	CI_low.Model 1	CI_high.Model
(Intercept)	conditional	0.00264	0.00302	0.95	-0.00328	0.0085
V1	conditional	0.20024	0.00302	0.95	0.19431	0.2061
V2	conditional	0.14897	0.00302	0.95	0.14307	0.1548
V3	conditional	0.12048	0.00300	0.95	0.11460	0.1263
V4	conditional	0.09941	0.00301	0.95	0.09352	0.1053

The effects of data structure on estimation for Normally distributed variables with linear regression are well-known and the results in Table X are, as a result, consistent with extant research in showing that these data tended to recover the correct parameter estimates.

...

data generated were a rather large population size (100,000) and our model was correctly specified.

Hence, it is not a surprise that the known model parameters were recovered very well from the data across all four data structures and models.

Recall that the true parameters were .2 for X_1, .15 for X_2, .125 for X_3, and .1 for X_4. Also recall that these linear regression equations underlie the generation of both the Poisson and negative Binomial DVs. I expect then that the linear regression results will serve as a useful comparison point for the CRM parameter estimation results.

Parameter	Component	Log-Mean.Model 1	SE.Model 1	CI.Model 1	CI_low.Model 1	CI_high.Model 1
(Intercept)	conditional	-0.03429	0.00327	0.95	-0.04070	-0.02789
V1	conditional	0.18262	0.00318	0.95	0.17640	0.18884
V2	conditional	0.13493	0.00316	0.95	0.12875	0.14112
V3	conditional	0.10691	0.00314	0.95	0.10075	0.11307
V4	conditional	0.09054	0.00315	0.95	0.08437	0.09671

Numerically, the Poisson results look rather similar to the linear regression results as was intended. Indeed, the linear regression prediction equation (with error) was used to generate the underlying Poisson distributed variable and both had the same underlying variance of 1. A further similarity between the linear and Poisson results is that the coefficients obtained across data structures were nearly identical with the exception of the intercept.

Parameter	Component	Log-Mean.Model 1	SE.Model 1	CI.Model 1	CI_low.Model 1	CI_high.Model 1
(Intercept)	conditional	-0.08308	0.00535	0.95	-0.09355	-0.07260
V1	conditional	0.28582	0.00535	0.95	0.27534	0.29629
V2	conditional	0.20828	0.00530	0.95	0.19789	0.21866
V3	conditional	0.16658	0.00526	0.95	0.15626	0.17690
V4	conditional	0.13695	0.00527	0.95	0.12663	0.14728

The negative Binomial regression results, by contrast to the Poisson, were numerically less similar to the linear regression and more noticeably sensitive to data

conditions during model estimation. For instance, the negative Binomial results tended to produce larger coefficients when the variables were both correlated and had unequal variances. The next section shifts to focus on the effect of the model, data structure, and fit metric chosen have on the determination of relative importance.

Offset Models

Zero-inflated Models

Dominance Analysis

The model parameter estimates reported on above can be used as a basis for interpretation of the critical set of results for this manuscript; the DA statistics. Above I note that the data structures chosen for this manuscript had little effect on the coefficient estimation broadly. I do not expect that the data structure will be as passive a factor for DA statistic computation and, in fact, the data structures chosen for this paper were chosen as they were expected to have a notable effect on the DA results.

In this section below, the linear, Poisson, and negative Binomial DVs were also submitted to DAs with their appropriate model to determine importance among the IVs. The results from each DA is presented in the table X.

vars	data	Linear Model	Poisson Model A	Poisson Model B	Negative Binomial Model A	Negative Bin
1	EC	0.05926	0.04362	0.05043	0.04607	
2	EC	0.04636	0.03364	0.03900	0.03565	
3	EC	0.03055	0.02145	0.02424	0.02275	
4	EC	0.02445	0.01726	0.01977	0.01887	
1	EU	0.04008	0.02902	0.03301	0.03047	
2	EU	0.02234	0.01604	0.01848	0.01652	
3	EU	0.01474	0.01015	0.01160	0.01051	
4	EU	0.00989	0.00717	0.00836	0.00718	
1	UC	0.07205	0.05260	0.05959	0.05804	
2	UC	0.06739	0.04926	0.05622	0.05246	
3	UC	0.06267	0.04523	0.05118	0.04887	
4	UC	0.07225	0.05192	0.05878	0.05657	
1	UU	0.04025	0.02848	0.03243	0.02942	
2	UU	0.03388	0.02492	0.02851	0.02557	
3	UU	0.03130	0.02272	0.02611	0.02370	
4	UU	0.03911	0.02835	0.03238	0.03024	

Table X shows the effect of the intended differences across data structures as realized in the DA results.

Table X shows stark differences between the data structures in terms of the dominance statistics obtained by each of the IVs. The EU data structure

The effect of the data structure is apparent in these results. The uncorrelated, equal variance results generally recovered the correct amount of explained variance for each IV. When introducing correlations to the model, ...

Whereas there were clear differences between the data structures in terms of importance determination, the results across fit metrics within a model were quite similar. Table X shows that the proportional contribution to the fit metric each IV was ascribed remained stable across fit metrics within a model. For instance, IV_1s proportional contribution for the Poisson model... Moreover, the dominance determinations did not differ across metrics within a model. IV_1 completely dominated all other...

One way to interpret the results presented above is to suggest that relative importance determination for CRMs is insensitive to fit metric choice. Despite their similarities however, the explained variance R^2 values are consistently discrepant from the deviance R^2 values with the Poisson model producing values that were too high and the negative Binomial producing values that were too low. This result then suggests that the primary risk of using the explained variance R^2 as opposed to the deviance R^2 is in mis-characterizing the extent of predictive usefulness of the model as a whole, which affects the the numeric magnitudes or results obtained by the DA, but metric choice does not, itself, affect relative importance determinations.

Modeling with Exposure

...

Modeling with Zero-inflation

Poisson and linear regression share many similarities and, as a result of these similarities, often give similar answers when a count DV is applied to either model. Indeed the DA results for the Poisson model with the deviance R^2 are strikingly similar to those obtained from the Gaussian DV DA both in terms of the magnitudes of the deviance R2. This result is as expected in that the Poisson results are an extension of the Gaussian DV-based results and the deviance R^2 is a direct generalization of the explained variance R^2 to generalized linear models.

When applying

...

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. By contrast, two hypothetical less ideal predicted values might be 4 and 6. In both cases, the log-likelihood values diverge from the best fitting one and increase, indicating a lower likelihood. Specifically, 4 gives $5 * \ln(4) - \ln(5) + 4 = -1.856$ and 6 gives $5 * \ln(6) - \ln(4) + 6 = 1.829$... This also shows an important point about the Poisson log likelihood–that it is not symmetric around the best fitting value. By contrast, ordinary least squares regression would penalize both 4 and 6 as squared deviations of 1 from the line of best fit (footnote that logged DV can do this too - but the loglik is still symmetrical and the model is now fitting the mean of a transformed DV as opposed to the original DV). That the model fit to the data is asymmetrical has implications for assessing fit to the model that will be discussed in greater detail later.

Negative Binomial: A More Complex and Flexible Case

A constraint with the Poisson model is that it has only a single parameter estimated from the data, it's mean. The variance of the Poisson distribution is, by assumption, identical to the mean.

One way to extend on the Poisson to make it more flexible is to conceptualize the mean of the Poisson distribution as a random variable. If the mean of the Poisson is conceptualized as Gamma distributed, the hybrid Poisson-Gamma distribution can be reduced to a unique distribution called the negative Binomial.

The negative Binomial model is a more complex generalization of the Poisson that relaxes the assumption that the variance and mean are identical

...negative binomial...

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

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.

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

References

Azen, R. & Traxel, N. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, *34(3)*, 319–347.

Blevins, D. P., Tsang, E. W. & Spain, S. M. (2015). Count-based research in management: Suggestions for improvement. *Organizational Research Methods*, *18(1)*, 47–69.

Cameron, A. C. & Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, *14(2)*, 209–220.

Luchman, J., Lei, X. & Kaplan, S. (2020). Relative importance analysis with multivariate models: Shifting the focus from independent variables to parameter estimates. *Journal of Applied Structural Equation Modeling*, *4(2)*, 1–20.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth Ed.). New York: Springer. ISBN 0-387-95457-0.